Improving knockoffs with conditional calibration

Yixiang Luo, William Fithian, and Lihua Lei

June 22, 2022

Abstract

The knockoff filter of Barber and Candès (2015) is a flexible framework for multiple testing in supervised learning models, based on introducing synthetic predictor variables to control the false discovery rate (FDR). Using the conditional calibration framework of Fithian and Lei (2020), we introduce the *calibrated knockoff procedure*, a method that uniformly improves the power of any knockoff procedure. We implement our method for fixed-X knockoffs and show theoretically and empirically that the improvement is especially notable in two contexts where knockoff methods can be nearly powerless: when the rejection set is small, and when the structure of the design matrix prevents us from constructing good knockoff variables. In these contexts, calibrated knockoffs even outperform competing FDR-controlling methods like the (dependence-adjusted) Benjamini–Hochberg procedure in many scenarios.

1 Introduction

The Gaussian linear regression model is one of the most versatile and best studied models in statistics, with myriad applications in experimental analysis, causal inference, and machine learning. In modern applications, there are commonly many explanatory variables, and we suspect that most of them have little to do with the response, i.e. that the true coefficient vector is (approximately) *sparse*. In such problems, multiple hypothesis testing methods are a natural tool for discovering a small number of variables with nonzero coefficients in the sea of noise variables, while controlling some error measure such as the false discovery rate (FDR), introduced by Benjamini and Hochberg (1995).

At present, however, the multiple testing literature offers practitioners little clarity regarding *how* they ought to perform the inference. There are at least two well-known methods for multiple testing with FDR control: the (fixed-X) *knockoff filter* of Barber and Candès (2015) and the Benjamini–Hochberg (BH) procedure of Benjamini and Hochberg (1995) (recently modified by Fithian and Lei (2020) to ensure provable FDR control in linear regression among other problems with dependent p-values). Knockoffs and BH use radically different approaches and can return very different rejection sets on the same data, and it is not uncommon for one method to dramatically outperform the other, depending on the problem context. For example, Section 1.2 illustrates a simple problem setting where BH has much higher power at FDR level $\alpha = 0.05$, but the knockoff filter recovers and outperforms BH at level $\alpha = 0.2$. In particular, the knockoff filter suffers from a so-called *threshold phenomenon*, explained in Section 2.1, that makes it nearly powerless when the number of discernibly non-null variables is smaller than $1/\alpha$, making it a risky choice for an analyst who aims for more stringent FDR control. In problems with enough rejections to avoid this issue, however, the knockoff filter often shines, since it can use efficient estimation methods like the lasso (Tibshirani, 1996) to guide its prioritization of variables.

In this work, we propose a new method, the *calibrated Knockoff procedure* (cKnockoff), which uniformly improves the knockoff filter's power while achieving finite-sample FDR control in the Gaussian linear model. Our method acts as a "wrapper" around any implementation of fixed-X knockoffs, augmenting its rejection set using a "fallback test" for each variable that is not already rejected by knockoffs. To set the power of the fallback tests without violating FDR control, we use the conditional

calibration framework proposed in Fithian and Lei (2020). cKnockoff is strictly more powerful than knockoffs in every problem instance, but the power gain is especially large in problems with a small number of non-null variables, resolving the threshold phenomenon while retaining the knockoff filter's advantages.

1.1 Multiple testing in the Gaussian linear model

We consider the linear model relating an observed response vector $\boldsymbol{y} = (y_1, \dots, y_n)^\mathsf{T}$ to fixed explanatory variables $\boldsymbol{X}_j = (X_{1j}, \dots, X_{nj})^T$, for $j = 1, \dots, m$ via

$$\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{X}_j \beta_j + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{1}$$

where the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ has $\boldsymbol{X}_j$ as its jth column. Both the coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^\mathsf{T}$ and the error variance σ^2 are assumed to be unknown. We assume throughout that $m < n$, and that X has full column rank, ensuring that $\boldsymbol{\beta}$ and σ^2 are identifiable.

A central inference question in this model is whether a given variable $\boldsymbol{X}_j$ helps to explain the response, after adjusting for the other variables. Formally, we will study the problem of testing the hypothesis $H_j : \beta_j = 0$ for each variable $\boldsymbol{X}_j$ simultaneously, while controlling the FDR.

Let $\mathcal{H}_0 = \{j : H_j \text{ is true}\}$ and $m_0 = |\mathcal{H}_0|$; we say $\boldsymbol{X}_j$ is a *null variable* if $j \in \mathcal{H}_0$. For a multiple testing procedure with rejection set $\mathcal{R} \subset \{1, \dots, m\}$, the *false discovery proportion* (FDP) and FDR are defined respectively as

$$\mathrm{FDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}, \quad \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

We write $R = |\mathcal{R}|$ and $V = |\mathcal{R} \cap \mathcal{H}_0|$ for the number of rejections and false rejections respectively. Our goal is to control FDR at a pre-specified threshold α while achieving a power as high as possible. Throughout this paper we define power in terms of the *true positive rate* (TPR), defined as the expectation of the *true positive proportion* (TPP), the fraction of the $m_1 = m - m_0$ non-null hypotheses rejected:

$$\mathrm{TPP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_0^\mathsf{c}|}{m_1}, \quad \mathrm{TPR} = \mathbb{E}[\mathrm{TPP}].$$

A traditional approach to multiple testing would start with the usual two-sided t-test statistics $|\hat{\beta}_j|/\hat{\sigma}$, which are calculated from the ordinary least squares (OLS) estimator and the unbiased estimator of the error variance

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}, \quad \text{and } \hat{\sigma}^2 = \mathrm{RSS}/(n-m),$$

where $\mathrm{RSS} = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$ is the residual sum of squares. Such t-tests are uniformly most powerful unbiased for the individual hypotheses. Then an appropriate multiplicity correction is applied to their corresponding p-values. The celebrated *Benjamini–Hochberg procedure* (BH), the best-known FDR-controlling method, orders the p-values from smallest to largest $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$, and rejects

$$\mathcal{R}^{\mathrm{BH}} = \left\{ j : p_j \leqslant \frac{\alpha R^{\mathrm{BH}}}{m} \right\}, \quad \text{where} \quad R^{\mathrm{BH}} = \max\left\{ r : \; p_{(r)} \leqslant \frac{\alpha r}{m} \right\}.$$

While BH does not provably control FDR in this context due to the dependence between p-values, a corrected version called the *dependence-adjusted BH procedure* (dBH) does, while achieving nearly identical power (Fithian and Lei, 2020).

The knockoff filter, described below in Section 2.1, is a flexible class of methods that take a radically different approach, completely bypassing the t-tests. Instead, these methods introduce a "knockoff" variable $\tilde{\boldsymbol{X}}_j$ to serve as a negative control for each real predictor variable $\boldsymbol{X}_j$, and then

apply a learning algorithm to rank the $2m$ variables according to some importance measure in the model. The knockoffs are constructed to ensure that, under H_j, $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ are indistinguishable in an appropriate sense. We emphasize that the knockoff filter does not relax the t-test assumptions in any way: it controls FDR in model (1) in finite samples, for any β and σ^2.

Knockoff methods enjoy substantially higher power than BH and dBH in some scenarios while struggling in others, with the relative performance depending on the problem dimensions, the structure of the design matrix, and the true β vector, among other considerations. Figure 2 illustrates one such stark contrast, where the knockoff filter outshines BH at FDR level $\alpha = 0.2$ but struggles at level $\alpha = 0.05$, in the same instance of a scenario we call *blockwise multiple comparisons to control* (MCC-Block), a variation on the classic MCC problem of Dunnett (1955), which we describe next.

1.2 Motivating example: blockwise multiple comparisons to control

To illustrate why dependence between test statistics can offer opportunities to improve on BH, we now introduce a very simple linear modeling problem that we will use as a running example throughout the remainder of this paper. We will see that, when the dimension m is large and the coefficient vector β is sparse, the variables whose p-values are smallest may not be the most favorable.

Example 1.1 (MCC-Block)**.** *For treatment group $g = 1, \dots, G$ in block $k = 1, \dots, K$, we observe r independent replicates*

$$z_{g,k,i} = \mu_k + \delta_{g,k} + \varepsilon_{g,k,i}, \qquad \textit{where} \quad \varepsilon_{g,k,i} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, r$$

along with r independent replicates for a control group in the same block:

$$z_{0,k,i} = \mu_k + \varepsilon_{0,k,i}, \qquad \textit{where} \quad \varepsilon_{0,k,i} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, r.$$

The effects of interest are the location shifts for each treatment group, $\delta_{g,k}$.

Example 1.1 can naturally arise in experimental contexts where μ_k represents a fixed or random "batch effect" for a set of observations. In applied settings, the number of groups in each block or the number of replicates per group may be variable. The assumption of Gaussian errors with common variance can easily be relaxed if r, the number of independent replicates per group, is large.

After "projecting out" the block effects, Example 1.1 can be equivalently expressed as a linear model of the form (1) with $m = KG$ variables and $n = K(r(G+1) - 1)$ independent observations, where $\beta_j = \delta_{g,k}$ for $j = (k-1)G + g$, and the design matrix X and response vector y are given by appropriate linear contrasts; see Appendix E.1 for details. The OLS estimator $\hat{\boldsymbol{\beta}}$ is easily shown to be

$$\hat{\beta}_{(k-1)G+g} = \hat{\delta}_{g,k} := \overline{z}_{g,k} - \overline{z}_{0,k} = \delta_{g,k} + \overline{\varepsilon}_{g,k} - \overline{\varepsilon}_{0,k} \;\sim\; N(\delta_{g,k},\, 2\sigma^2/r), \tag{2}$$

where $\overline{z}_{g,k}$ and $\overline{z}_{0,k}$ respectively represent the sample means for the gth treatment group and the control group in block k, and $\overline{\varepsilon}_{g,k}$ and $\overline{\varepsilon}_{0,k}$ are defined similarly. Because estimates for the same block use the same control group, the correlation matrix of $\hat{\boldsymbol{\beta}}$ is block-diagonal with correlation 0.5 for pairs of entries in the same block, and zero correlation across blocks. In the special case $K = 1$, Example 1.1 reduces to the classical MCC problem.

To test any individual hypothesis, the two-sided t-test using t-statistic $\sqrt{r/2\hat{\sigma}^2} \cdot \hat{\beta}_j$ seems virtually unassailable. As a result, it would be quite natural to apply the BH method, or its close cousin dBH, with the t-test p-values. To understand why BH is sub-optimal in this problem, we must carefully consider the implications of *sparsity*: that we may expect $\delta_{g,k} \approx 0$ for the vast majority of groups. Hence if we observe

$$\hat{\delta}_{1,k} \approx 0 \quad \text{and} \quad \hat{\delta}_{g,k} \approx -1, \quad g = 2, \dots, G$$

for some k, this would be a strong hint for $\delta_{1,k} \approx 1 \approx \overline{\varepsilon}_{0,k} - \overline{\varepsilon}_{g,k}$ under the sparsity assumption. While the expectation of sparsity is natural and powerful in many multiple testing contexts, the OLS

estimator that forms the backbone of the BH method does not incorporate a sparsity assumption in any way, since the distributions of the jth t-statistic, $\sqrt{r/2\hat{\sigma}^2} \cdot \hat{\beta}_j$ and its p-value p_j depend only on β_j/σ.

If $\boldsymbol{\beta}$ is sparse, however, we should be able to exploit the sparsity to improve our estimator of $\boldsymbol{\beta}$, for example by using the *lasso estimator* of Tibshirani (1996), defined by

$$\hat{\boldsymbol{\beta}}^{\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^m}{\operatorname{argmin}} \ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \cdot \|\boldsymbol{\beta}\|_1 . \tag{3}$$

In particular, Figure 1 shows that the lasso estimator does a better job of ordering the variables in an instance of the MCC-Block problem with sparse $\boldsymbol{\beta}$. This suggests that the BH method, which is restricted to rejecting variables in order of their t-statistics, will be inherently limited in its statistical power relative to a method that tracks the lasso estimator instead.

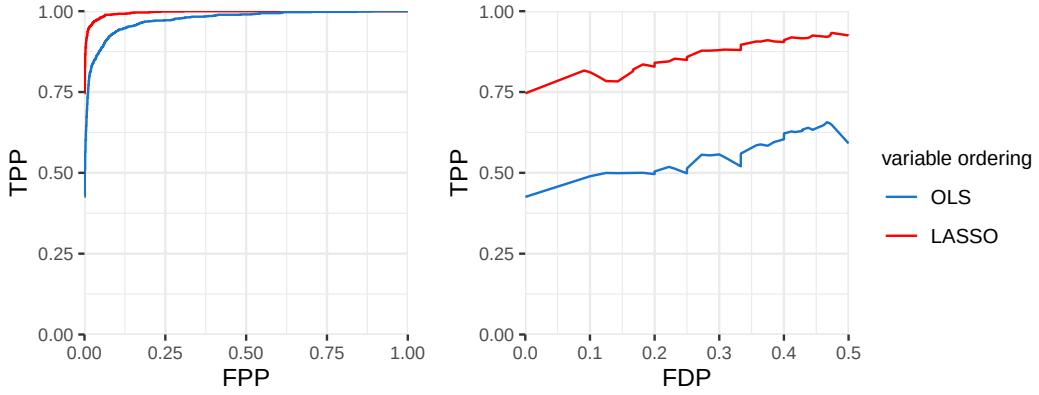

Figure 1: Lasso regression (3) achieves a better variable ordering than OLS regression (2) in the MCC-Block problem with sparse $\boldsymbol{\beta}$. Left: receiver operator characteristic (ROC) curve, TPP versus FPP $= V/m_0$; Right: TPP versus FDP curve, both averaged over 100 realizations. In each realization the variables are "rejected" in decreasing order of $|\hat{\beta}_j|$. For the lasso, we use $\lambda = 2\hat{\sigma}$, where $\hat{\sigma}^2$ is the unbiased variance estimator, and break ties among variables with $\hat{\beta}_j = 0$ according to the magnitude of their correlation with the lasso residual. We simulate $K = 200$ blocks, $G = 5$ treatment groups per block, and $r = 3$ replicates per group. There are $m_1 = 10$ nonzero effects with equal strength, distributed at random across the $m = KG = 1000$ total hypotheses, with the signal strength calibrated so that BH at level $\alpha = 0.2$ attains TPR $= 0.5$.

As shown in Figure 2, a version of the knockoff filter based on the lasso can likewise outperform the OLS-based BH method, but its superior performance is only observed in this instance for $\alpha = 0.2$. For smaller α values, a specific drawback of knockoffs — ironically, that knockoff methods break down when the coefficient vector is *too* sparse — prevents the method from realizing its potential. This drawback is resolved by our calibrated knockoff method, the main subject of this work.

1.3 Outline and contributions

In this work, we propose the *calibrated Knockoff procedure* (cKnockoff), a method that controls finite-sample FDR in the Gaussian linear model with fixed design. Our method acts as a "wrapper" around any implementation of fixed-X knockoffs, uniformly improving its power by means of a *fallback test* that allows for the rejection of variables not rejected by knockoffs.

For a generic fallback test statistic $T_j(\boldsymbol{y})$, we calibrate a data-adaptive rejection threshold $\hat{c}_j(\boldsymbol{y})$, and reject H_j for any j in the knockoff rejection set *or* which has $T_j(\boldsymbol{y}) \geqslant \hat{c}_j(\boldsymbol{y})$. That is,

$$\mathcal{R}^{\mathrm{cKn}} \ = \ \mathcal{R}^{\mathrm{Kn}} \ \cup \ \{j : \ T_j(\boldsymbol{y}) \geqslant \hat{c}_j(\boldsymbol{y})\},$$

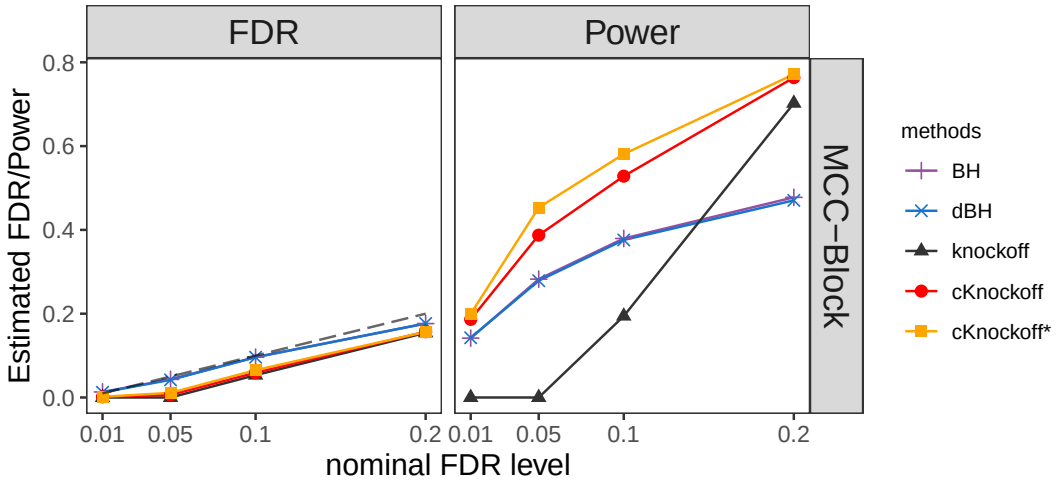

Figure 2: Performance of several multiple testing methods for the same instance of the MCC-Block problem as in Figure 1, for varying FDR significance levels, $\alpha = 0.01, 0.05, 0.1$, and 0.2. The knockoff method using lasso-based LCD feature statistics outperforms BH by a wide margin when $\alpha = 0.2$ by successfully exploiting sparsity, but it fails for smaller values of α due to the threshold phenomenon. The cKnockoff method described in this paper outperforms both BH and knockoffs.

where $\mathcal{R}^{\mathrm{Kn}}$ and $\mathcal{R}^{\mathrm{cKn}}$ are respectively the rejection sets for the baseline and calibrated knockoff methods.

The threshold $\hat{c}_j$ is calibrated to control H_j's contribution to the overall FDR, using the conditional FDR calibration method of Fithian and Lei (2020), which we review in more detail in Section 2.3. In brief, define $S_j = (\boldsymbol{X}^{\mathsf{T}}_{-j}\boldsymbol{y},\ \|\boldsymbol{y}\|_2^2)$, the complete sufficient statistic for the submodel described by H_j. Then under H_j, the distribution of y given S_j is known so that, for any fallback test threshold c_j, we can calculate the resulting *conditional FDR contribution* of H_j, defined as

$$\mathrm{FDR}_j(\mathcal{R}^{\mathrm{cKn}} \mid S_j) \ := \ \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{cKn}}\right\}}{|\mathcal{R}^{\mathrm{cKn}}| \vee 1} \,\middle|\, S_j\right] \ \leqslant \ \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant c_j\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \,\middle|\, S_j\right].$$

We will choose the threshold $\hat{c}_j(S_j(\boldsymbol{y}))$ to set the last expression equal to a variable-specific, data-adaptive budget that we obtain by analyzing the FDR control proof for knockoffs.

Because the cKnockoff rejection set almost surely includes the knockoff rejection set and sometimes exceeds it, the method is uniformly more powerful than fixed-X knockoffs. We find in simulations that the power gain is especially large when the true $\boldsymbol{\beta}$ vector is very sparse, in particular when we do not have $m_1 \gg 1/\alpha$. The same ideas can be applied to model-X knockoffs, but efficiently extending them to that context will require significant computational finesse, which we leave to future work.

The only downside of cKnockoff is the additional computation it requires. To reduce this burden, we only carry out the fallback test on hypotheses that appear promising, and we use a conservative approximation to speed the fallback test calculation. We prove that these speedup techniques do not inflate the FDR, and we find numerically that the computation time of our implementation is a small multiple of the knockoff computation time, which is further improved when parallel computing is available.

Section 2 reviews the basics of knockoffs and conditional calibration, and Section 3 defines our method in full detail. Section 4 gives more detail about how we implement the fallback test for a given variable, using a single Monte Carlo integral. Sections 5–6 illustrate our method's performance on selected simulation scenarios, in addition to the HIV data from the original knockoff paper (Barber and Candès, 2015), and Section 7 concludes.

2 Review: knockoffs and conditional calibration

2.1 Knockoffs: a flexible framework

This section reviews the elements of knockoffs and conditional calibration. Our focus in this paper is on *fixed-X knockoffs*, the original version of the knockoff filter proposed in Barber and Candès (2015) for the Gaussian linear model with fixed design. In this setting, the requisite "indistinguishability" is defined by the pairwise correlations between variables. Specifically, the knockoff design matrix $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_d) \in \mathbb{R}^{n \times m}$ must satisfy

$$\tilde{\boldsymbol{X}}^\mathsf{T} \tilde{\boldsymbol{X}} = \boldsymbol{X}^\mathsf{T} \boldsymbol{X}, \quad \text{and} \quad \boldsymbol{X}^\mathsf{T} \tilde{\boldsymbol{X}} = \boldsymbol{X}^\mathsf{T} \boldsymbol{X} - \boldsymbol{D}, \tag{4}$$

for some diagonal matrix $\boldsymbol{D}$. Following Barber and Candès (2015), we require $m \geqslant 2n$; in some cases we will require further that $m \geqslant 2n + 1$ so that σ^2 is identifiable in the augmented linear model with design matrix $\boldsymbol{X}_+ = (\boldsymbol{X}, \tilde{\boldsymbol{X}}) \in \mathbb{R}^{n \times 2m}$.

There are many ways to implement knockoffs, but every knockoff method yields common intermediate outputs called *feature statistics* $W_1, \ldots, W_m \in \mathbb{R}$ which are inputs to an ordered multiple testing algorithm called *Selective SeqStep*, also proposed in Barber and Candès (2015). The absolute value $|W_j|$ roughly quantifies how much overall importance the learning algorithm assigns to the pair $\{\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j\}$, while the sign $\operatorname{sgn}(W_j)$ is positive if the algorithm assigns a greater importance to $\boldsymbol{X}_j$ than $\tilde{\boldsymbol{X}}_j$, and negative otherwise. Formally, each W_j must be a function of $\boldsymbol{X}_+^\mathsf{T} \boldsymbol{X}_+$ and $\boldsymbol{X}_+^\mathsf{T} \boldsymbol{y}$ (the *sufficiency condition*) and W_j must have the same absolute value but opposite sign whenever we swap $\boldsymbol{X}_j$ with $\tilde{\boldsymbol{X}}_j$ (the *anti-symmetry condition*).

The two most popular feature statistics in practice, proposed by Barber and Candès (2015) and Candès et al. (2018) respectively, are both based on the Lasso estimator for the augmented model:

$$\hat{\boldsymbol{\beta}}^\lambda = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2m}}{\operatorname{argmin}} \ \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X}_+ \boldsymbol{\beta} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\beta} \right\|_1 .$$

The *lasso signed-max* (LSM) statistics are defined by variables' entry points on the regularization path:

$$W_j^{\mathrm{LSM}} = (\lambda_j^* \vee \lambda_{j+m}^*) \cdot \operatorname{sgn}(\lambda_j^* - \lambda_{j+m}^*), \quad \text{for} \quad \lambda_j^* = \sup \left\{ \lambda : \ \hat{\beta}_j^\lambda \neq 0 \right\}, \tag{5}$$

while the *lasso coefficient-difference* (LCD) statistics are defined by the lasso estimator for a fixed λ:

$$W_j^{\mathrm{LCD}} = |\hat{\beta}_j^\lambda| - |\hat{\beta}_{j+m}^\lambda|. \tag{6}$$

If $\beta_j^\lambda = \beta_{j+m}^\lambda = 0$, then $W_j^{\mathrm{LCD}} = 0$. For the simulations in this paper we use a minor modification of $\boldsymbol{W}^{\mathrm{LCD}}$ that breaks ties using the variables' correlations with the lasso residuals

$$\boldsymbol{r}^\lambda = \boldsymbol{y} - \boldsymbol{X}_+ \hat{\boldsymbol{\beta}}^\lambda. \tag{7}$$

Formally, we define the *LCD with tiebreaker* (LCD-T) statistics as

$$W_j^{\text{LCD-T}} \ = \ \begin{cases} W_j^{\mathrm{LCD}} + 2\lambda \operatorname{sgn}(W_j^{\mathrm{LCD}}) & \text{if } W_j^{\mathrm{LCD}} \neq 0 \\ |\boldsymbol{X}_j^\mathsf{T} \boldsymbol{r}^\lambda| - |\tilde{\boldsymbol{X}}_j^\mathsf{T} \boldsymbol{r}^\lambda| & \text{otherwise.} \end{cases} \tag{8}$$

Applying the Karush–Kuhn–Tucker (KKT) condition, it is easy to verify that $|W_j^{\text{LCD-T}}| > 2\lambda$ if and only if $W_j^{\mathrm{LCD}} \neq 0$.

The sufficiency and antisymmetry conditions can be relaxed slightly. If $n \geqslant 2m + 1$, then W can also take as input the unbiased variance estimator $\tilde{\sigma}^2 = \|\boldsymbol{r}^0\|_2^2 / (n - 2m)$, which is independent of $\boldsymbol{X}_+^\mathsf{T} \boldsymbol{y}$ (Li and Fithian, 2021). This can help us to select λ in (6); we find that $\lambda = 2\tilde{\sigma}$ is a practical choice, where the predictor variables are standardized to have unit norm.

Knockoff methods' FDR control guarantee arises from a crucial stochastic property of the feature statistics: conditional on their absolute values $|\boldsymbol{W}| = (|W_1|, \ldots, |W_m|)$, the signs for null variables are independent Rademacher random variables:

$$\operatorname{sgn}(W_j) \mid |W_j|, \boldsymbol{W}_{-j} \overset{H_j}{\sim} \operatorname{Unif}\{-1, +1\}, \tag{9}$$

where $\boldsymbol{W}_{-j}$ encodes all entries other than W_j. To avoid trivialities, we assume all $W_j \neq 0$ and $W_i \neq W_j$ for any $i \neq j$.

Once the feature statistics are calculated, Selective SeqStep rejects H_j if $W_j \geqslant \widehat{w}$, for an adaptive rejection threshold $\widehat{w} \geqslant 0$ that is based on a running estimator of FDP:

$$\widehat{w} = \min\left\{w \geqslant 0 : \widehat{\mathrm{FDP}}(w) \leqslant \alpha\right\}, \quad \text{for } \widehat{\mathrm{FDP}}(w) = \frac{1 + |\{j : W_j \leqslant -w\}|}{|\{j : W_j \geqslant w\}|}, \tag{10}$$

where $\widehat{w} = \infty$ (no rejections) if $\widehat{\mathrm{FDP}}(w) > \alpha$ for all w. Let $\mathcal{R}^{\mathrm{Kn}} = \{W_j \geqslant \widehat{w}\}$ denote the knockoff rejection set. This rejection rule controls FDR at level α whenever the feature statistics satisfy (9), as Section 3.2 discusses in detail. Appendix A.1 includes a proof of (9) for fixed-X knockoffs.

2.2 Two limitations of knockoffs

Despite its deft exploitation of sparsity, the fixed-X knockoff filter has two major limitations that can inhibit its performance in certain settings. One limitation, reflected in Figure 2, is the so-called *threshold phenomenon*: because the denominator of $\widehat{\mathrm{FDP}}_w$ is the size of the candidate rejection set, we cannot make any rejections at all unless we have $R \geqslant 1/\alpha$. For example, if $\alpha = 0.1$, we must make at least 10 rejections or none at all, even if several p-values lie well below the Bonferroni threshold α/m. Even when the number of potential rejections is above the $1/\alpha$ threshold, the FDP estimator can be highly variable and upwardly biased, adversely affecting the method's power and stability.

Some recent proposals ameliorate this limitation by generating multiple negative controls (Gimenez and Zou, 2019; Emery and Keich, 2019; Nguyen et al., 2020). However, they come at the price of a higher correlation between the original variables and negative controls and a noisier ordering of the variables passed into the Selective Seqstep filter, both of which potentially lead to reduced power (Nguyen et al., 2020). Another proposal by Sarkar and Tang (2021) views fixed-X knockoffs as "splitting" the data into two unbiased estimators of β, one of which has independent coordinates, and applies a hybrid data-splitting method. This proposal can also mitigate the threshold phenomenon, but is often less powerful than knockoffs. By contrast, our calibrated knockoff method is always more powerful than a baseline knockoff method, and we find in simulations that it usually outperforms both multiple knockoffs and the method of Sarkar and Tang (2021) as well.

A second issue is the *whiteout phenomenon* discussed by Li and Fithian (2021), who prove finite-sample bounds on the power of any fixed-X knockoff method in terms of the eigenvalues and eigenvectors of $X^{\mathsf{T}}X$, and the coefficient vector β. When the eigenstructure is unfavorable, we may be forced to make all but a few knockoff variables very highly correlated with their real variable counterparts, and as a result $\operatorname{sgn}(W_j)$ can be very noisy even for strong signal variables, severely biasing $\widehat{\mathrm{FDP}}_w$ upwards. The MCC problem (Example 1.1 with $K = 1$) is a prototypical example exhibiting the whiteout phenomenon, and Li and Fithian (2021) prove that in large MCC problems even the Bonferroni method is dramatically more powerful than the best possible knockoff method. More prosaically, even when the results of Li and Fithian (2021) do not cause catastrophic failure, the knockoff variables still tend to interfere with one another, degrading each other's quality. Multiple knockoff methods tend to exacerbate these problems. As we will see, calibrated knockoffs partially address this issue, delivering high power under some circumstances, but giving limited performance gains in other settings.

Note that these limitations do not conflict with recent theoretical analyses establishing positive results in regimes where the design matrix is well-conditioned and the number of non-nulls diverges; see e.g. Weinstein et al. (2017); Fan et al. (2019); Liu and Rigollet (2019); Wang and Janson (2020).

2.3 Conditional calibration and dBH

Fithian and Lei (2020) introduced a novel technique called *conditional calibration* for proving and achieving FDR control under dependence. They begin by decomposing the FDR into the contributions from each null hypothesis:

$$\mathrm{FDR}(\mathcal{R}) \;=\; \sum_{j\in\mathcal{H}_0} \mathrm{FDR}_j(\mathcal{R}), \quad \text{where} \quad \mathrm{FDR}_j(\mathcal{R}) \;=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\{j\in\mathcal{R}\}}{R\vee 1}\right], \tag{11}$$

and propose controlling each FDR contribution at level α/m, so that $\mathrm{FDR} \leqslant \alpha m_0/m \leqslant \alpha$.

Just as we decomposed the FDR into the contributions from each null hypothesis, we can likewise decompose the FDP as

$$\mathrm{FDP} \;=\; \sum_{j\in\mathcal{H}_0} \mathrm{DP}_j, \quad \text{where} \quad \mathrm{DP}_j(\mathcal{R}) \;=\; \frac{\mathbf{1}\{j\in\mathcal{R}\}}{R\vee 1} \;=\; \frac{\mathbf{1}\{j\in\mathcal{R}\}}{|\mathcal{R}\cup\{j\}|}. \tag{12}$$

We will call DP_j the *realized discovery proportion* for H_j; then $\mathrm{FDR}_j = \mathbb{E}_{H_j}[\mathrm{DP}_j]$. Note that DP_j only contributes to FDP if H_j is true, but it is a well-defined statistic whether H_j is true or false.

To control FDR_j at level α/m, Fithian and Lei (2020) first condition on a sufficient statistic S_j for the submodel described by H_j. By the sufficiency of S_j, the conditional expectation $\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}) \mid S_j]$ can be calculated for any rejection rule $\mathcal{R}$; as a result, a rejection rule with a tuning parameter can also be *calibrated* to control the conditional expectation at α/m.

In particular, the dBH procedure thresholds the BH-adjusted p-value for H_j at an adaptive rejection cutoff $\hat{c}_j^{\mathrm{dBH}}$. When $\hat{c}_j^{\mathrm{dBH}} \geqslant \alpha$, dBH is more liberal than BH (at least concerning H_j), and when $\hat{c}_j^{\mathrm{dBH}} \leqslant \alpha$ it is more conservative.

3 Our method: calibrated knockoffs

3.1 Conditional calibration for knockoffs

Our method is built upon the knockoff procedure. We reject any hypothesis that is rejected by knockoff or by a *fallback test* with a data-adaptive threshold. The FDR is guaranteed to be controlled via conditional calibration.

Formally, our *calibrated knockoff procedure* (cKnockoff) rejects H_j if j is in the index set

$$\mathcal{R}^{\mathrm{cKn}} \;=\; \mathcal{R}^{\mathrm{Kn}} \;\cup\; \{j:\; T_j(\boldsymbol{y}) \geqslant \hat{c}_j(\boldsymbol{y})\}, \tag{13}$$

where $T_j(\boldsymbol{y})$ is some test statistic and $\hat{c}_j(\boldsymbol{y})$ is a data-adaptive threshold which is calibrated to achieve FDR control. For notational convenience, we will suppress the dependence on $\boldsymbol{y}$ when no confusion can arise. We describe T_j and $\hat{c}_j$ in detail next.

In principle, the test statistic T_j can be chosen arbitrarily by the analyst, with larger values representing stronger evidence against the null. To avoid trivialities, we assume that T_j is non-negative and continuously distributed. Our implementation of cKnockoff uses the test statistic

$$T_j = \left|\boldsymbol{X}_j^{\mathsf{T}}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}^{(j)}\right)\right|, \tag{14}$$

where $\hat{\boldsymbol{y}}^{(j)}$ is a vector of fitted values from lasso regression of $\boldsymbol{y}$ on $\boldsymbol{X}_{-j}$ with regularization parameter $\lambda^{(j)}$. Note that if we set $\lambda^{(j)} = 0$, then $\boldsymbol{y} - \hat{\boldsymbol{y}}^{(j)}$ is the vector of OLS residuals under H_j and, holding S_j fixed, T_j is an increasing function of the OLS t-statistic's absolute value. In this sense, (14) generalizes the usual two-tailed t-statistic. Our reason for using $\lambda^{(j)} > 0$ is that, in the sparse setting, the lasso fitted values will likely yield a more accurate adjustment for the effects of the other predictor variables.

As we will see in Section 4.2, it is computationally convenient for $\hat{\boldsymbol{y}}^{(j)}$ to be a function of S_j only. For this reason, we take $\lambda^{(j)} = 2\hat{\sigma}^{(j)}$, using the unbiased estimate of σ^2 under H_j:

$$(\hat{\sigma}^{(j)})^2 = \frac{\|\boldsymbol{y} - \boldsymbol{\Pi}_{-j}\boldsymbol{y}\|^2}{n-m+1}, \quad \text{where} \quad \boldsymbol{\Pi}_{-j} = \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^{\mathsf{T}}\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^{\mathsf{T}}.$$

The rejection threshold $\hat{c}_j$ is calibrated to ensure that the conditional expectation of DP_j of our method is below certain budget. Let $\hat{c}_j$ be the minimal value $c_j \in [0, \infty]$ satisfying

$$\mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant c_j\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right] \leqslant \mathbb{E}_{H_j}[b_j \mid S_j], \tag{15}$$

where b_j, which we call *adaptive budget* of DP_j, is a function of $\boldsymbol{y}$ that satisfies two conditions:

$$\text{(i)}\ \ \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}}) \leqslant b_j \text{ almost surely, and (ii)} \ \ \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\, b_j \,\right] \leqslant \alpha. \tag{16}$$

We could trivially satisfy both conditions by choosing $b_j = \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}})$. But this would not yield any improvement over knockoff since we would be forced to take $\hat{c}_j = \infty$ for all j. For any other choice of budgets satisfying (i), $c_j = \infty$ always satisfies (15), so the calibration problem is always solvable. We defer the construction of the budgets to Section 3.2, giving the explicit formula in (19).

Recall $S_j = (\boldsymbol{X}_{-j}^{\mathsf{T}}\boldsymbol{y}, \|\boldsymbol{y}\|^2)$ is a complete sufficient statistic for the submodel described by H_j. Hence the conditional distribution of y given S_j is fully known under H_j (see Appendix E.2) and the conditional expectations in (15) are computable for any given c_j.

Our method controls FDR in finite samples, as we see next.

Theorem 3.1. *Assume the budgets $b_1, \ldots, b_m$ satisfy the two conditions in* (16)*, and $\hat{c}_j$ are chosen to satisfy* (15)*. Then* $\mathrm{FDR}(\mathcal{R}^{\mathrm{cKn}}) \leqslant \alpha$.

Proof. By construction, $\mathcal{R}^{\mathrm{cKn}} \supseteq \mathcal{R}^{\mathrm{Kn}}$, so

$$\begin{aligned}
\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}^{\mathrm{cKn}}) \mid S_j] \;&=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}}{|\mathcal{R}^{\mathrm{cKn}} \cup \{j\}|} \;\middle|\; S_j\right] \\
&\leqslant\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right] \\
&\leqslant\; \mathbb{E}_{H_j}[b_j \mid S_j].
\end{aligned}$$

Marginalizing over S_j and applying condition (ii), we have

$$\mathrm{FDR}(\mathcal{R}^{\mathrm{cKn}}) \;=\; \sum_{j \in \mathcal{H}_0} \mathbb{E}[\,\mathrm{DP}_j(\mathcal{R}^{\mathrm{cKn}})\,] \;\leqslant\; \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\, b_j \,\right] \;\leqslant\; \alpha.$$

□

To implement cKnockoff efficiently, we will not calculate $\hat{c}_j$ directly. Subtracting the right-hand side of (15) yields an equivalent inequality for the *excess FDR contribution* of variable j, given by

$$E_j(c\,; S_j) \;:=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant c\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;-\; b_j \;\middle|\; S_j\right] \;\leqslant\; 0. \tag{17}$$

Because E_j is a continuous, non-increasing function of c, we have $T_j \geqslant \hat{c}_j$ if and only if $E_j(T_j; S_j) \leqslant 0$. We thus obtain an equivalent, but more computationally useful, definition of calibrated knockoffs as

$$\mathcal{R}^{\mathrm{cKn}} \;=\; \mathcal{R}^{\mathrm{Kn}} \cup \{j :\; E_j(T_j; S_j) \leqslant 0\}.$$

The cKnockoff procedure is adaptive to the choice of knockoff matrix $\tilde{\boldsymbol{X}}$ and the choice of feature statistics, and uniformly improves on any implementation of $\mathcal{R}^{\mathrm{Kn}}$ we might choose. As we will see in Section 3.2, the power is strictly larger than the power of knockoffs when b_j are defined as in (19).

Remark 3.1. *The same calibration scheme can be applied to any baseline FDR-controlling method $\mathcal{R}$ as long as we can find budgets $b_1, \ldots, b_m$ satisfying* (16)*. While we have assumed $\mathrm{DP}_j(\mathcal{R}) \leqslant b_j$, it is enough to have $\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}) - b_j \mid S_j] \leqslant 0$ almost surely.*

3.2 Finding budgets

We now review the proof that knockoff methods control FDR, with a view toward finding slack in the proof that we can use to devise good budgets b_j satisfying the two conditions in (16).

Recall that in knockoffs, we calculate the feature statistic W_j and reject H_j if W_j is above a certain cutoff. Define the *candidate set* for rejection cutoff w as $\mathcal{C}(w) = \{j : \ W_j \geqslant w\}$, and let $\mathcal{A}(w) = \{j : \ W_j \leqslant -w\}$, so that

$$\widehat{\mathrm{FDP}}(w) \ = \ \frac{1 + |\mathcal{A}(w)|}{|\mathcal{C}(w)|}.$$

Let $w_1 < \cdots < w_m$ denote the order statistics of $|W_1|, \ldots, |W_m|$. It suffices to restrict our attention to these order statistics because they are the only values of w where $\mathcal{C}(w)$ or $\widehat{\mathrm{FDP}}(w)$ change. Then we can equivalently write

$$\widehat{w} = w_\tau, \quad \text{where} \quad \tau = \min\Big\{t \in \{1, \ldots, m+1\} : \ \widehat{\mathrm{FDP}}(w_t) \leqslant \alpha\Big\},$$

where we set $w_{m+1} = \infty$ and $\widehat{\mathrm{FDP}}(\infty) = 0$ to cover the case where no rejections are made. In these terms, we can consider knockoffs as a stepwise algorithm with discrete "time" index t, which calculates $\widehat{\mathrm{FDP}}(w_t)$ for each $t = 1, 2, \ldots$, and stops and rejects $\mathcal{C}(w_t)$ the first time $\widehat{\mathrm{FDP}}(w_t) \leqslant \alpha$.

The FDR control proof for the knockoff filter is based on an optional stopping argument. Define

$$M_t \ := \ \frac{|\mathcal{C}(w_t) \cap \mathcal{H}_0|}{1 + |\mathcal{A}(w_t) \cap \mathcal{H}_0|} \ \geqslant \ \frac{|\mathcal{C}(w_t) \cap \mathcal{H}_0|}{1 + |\mathcal{A}(w_t)|} \ = \ \frac{\mathrm{FDP}(\mathcal{C}(w_t))}{\widehat{\mathrm{FDP}}(w_t)},$$

where we take the last expression to be zero by convention if $\mathcal{C}(w_t) = \varnothing$. Barber and Candès (2015) show that M_t is a super-martingale with respect to the discrete-time filtration given by

$$\mathcal{F}_t = \sigma\Big(\, |\boldsymbol{W}|,\ (W_j : j \in \mathcal{H}_0^{\mathsf{c}} \text{ or } |W_j| < w_t)\,,\ |\mathcal{C}(w_t)|\,\Big), \qquad \text{for } t = 1, \ldots, m+1,$$

and they also show that $\mathbb{E}[M_1] \leqslant 1$. We include proofs of both facts in Appendix A.1 for completeness. Because τ is also a stopping time with respect to the same filtration, we have the chain of inequalities

$$\mathrm{FDR}(\mathcal{R}^{\mathrm{Kn}}) \ = \ \mathbb{E}[\mathrm{FDP}(\mathcal{C}(w_\tau))] \ \leqslant \ \alpha\, \mathbb{E}\left[\frac{\mathrm{FDP}(\mathcal{C}(w_\tau))}{\widehat{\mathrm{FDP}}(w_\tau)}\right] \ \leqslant \ \alpha\, \mathbb{E}\left[M_\tau\right] \ \leqslant \ \alpha\, \mathbb{E}\left[M_1\right] \ \leqslant \ \alpha. \tag{18}$$

Because our goal is to find large budgets whose sum is controlled at α in expectation, the intermediate expressions in (18) are natural places to look. Although we cannot calculate αM_τ or αM_1 without knowing $\mathcal{H}_0$, we can decompose the next largest expression to obtain the budgets

$$b_j^0 \ = \ \alpha\, \frac{\mathrm{DP}_j(\mathcal{C}(w_\tau))}{\widehat{\mathrm{FDP}}(w_\tau)} \ = \ \alpha\, \frac{\mathbf{1}\{j \in \mathcal{C}(w_\tau)\}}{1 + |\mathcal{A}(w_\tau)|}, \quad \text{since} \quad \sum_{j \in \mathcal{H}_0} b_j^0 \ = \ \alpha\, \frac{\mathrm{FDP}(\mathcal{C}(w_\tau))}{\widehat{\mathrm{FDP}}(w_\tau)}.$$

These budgets satisfy (i) because $\widehat{\mathrm{FDP}}(w_\tau) \leqslant \alpha$ almost surely, and (ii) by the inequalities in (18).

The budgets b_j^0 do yield a small improvement over baseline knockoffs by taking up the slack in the first inequality of (18), but they do not resolve the main failure modes we discussed in Section 2.2,

where knockoffs usually makes no rejections. In that case, most of the slack is in the second-to-last inequality, since $\mathbb{E}[M_\tau] \approx 0$ while $\mathbb{E}[M_1]$ may be close to 1.

To understand how we might find better budgets, consider the method's behavior, checking if $\widehat{\mathrm{FDP}}(w_t) \leqslant \alpha$ for each $t = 1, 2, \ldots$, in realizations where no rejections are made. Then, as soon as $|\mathcal{C}(w_t)|$ falls below $1/\alpha$, it becomes a foregone conclusion that $\tau = m + 1$ and $M_\tau = 0$, even while the current value of M_t may still be fairly large. In that case, we should stop the algorithm early and harvest as much of M_t as we can. That is, we can obtain larger budgets by replacing τ with another stopping time τ_1 that halts early in hopeless cases:

$$b_j \;=\; \alpha \,\frac{\mathrm{DP}_j(\mathcal{C}(w_{\tau_1}))}{\widehat{\mathrm{FDP}}(w_{\tau_1})} \;=\; \alpha \,\frac{\mathbf{1}\{j \in \mathcal{C}(w_{\tau_1})\}}{1 + |\mathcal{A}(w_{\tau_1})|}, \quad \text{for} \quad \tau_1 = \tau \wedge \min\left\{t : \; |\mathcal{C}(w_t)| < 1/\alpha\right\}. \tag{19}$$

We have $b_j \geqslant b_j^0 \geqslant \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}})$ almost surely because $\tau_1 = \tau$ unless $\mathcal{R}^{\mathrm{Kn}} = \varnothing$. Further, we have

$$\sum_{j \in \mathcal{H}_0} b_j = \frac{\mathrm{FDP}(\mathcal{C}(w_{\tau_1}))}{\widehat{\mathrm{FDP}}(w_{\tau_1})} \leqslant \alpha M_{\tau_1},$$

whose expectation is below α by optional stopping. As a result, the budgets defined in (19) satisfy both conditions in (16).

To illustrate the improvement of b_j over b_j^0, consider the problem setting in Figure 1. Averaging over 100 simulations with $\alpha = 0.05$, we estimate $\sum_{j \in \mathcal{H}_0} \mathbb{E}[b_j] \approx 0.99\alpha$, while $\sum_{j \in \mathcal{H}_0} \mathbb{E}[b_j^0] \approx 0$. While the increase is not always so dramatic, b_j always yields a uniform power improvement, as we see next.

Theorem 3.2. *Assume $\mathcal{H}_0^{\mathrm{c}} \neq \varnothing$. Let the budget be defined as in* (19) *and the nominal FDR level $\alpha \in (0, 0.5]$. Then*

$$\mathrm{TPR}(\mathcal{R}^{\mathrm{cKn}}) > \mathrm{TPR}(\mathcal{R}^{\mathrm{Kn}}).$$

In short, the theorem follows from the fact that our fallback test always makes each hypothesis strictly more likely to be rejected than the knockoffs, due to the construction of b_j in (19). We defer the detailed proof to Appendix E.3. It's worth noticing that although theoretically, the null hypotheses also get more likely to be rejected, the realized FDR is almost the same as knockoffs in our simulation studies in Section 5, even when the power gain is significant. This is because most hypotheses rejected by the fallback test are non-null.

3.3 Refined cKnockoff procedure

The proof of Theorem 3.1 indicates that $\mathcal{R}^{\mathrm{Kn}}$ in the denominator in (15) can be replaced by any $\mathcal{R}^*$ satisfying $\mathcal{R}^{\mathrm{Kn}} \subseteq \mathcal{R}^* \subseteq \mathcal{R}^{\mathrm{cKn}}$ to obtain an even more powerful procedure. In particular, we could use $\mathcal{R}^* = \mathcal{R}^{\mathrm{cKn}}$ and apply the calibration scheme recursively; this would be an example of *recursive refinement* as proposed in Fithian and Lei (2020). However, the computational cost of recursive refinement may be prohibitive since $\mathcal{R}^{\mathrm{cKn}}$, which is already a computationally intensive method, becomes part of the integrand.

A computationally feasible alternative is for $\mathcal{R}^*$ to augment $\mathcal{R}^{\mathrm{Kn}}$ only with a set of very promising variables whose inclusion in $\mathcal{R}^{\mathrm{cKn}}$ can be quickly verified. Informally, we use

$$\mathcal{R}^* \;=\; \mathcal{R}^{\mathrm{Kn}} \cup \{j : \; E_j(T_j; S_j) \leqslant 0, \text{ and } p_j \text{ is tiny}\} \;\subseteq\; \mathcal{R}^{\mathrm{cKn}},$$

where p_j is the p-value from the standard two-sided t-test, a computationally cheap substitute for T_j. We defer our exact formulation of $\mathcal{R}^*$ and additional computational tricks to Appendix D. Using $\mathcal{R}^*$ leads to the *refined calibrated knockoff* (cKnockoff*) procedure rejecting

$$\mathcal{R}^{\mathrm{cKn}*} \;=\; \mathcal{R}^{\mathrm{Kn}} \cup \left\{j : \; T_j \geqslant \hat{c}_j^*\right\} \;=\; \mathcal{R}^{\mathrm{Kn}} \cup \left\{j : \; E_j^*(T_j; S_j) \leqslant 0\right\}, \tag{20}$$

where

$$E_j^*(c\,;S_j) \;=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j\in\mathcal{R}^{\mathrm{Kn}}\right\}\vee\mathbf{1}\left\{T_j\geqslant c\right\}}{|\mathcal{R}^*\cup\{j\}|}-b_j\;\middle|\;S_j\right]\;\leqslant\;E_j(c\,;S_j), \tag{21}$$

and $\hat{c}_j^* = \min\left\{c:\; E_j^*(c\,;S_j)\leqslant 0\right\}\leqslant\hat{c}_j$.

cKnockoff* controls FDR and is uniformly more powerful than cKnockoff, as we show next. However, as a price of handling its additional computational complexity, we will lose the theoretical upper bound of the numerical error in our implementation of cKnockoff*, although simulation studies show the calculation is precise and reliable.

Theorem 3.3. *Assume $\mathcal{R}^{\mathrm{Kn}}\subseteq\mathcal{R}^*\subseteq\mathcal{R}^{\mathrm{cKn}}$, and the budgets $b_1,\ldots,b_m$ satisfy the two conditions in* (16)*. Then $\mathcal{R}^{\mathrm{cKn}*}\supseteq\mathcal{R}^{\mathrm{cKn}}$, and $\mathcal{R}^{\mathrm{cKn}*}$ controls FDR at level α.*

Proof. Because $E_j^*(c\,;S_j)\leqslant E_j(c\,;S_j)$, we have

$$\mathcal{R}^{\mathrm{cKn}*}\supseteq\mathcal{R}^{\mathrm{cKn}}\supseteq\mathcal{R}^*.$$

Recall $\hat{c}_j^* = \min\left\{c:\; E_j^*(c\,;S_j)\leqslant 0\right\}$, so that $E_j^*\leqslant 0$ if and only if $T_j\geqslant\hat{c}_j^*$. Then we have

$$\begin{aligned}
\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}^{\mathrm{cKn}*})\mid S_j] \;&\leqslant\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j\in\mathcal{R}^{\mathrm{cKn}*}\right\}}{|\mathcal{R}^*\cup\{j\}|}\;\middle|\;S_j\right]\\
&=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j\in\mathcal{R}^{\mathrm{Kn}}\right\}\vee\mathbf{1}\left\{T_j\geqslant\hat{c}_j^*\right\}}{|\mathcal{R}^*\cup\{j\}|}\;\middle|\;S_j\right]\\
&\leqslant\; \mathbb{E}_{H_j}[b_j\mid S_j],
\end{aligned}$$

so that $\mathrm{FDR}(\mathcal{R}^{\mathrm{cKn}*})\leqslant\sum_{j\in\mathcal{H}_0}\mathbb{E}[b_j]\leqslant\alpha$.

□

3.4 Robustness to filtering

A nice property of our methods is that they are robust to filtering. Namely, we can filter the fallback test rejections arbitrarily without damaging the FDR control, as stated formally for cKnockoff in Theorem 3.4. This is nontrivial because, in general, the FDR can increase after filtering the rejection set; see e.g. Katsevich et al. (2021).

Theorem 3.4 (Sandwich)**.** *For any rejection rule $\mathcal{R}$ with $\mathcal{R}^{\mathrm{Kn}}\subseteq\mathcal{R}\subseteq\mathcal{R}^{\mathrm{cKn}}$ almost surely, we have* $\mathrm{FDR}(\mathcal{R})\leqslant\alpha$.

Proof. Recall

$$\mathcal{R}^{\mathrm{cKn}} \;=\; \mathcal{R}^{\mathrm{Kn}}\cup\{j:\; T_j\geqslant\hat{c}_j\}.$$

Hence $\mathcal{R}^{\mathrm{Kn}}\subseteq\mathcal{R}\subseteq\mathcal{R}^{\mathrm{cKn}}$ implies

$$\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R})\mid S_j] \;\leqslant\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j\in\mathcal{R}^{\mathrm{Kn}}\right\}\vee\mathbf{1}\left\{T_j\geqslant\hat{c}_j\right\}}{|\mathcal{R}^{\mathrm{Kn}}\cup\{j\}|}\;\middle|\;S_j\right]\;\leqslant\;\mathbb{E}_{H_j}[b_j\mid S_j],$$

so that $\mathrm{FDR}(\mathcal{R})\leqslant\sum_{j\in\mathcal{H}_0}\mathbb{E}[b_j]\leqslant\alpha$. □

The Sandwich property plays a central role in implementing our methods in a fast and reliable way. One important and direct example is filtering. Formally, we reject

$$\mathcal{R}=\mathcal{R}^{\mathrm{Kn}}\cup\{j\in\mathcal{S}:\; E_j(T_j;S_j)\leqslant 0\}\;\subseteq\;\mathcal{R}^{\mathrm{cKn}}$$

for a subset $\mathcal{S}$ of hypotheses for which a simple calculation suggests a high likelihood of rejection by the fallback test. For example, we need not invest computational resources in calculating E_j if $p_j \approx 1$. We defer discussion of our particular choice of $\mathcal{S}$ to Appendix C, where we also prove a generalized version of Theorem 3.4 that also applies to cKnockoff*.

4 Implementation

4.1 Integrating excess FDR

This section discusses implementation details for the core calculation for cKnockoff: evaluating the conditional expectation $E_j(T_j; S_j)$ for each variable.

Let $Q_j(\cdot \mid S_j)$ denote the conditional distribution of the response vector $\boldsymbol{y}$ given $S_j(\boldsymbol{y}) = (\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{y},\ \|\boldsymbol{y}\|_2^2)$. The support of Q_j is the preimage of $S_j(\boldsymbol{y})$, a sphere of dimension $n-m$ embedded in $\mathbb{R}^n$, on which $\boldsymbol{y}$ is conditionally uniform under H_j; see Appendix B. We can write the conditional expectation as

$$E_j(c\,; S_j) \;=\; \mathbb{E}_{H_j}\left[f_j(\boldsymbol{y}; c) \mid S_j\right] \;=\; \int f_j(\boldsymbol{z}; c)\, dQ_j(\boldsymbol{z} \mid S_j), \tag{22}$$

where the integrand is given by

$$f_j(\boldsymbol{z}; c) \;=\; \frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z})\right\} \,\vee\, \mathbf{1}\left\{T_j(\boldsymbol{z}) \geqslant c\right\}}{|\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}) \cup \{j\}|} \;-\; b_j(\boldsymbol{z}), \tag{23}$$

with b_j as defined in (19). Note that the analogous calculation for cKnockoff*, where we calculate $E_j^*(T_j; S_j)$ instead, is exactly the same except with $\mathcal{R}^*$ replacing $\mathcal{R}^{\mathrm{Kn}}$ in the denominator of (23).

Throughout this section we are only concerned with calculating $E_j(T_j(\boldsymbol{y}); S_j(\boldsymbol{y}))$ once the response $\boldsymbol{y}$ has already been observed. As such, we regard $S_j(\boldsymbol{y})$, $c = T_j(\boldsymbol{y})$, and $Q_j(\cdot \mid S_j(\boldsymbol{y}))$ as fixed inputs to the integral $E_j(c\,; S_j(\boldsymbol{y}))$, and suppress their dependence on $\boldsymbol{y}$. To avoid confusion, we use z to denote a generic response vector drawn from the conditional null distribution Q_j. We will use Monte Carlo methods to evaluate the integral (22), using an importance sampling scheme we describe next.

4.2 Conservative importance sampling

While it is easy to sample $\boldsymbol{z} \sim Q_j$, standard Monte Carlo sampling is highly inefficient since we commonly have $f_j(\boldsymbol{z}) = 0$ over most of $\mathrm{Supp}(Q_j)$. Instead, it would be more efficient to restrict our sampling to the region where the integrand is nonzero:

$$\Omega_j(\boldsymbol{y}) \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; f_j(\boldsymbol{z}; c) \neq 0\}. \tag{24}$$

Again, we will suppress the dependence on $\boldsymbol{y}$ when no confusion can arise. The function $f_j(\boldsymbol{z}; c)$ has two terms. The budget $b_j(\boldsymbol{z})$ is zero unless $j \in \mathcal{C}(w_{\tau_1})$, and the other term is zero unless $T_j(\boldsymbol{z}) \geqslant c$ or $j \in \mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}) \subseteq \mathcal{C}(w_{\tau_1})$. Recall the definition of set $\mathcal{C}(w_{\tau_1})$ as a function of $\boldsymbol{z}$ in Section 3.2. As a result, we have $\Omega_j \;=\; \Omega_j^{(1)} \,\cup\, \Omega_j^{(2)}$ with

$$\Omega_j^{(1)} \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; T_j(\boldsymbol{z}) \geqslant c\}, \quad \text{and} \quad \Omega_j^{(2)} \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; j \in \mathcal{C}(w_{\tau_1})\}. \tag{25}$$

For our fallback test statistic defined in (14), $\Omega_j^{(1)}$ amounts to a simple constraint on $\boldsymbol{X}_j^\mathsf{T}\boldsymbol{z}$:

$$T_j(\boldsymbol{z}) \geqslant c \quad\Longleftrightarrow\quad \left|\boldsymbol{X}_j^\mathsf{T}\left(\boldsymbol{z} - \hat{\boldsymbol{y}}^{(j)}(S_j)\right)\right| \geqslant c \quad\Longleftrightarrow\quad \boldsymbol{X}_j^\mathsf{T}\boldsymbol{z} \;\notin\; \left(a_j^{(1)},\, a_j^{(2)}\right). \tag{26}$$

where the bounds of the interval depend only on $\boldsymbol{y}$. Unfortunately, however, $\Omega_j^{(2)}$ admits no such simple description since it is defined implicitly in terms of the feature statistics. Instead, we will use a

conservative importance sampling scheme that approximates $\Omega_j^{(2)}$ by an estimator $\widetilde{\Omega}_j^{(2)}$ and resulting in the approximate integral

$$\widetilde{E}_j(c\,;S_j) \;=\; \int_{\widetilde{\Omega}_j} f_j(\boldsymbol{z};c)\,dQ_j(\boldsymbol{z}\mid S_j), \quad \text{for } \widetilde{\Omega}_j \;=\; \Omega_j^{(1)} \cup \widetilde{\Omega}_j^{(2)}. \tag{27}$$

Proposition 4.1 shows that this approximation does not inflate the FDR, since $\Omega_j^{(1)}$ covers the entire region where $f_j(\boldsymbol{z};c) > 0$.

Proposition 4.1 (Conservative calibration)**.** *Assume that, for each $j = 1,\dots,m$, we calculate $\widetilde{E}_j$ using $\Omega_j^{(1)}$ and a (possibly randomized) estimate $\widetilde{\Omega}_j^{(2)}$. Define the approximate rejection set*

$$\widetilde{\mathcal{R}} \;=\; \mathcal{R}^{\mathrm{Kn}} \,\cup\, \left\{j:\; \widetilde{E}_j(T_j(\boldsymbol{y});S_j(\boldsymbol{y})) \leqslant 0\right\}.$$

Then $\mathrm{FDR}(\widetilde{\mathcal{R}}) \leqslant \alpha$*, and* $\mathcal{R}^{\mathrm{Kn}} \subseteq \widetilde{\mathcal{R}} \subseteq \mathcal{R}^{\mathrm{cKn}}$.

Proof. Fix some j and let $c = T_j(\boldsymbol{y})$. If $T_j(\boldsymbol{z}) < c$, then we have $f_j(\boldsymbol{z};c) = \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z})) - b_j(\boldsymbol{z}) \leqslant 0$ almost surely. As a result,

$$\{\boldsymbol{z} \in \mathrm{Supp}(Q_j):\; f_j(\boldsymbol{z};c) > 0\} \subseteq \Omega_j^{(1)} \subseteq \widetilde{\Omega}_j,$$

so $\widetilde{E}_j$ is a conservative approximation for E_j:

$$E_j(c\,;S_j) - \widetilde{E}_j(c\,;S_j) \;=\; \int_{\mathrm{Supp}(Q_j)\setminus\widetilde{\Omega}_j} f_j(\boldsymbol{z};c)\,dQ_j(\boldsymbol{z}\mid S_j) \;\leqslant\; 0.$$

Since j was arbitrary, this establishes that $\mathcal{R}^{\mathrm{Kn}} \subseteq \widetilde{\mathcal{R}} \subseteq \mathcal{R}^{\mathrm{cKn}}$, hence $\mathrm{FDR}(\widetilde{\mathcal{R}}) \leqslant \alpha$ by Theorem 3.4. □

In practice, we approximate $\Omega_j^{(2)}$ as another constraint on $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z}$ using local linear regression:

$$\widetilde{\Omega}_j^{(2)} \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j):\; \boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z} \in A_j^{(2)}\}$$

for some $A_j^{(2)}$ specified in Appendix B. Our approximation yields the set

$$\widetilde{\Omega}_j \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j):\; \boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z} \in A_j\}, \quad \text{for } A_j = A_j^{(2)} \cup (a_j^{(1)}, a_j^{(2)})^{\mathsf{c}}.$$

As long as A_j is simple enough to quickly evaluate whether $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z} \in A_j$, we can use rejection sampling to rapidly generate a stream of independent samples $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots$ from Q_j conditional on $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z}_k \in A_j$.

After evaluating (23) on each $\boldsymbol{z}_k$, we obtain an independent stream of values $f_j(\boldsymbol{z}_1;c), f_j(\boldsymbol{z}_2;c), \dots$ from the conditional distribution of $f_j(\boldsymbol{z}_k;c)$ given $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z}_k \in A_j$. As a result, $\mathbb{E}[f_j(z_k;c)] = \widetilde{E}_j/Q_j(\widetilde{\Omega}_j)$. We then average them to obtain a Monte Carlo estimate of $\mathbb{E}[f_j(z_k;c)]$ to decide if $\widetilde{E}_j \leqslant 0$.

The naive Monte-Carlo estimation requires a sufficiently large sample size to achieve desired accuracy. In the case where most $f_j(z_k;c)$s are positive with large magnitude, one should be able to declare $\widetilde{E}_j > 0$ with high confidence even with a handful of samples. To be more prudent in sampling, we formulate the problem of deciding if $\widetilde{E}_j \leqslant 0$ into a one-sided hypothesis test that $\mathbb{E}[f_j(z_k;c)] \leqslant 0$ and apply a sequential testing method proposed by Waudby-Smith and Ramdas (2020). We observe empirically that it reduces the Monte-Carlo samples substantially for variables with a sizable $\widetilde{E}_j$.

Appendix B gives further details on the local linear regression algorithm to compute $A_j^{(2)}$, the Monte Carlo sampler for z_k, the sequential testing method, and a theoretical analysis of the FDR accounting for the Monte-Carlo uncertainty.

5 Numerical Studies

In this section we provide selected experiments that compare cKnockoff with competing procedures. Extensions of these simulations under other settings can be found in Appendix F.

5.1 FDR and TPR

We show simulations on the following design matrices $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ with $m = 1000$ and $n = 3000$.

1. **IID normal**: $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.
2. **MCC**: the setting in Example 1.1 with $G = 1000$ and $K = 1$.[1]
3. **MCC-Block**: the setting in Example 1.1 with $G = 5$ and $K = 200$.

The response vector is generated by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{1000}),$$

where $\boldsymbol{\beta}$ has

$$\beta_j = \beta^*, \quad \forall j \in \mathcal{H}_0^{\mathsf{c}}.$$

The signal strength β^* is calibrated such that the BH procedure $\mathcal{R}^{\text{BH}}$ with nominal FDR level $\alpha = 0.2$ will have power TPR $= 0.5$ under the particular design matrix setting. And the alternative hypotheses set $\mathcal{H}_0^{\mathsf{c}}$ is a random subset of $[m]$ that has cardinality 10, uniformly distributed among all such subsets.

The following procedures will be compared in our experiments:

1. **BH**: The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). It has no provable FDR control for all these design matrix settings we consider.
2. **dBH**: The dependence-adjusted Benjamini-Hochberg procedure introduced by Fithian and Lei (2020). We set $\gamma = 0.9$ in the method and do no recursive refinement. It performs similarly to BH but has provable FDR control in this context.
3. **knockoff**: The fixed-X Knockoff method (Barber and Candès, 2015).
4. **BonBH**: The adaptive Bonferroni-BH method (Sarkar and Tang, 2021).
5. **cKnockoff**: Our method as defined in Section 3.1.
6. **cKnockoff***: Our refined method using $\mathcal{R}^*$ as defined in Section 3.3.

For knockoff, cKnockoff, and cKnockoff*, we construct the knockoff matrix via the default semidefinite programming procedure and employ the LCD-T feature statistics (8).

For each trial, we generate $\boldsymbol{X}$ (a realization if it is random), $\boldsymbol{\beta}$, and $\boldsymbol{y}$, and then apply all procedures above. We estimate the FDR and TPR of the results from each procedure by averaging over 400 independent trials. The results are shown in Figure 3.

We observe that cKnockoff controls FDR and dominates knockoff as indicated by our theory. In particular, when knockoff suffers from the threshold phenomenon (small α) or the whiteout phenomenon (MCC problem), cKnockoff and cKnockoff* are able to make as many as or even more correct rejections than BH/dBH in average; when knockoff performs well, cKnockoff/cKnockoff* is even better.

The readers might be puzzled by the non-monotone power curve for cKnockoff in the MCC case. This is mainly driven by the shrinking advantage of cKnockoff over knockoffs as α increases. The

[1] Formally, Example 1.1 would give $n = 3002$. See Appendix E.1 for how we can set $n = 3000$. Similar processing applies to the MCC-Block case.

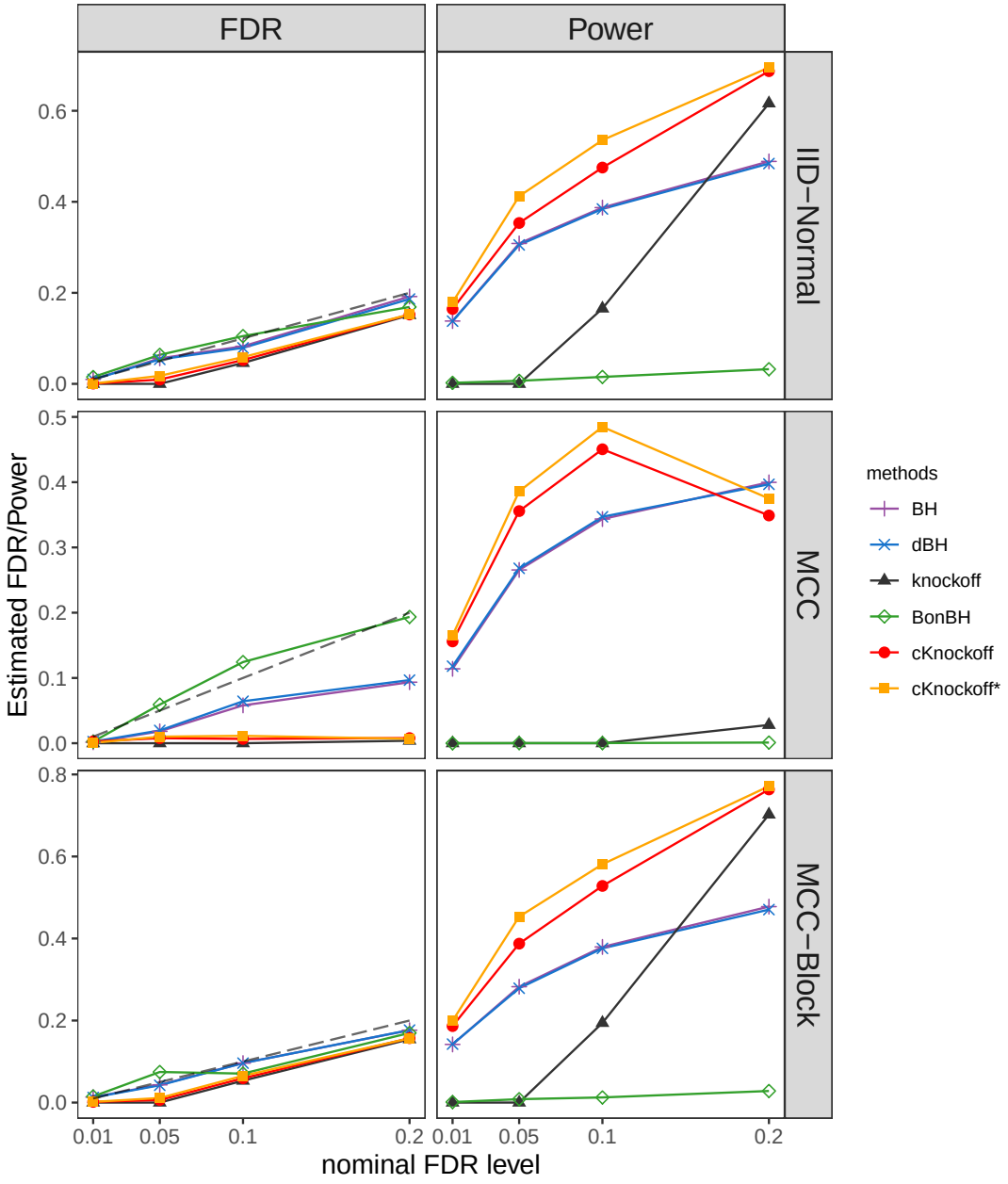

Figure 3: Estimated FDR and TPR under different design matrix settings. cKnockoff/cKnockoff* outperform the other procedures in general.

power gain of cKnockoff over knockoffs is mostly given by the realizations for which $\tau_1 < \tau$ (i.e., $\mathcal{R}^{\mathrm{Kn}} = \varnothing$) and hence b_j is substantially larger than b_j^0. This event happens less likely with a larger α. Furthermore, when the signal-to-noise ratio is large to the extent that the ordering of knockoff statistics is relatively stable, only the top $O(1/\alpha)$ variables could gain an extra budget, thus limiting the power boost. This heuristic analysis also suggests that cKnockoff/cKnockoff* only alleviates the whiteout phenomenon to a limited extent because knockoffs, the baseline procedure that cKnockoff wraps around, suffers even when $m_1 \gg 1/\alpha$. See Appendix F.1.2 for a numerical study. We briefly

discuss alternative strategies to handle whiteout in Section 7.

We do not include multiple knockoffs in the comparison here since it requires a larger aspect ratio than 3. We show the comparison with multiple knockoffs under a different problem setting in Appendix F.1.3. To summarize the results, when $m_1 < 1/\alpha$, multiple knockoffs relieves the threshold phenomenon but underperforms cKnockoff/cKnockoff*; when $m_1 \gg 1/\alpha$, multiple knockoffs is even less powerful than the vanilla knockoffs. Therefore, multiple knockoffs is not as competent as our method in spite of the stronger condition on the sample size.

5.2 Distributions of FDP and TPP

Figure 4 and 5 show the empirical cumulative distribution functions (ECDF) of the FDP and TPP, respectively, of selected procedures, under the same setting as the experiments in Section 5.1.

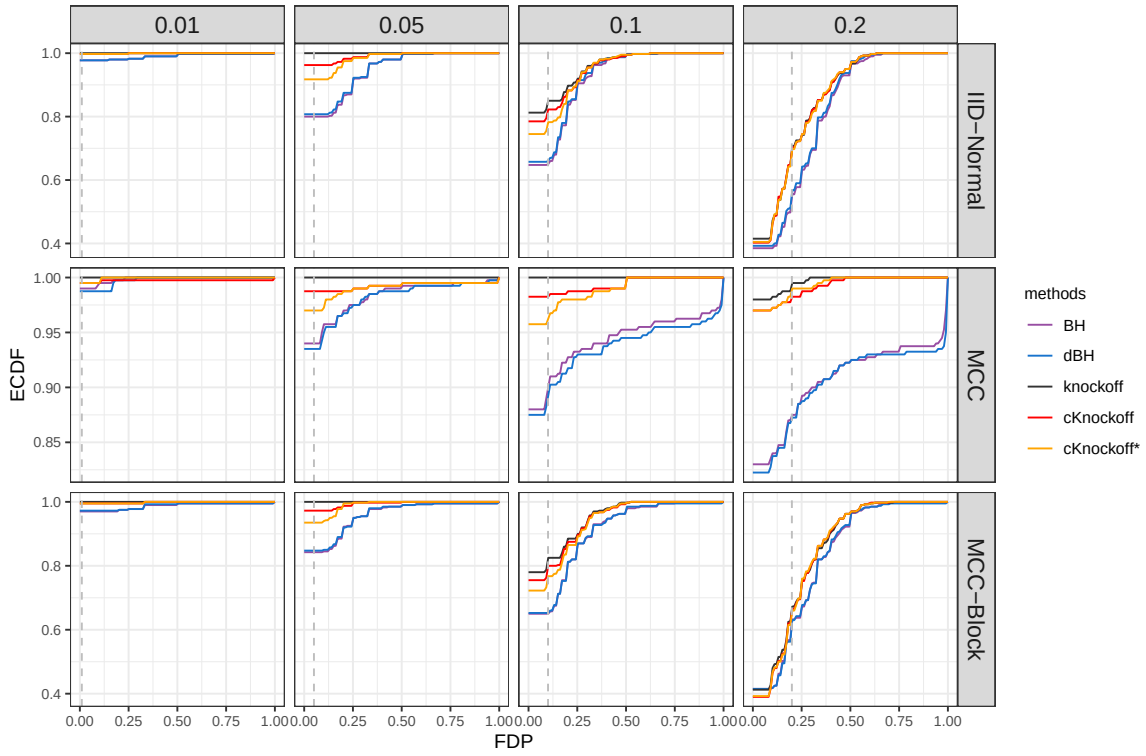

Figure 4: Empirical CDF of FDP from different procedures. Different columns represent different nominal FDR level α, whose value is also indicated by a vertical dashed line.

For FDP, we see that knockoffs, cKnockoff, and cKnockoff* all work well, in the sense that the FDP is smaller than α with high probability in most cases. By contrast, BH and dBH both have stochastically larger FDP in all these cases even when their power is lower. In particular, in the MCC problem, the FDP distribution of BH/dBH puts large mass at both 0 and 1, rendering the rejection sets less reliable.

For TPP, all procedures perform similarly and the TPP is not concentrated at the TPR. Moreover, when knockoffs make some rejections, cKnockoff/cKnockoff* makes a bit more; and when knockoffs fails to reject anything (a flat TPP CDF towards TPP = 0), the CDF of cKnockoff/cKnockoff* keeps its trend. This indicates that our methods fully unleash the potential power of the knockoffs when it suffers.

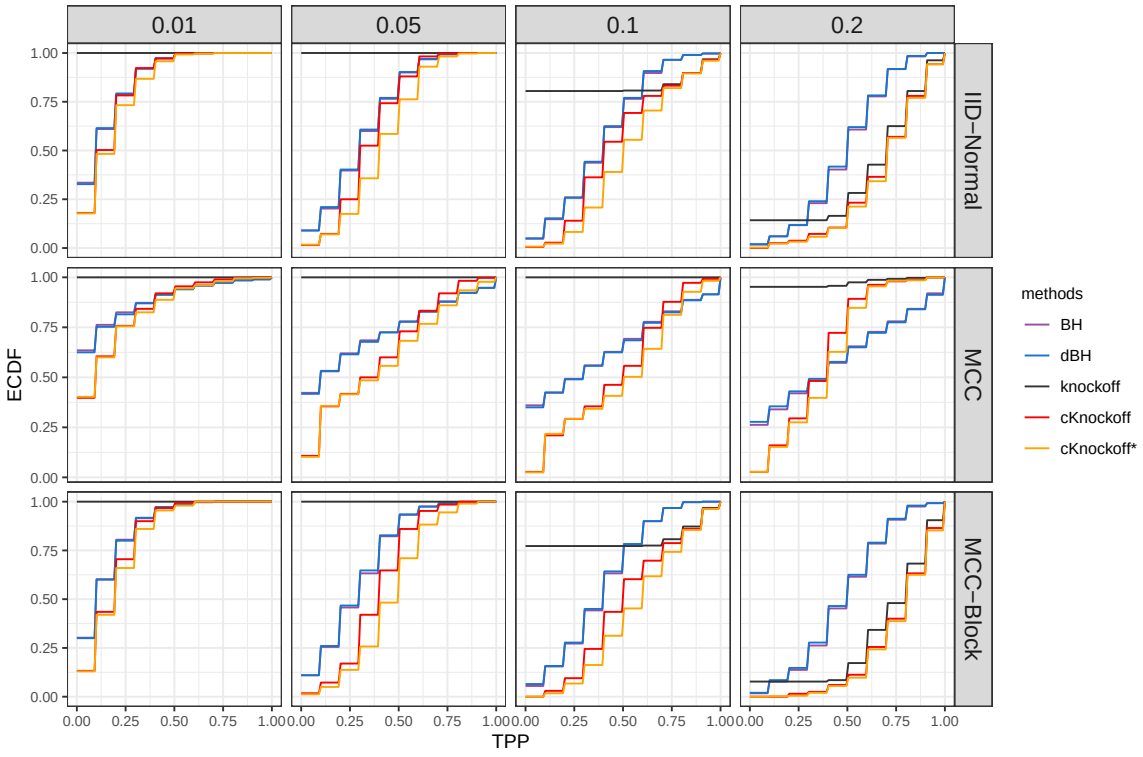

Figure 5: Empirical CDF of TPP from different procedures. Different columns represent different nominal FDR level α.

5.3 Scalability

The computation complexity of cKnockoff/cKnockoff* is highly instance-specific and the worst-case complexity is uninformative. Nevertheless, we can provide a rough analysis of the instance-specific complexity. For every variable that is being examined after the filtering step described in Appendix D, the amount of computation is roughly the same as running a bounded number of rounds of knockoffs with a given knockoff matrix; see Appendix B-D for detail. As a result, if we let A denote the number of variables after filtering, $C_{\mathrm{Kn,f}}$ denote the complexity of knockoffs with a given knockoff matrix, and $C_{\mathrm{Kn,m}}$ denote the complexity of generating a knockoff matrix, the complexity of our methods is

$$O(A \cdot C_{\mathrm{Kn,f}}) + C_{\mathrm{Kn,m}},$$

because the knockoff matrix is only computed once. By contrast, the complexity of knockoffs is $O(C_{\mathrm{Kn,f}}) + C_{\mathrm{Kn,m}}$. Note that when $A = O(1)$, our method has the same complexity as knockoffs up to a multiplicative constant. In many cases, A is small because our method would only examine a handful of promising variables not rejected by knockoffs. Furthermore, we can force A to be small by exploiting a more stringent filtering step.

Figure 6 shows the averaged running time of knockoffs and cKnockoff on a single-core 3.6GHz CPU. In these experiments, we set the signal strength, construct the knockoff matrix, and produce the feature statistics in the same way as in Section 5.1. We set $\alpha = 0.05$ and $n = 3m$ where m varies from 100 to 1000. The left panel of the figure has a fixed number of true alternatives $= 10$ as m increases; while the right panel has a fixed proportion of true alternatives $\pi_1 := m_1/m = 0.1$. As suggested by the heuristic complexity analysis above, the computation time of cKnockoff/cKnockoff* is a small multiple of that of knockoffs in all settings, even with a single core. When multiple cores

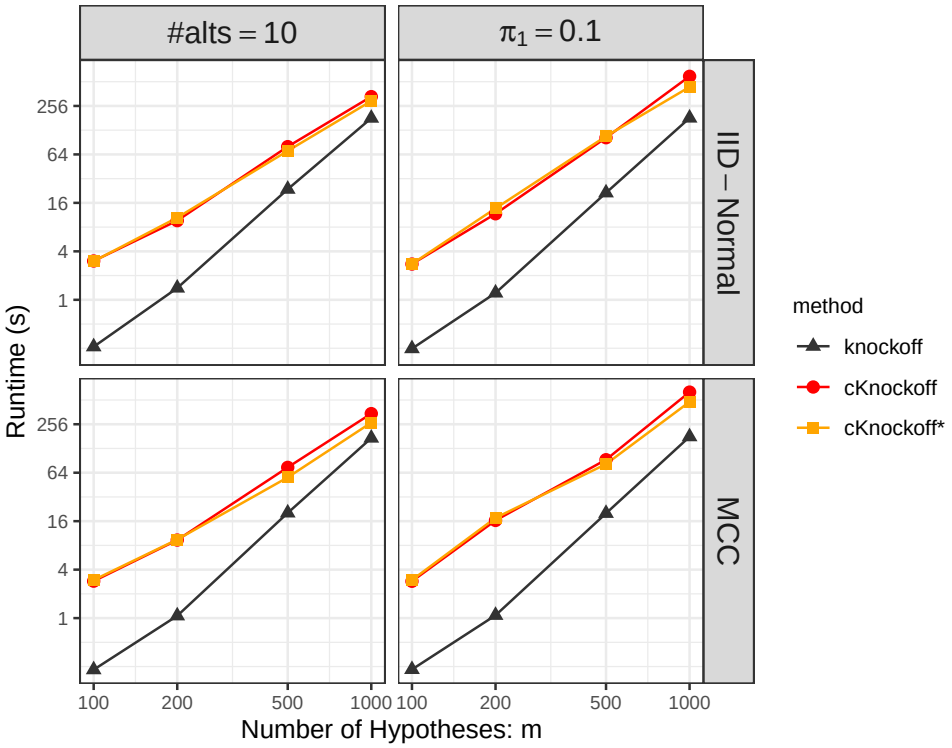

Figure 6: Averaged running time of knockoff, cKnockoff, and cKnockoff* as the problem size increases. The left panel has a fixed number of true alternatives = 10 as m increases; the right panel has a fixed proportion of true alternatives = 0.1. All three methods show a roughly $O(m^3)$ time complexity in the figure.

are available, we can easily parallelize the computation to decide $E_j \leqslant 0$ for different variables or the computation to calculate each E_j; see our `R` package for detail.

Figure 7 demonstrates the scalability in the sample size in an experiment with $\alpha = 0.05$, $m = 100$, and n varies from 300 to 2000. The computation time of our methods is almost flat because cKnockoff/cKnockoff* only depends on the sufficient statistic $(\boldsymbol{X}^\mathsf{T}\boldsymbol{y}, \tilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{y}, \|\boldsymbol{y}\|)$, which is of dimension $2m+1$. Unlike the implementation of knockoffs in `R`, which runs LASSO on the full data $(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{y})$, we implement it to be only based on the lower-dimensional sufficient statistic. Thus, only the computation time of $(\boldsymbol{X}^\mathsf{T}\boldsymbol{y}, \tilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{y})$ grows linearly in n while the majority of computation is independent of n.

6 HIV drug resistance data

In this section we apply cKnockoff and cKnockoff* to detect the mutations in the Human Immunodeficiency Virus (HIV) associated with drug resistance (Rhee et al., 2006), following the analysis in Barber and Candès (2015) and Fithian and Lei (2020).

The dataset include experimental results on 16 different drugs, each falling into one of three different categories: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs). In each experiment, we have access to a set of genetic mutations and a measure of resistance to each drug for a sample of HIV patients. Following Barber and Candès (2015), we construct a design matrix, without an intercept term, by one-hot encoding the mutation so that $X_{ij} = 1$ iff the jth mutation is present in the ith sample

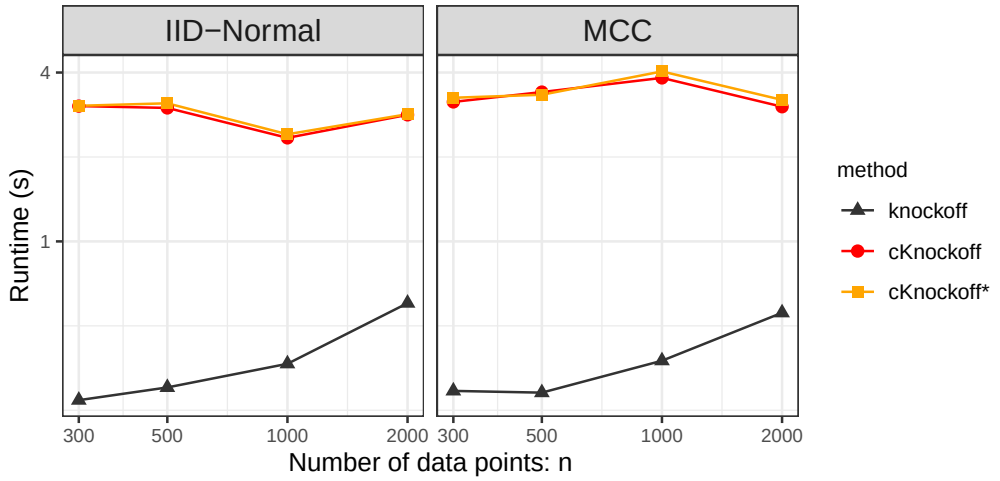

Figure 7: Averaged running time of knockoff, cKnockoff and cKnockoff* as the number of data points n increases. cKnockoff and cKnockoff* scale constantly in n.

and preprocess the data by discarding mutations that occur fewer than three times and removing duplicated columns. Since the ground truth is not available, we evaluate replicability in the same way as Barber and Candès (2015) by comparing the selected mutations to those identified in an independent treatment-selected mutation (TSM) panel of Rhee et al. (2006); see Section 4 of Barber and Candès (2015) for further detail. For each dataset, we will compare BH, knockoffs, cKnockoff, and cKnockoff* described in Section 5.1.

Figure 8 presents the results for $\alpha = 0.05$. Not surprisingly, knockoffs suffers from the threshold phenomenon and makes no rejections for nearly all drugs. Instead, BH makes many rejections but the fraction of rejections that are not replicated in the TSM panel is high for certain drugs. By contrast, cKnockoff and cKnockoff* make a decent number of rejections with a small fraction of non-replicable ones for most drugs.

Figure 9 presents the results for $\alpha = 0.2$. [2] In this case, knockoffs is able to make rejections in half of the problems but is dominated by cKnockoff and cKnockoff*. The other comparisons are qualitatively similar to Figure 8.

7 Discussion

7.1 Summary

We have presented a new approach, the calibrated knockoff procedure, for simultaneously testing if the explanatory variables are relevant to the outcome in the Gaussian linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Our cKnockoff procedure controls FDR and is strictly more powerful than the fixed-X knockoff procedure. And the power gain is especially large when the unknown $\boldsymbol{\beta}$ vector is very sparse, in particular when the number of nonzeros in $\boldsymbol{\beta}$ is not much larger than $1/\alpha$. While our new approach is more computationally intensive in principle, we introduce computational tricks that accelerates the method substantially without sacrificing FDR control in theory. Our implementation of cKnockoff turns out

[2] Readers may have noticed that the rejections made by knockoffs shown here are not exactly the same as the ones shown in Fithian and Lei (2020) or Barber and Candès (2015). This is because that knockoff is implemented as a random method in their R package. In particular, they randomly swap $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ to protect the FDR control from the bias that Lasso, implemented in the glmnet R package, prefers to select a feature with a smaller index. To avoid the interference of such random noise, the results we show are averaged over 20 times applying each procedure.

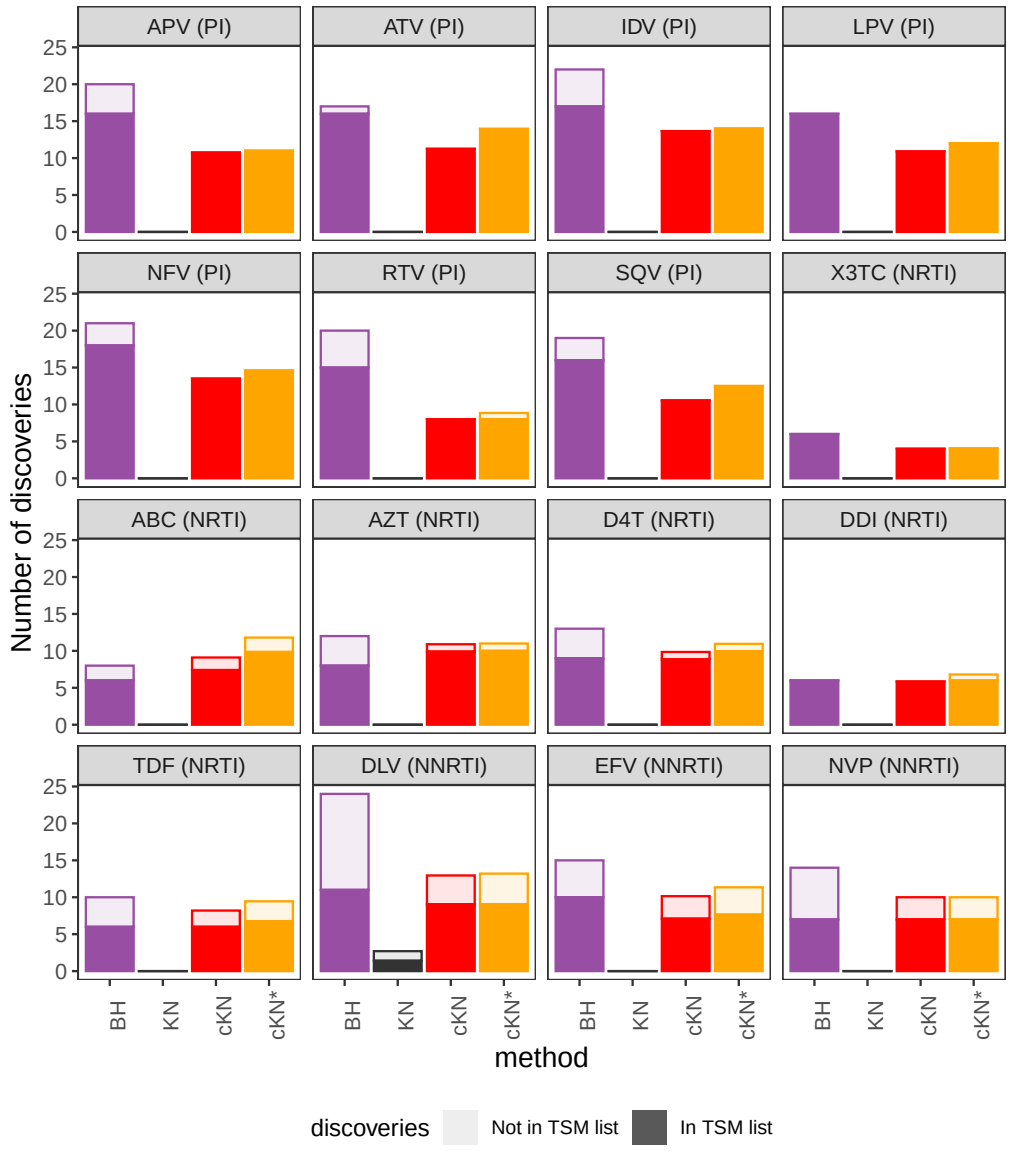

Figure 8: Results on the HIV drug resistance data with $\alpha = 0.05$. The darker segments represent the number of discoveries that were replicated in the TSM panel, while the lighter segments represent the number that were not. Results are shown for the BH, fixed-X knockoffs, cKnockoff and cKnockoff*.

to be quite efficient in our numerical experiments in the sense that the computation time is only a small multiple of that of knockoffs, and it can be further accelerated by parallelization.

7.2 Generalization to model-X knockoffs

Candès et al. (2018) introduced a different version of knockoffs called *model-X knockoffs* under the model-X setting where $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{m+1}$ are i.i.d. with a known distribution of $x_i \in \mathbb{R}^m$ and no assumption on the conditional distribution of y_i given x_i (Candès et al., 2018; Katsevich and Ramdas, 2020; Ren and Candès, 2020; Zhang and Janson, 2020; Li et al., 2021). For $j = 1, \ldots, m$, we

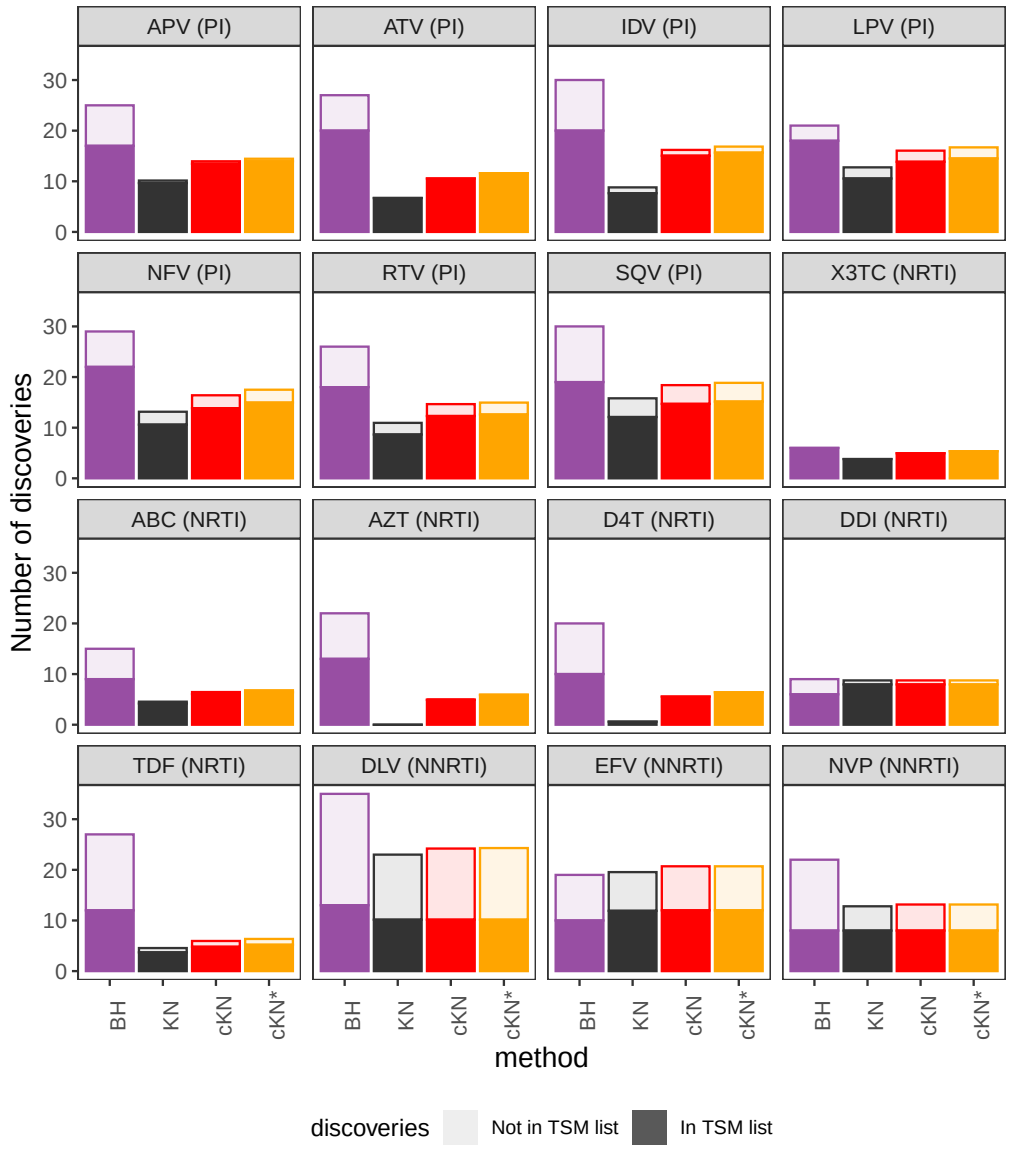

Figure 9: Results on the HIV drug resistance data with $\alpha = 0.2$. The darker segments represent the number of discoveries that were replicated in the TSM panel, while the lighter segments represent the number that were not. Results are shown for the BH, fixed-X knockoffs, cKnockoff and cKnockoff*.

test the null hypothesis H_j that $\boldsymbol{X}_j$ and $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{X}_{-j}$.

Under the model-X setting, $S_j = (\boldsymbol{X}_{-j}, \boldsymbol{y})$ is a sufficient statistic for the null model under H_j. Following the same argument as in Section 3, we can calibrate the model-X knockoffs with any fallback test statistic $T_j(\boldsymbol{X}, \boldsymbol{y})$ by rejecting

$$\mathcal{R}^{\mathrm{cKn}} = \mathcal{R}^{\mathrm{Kn}} \cup \{j : T_j \geqslant \hat{c}_j\} = \mathcal{R}^{\mathrm{Kn}} \cup \{j : E_j(T_j; S_j) \leqslant 0\},$$

where E_j and $\hat{c}_j$ are defined analogous to the fixed-X knockoffs.

Although the generalization from fixed-X to model-X knockoffs is straightforward, efficient implementation is nontrivial. For example, it is unclear which fallback test statistic would be powerful.

While we can continue to use the same one as in calibrated fixed-X knockoffs, there are other potentially powerful alternatives such as the p-value from the conditional randomization test (Candès et al., 2018). Moreover, many computational tricks we developed for calibrated fixed-X knockoffs exploit the rotational invariance of Gaussian errors, which is no longer available under a general model-X setting. We leave these problems for future work.

7.3 Remedies for whiteout

As discussed earlier, cKnockoff only alleviates the whiteout issue (Li and Fithian, 2021) to a limited extent because the signs of knockoff feature statistics are too noisy to be useful for the Selective-Seqstep filter even though the ordering of variables is satisfactory. Meanwhile, BH and dBH are not ideal either because they have bimodal FDP distributions with large masses at around 0 and 1. On the other hand, Example 1.1 indicates that the high correlation could help, instead of hurt, inference largely in the presence of sparsity. It would be interesting to investigate the possibility to take advantage of the sparsity which can help inform the ordering of variables without relying making the sparsity assumption explicitly and resorting to asymptotics.

Reproducibility

Calibrated knockoffs are implemented in an `R` package available online at the Github repository `https://github.com/yixiangLuo/cknockoff`. And the `R` code to reproduce all simulations and figures in this paper can be found at `https://github.com/yixiangLuo/cknockoff_expr`.

Acknowledgments

William Fithian and Yixiang Luo were partially supported by National Science Foundation grant DMS-1916220, and a Hellman Fellowship from UC Berkeley.

References

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

Yoav Benjamini and Yosef Hochberg. 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, 1995.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

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

Kristen Emery and Uri Keich. Controlling the fdr in variable selection via multiple knockoffs. *arXiv preprint arXiv:1911.09442*, 2019.

Yingying Fan, Emre Demirkaya, Gaorong Li, and Jinchi Lv. Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 2019.

William Fithian and Lihua Lei. Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438*, 2020.

Jaime Roquero Gimenez and James Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2184–2192. PMLR, 2019.

Eugene Katsevich and Aaditya Ramdas. A theoretical treatment of conditional independence testing under model-X. *arXiv preprint arXiv:2005.05506*, 2020.

Eugene Katsevich, Chiara Sabatti, and Marina Bogomolov. Filtering the rejection set while preserving false discovery rate control. *Journal of the American Statistical Association*, pages 1–12, 2021.

Shuangning Li, Matteo Sesia, Yaniv Romano, Emmanuel Candès, and Chiara Sabatti. Searching for consistent associations with a multi-environment knockoff filter. *arXiv preprint arXiv:2106.04118*, 2021.

Xiao Li and William Fithian. Whiteout: when do fixed-x knockoffs fail? *arXiv preprint arXiv:2107.06388*, 2021.

Jingbo Liu and Philippe Rigollet. Power analysis of knockoff filters for correlated designs. *arXiv preprint arXiv:1910.12428*, 2019.

Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, and Sylvain Arlot. Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR, 2020.

Zhimei Ren and Emmanuel Candès. Knockoffs with side information. *arXiv preprint arXiv:2001.07835*, 2020.

Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L Brutlag, and Robert W Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

Sanat K Sarkar and Cheng Yong Tang. Adjusting the benjamini-hochberg method for controlling the false discovery rate in knockoff assisted variable selection. *arXiv preprint arXiv:2102.09080*, 2021.

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

Wenshuo Wang and Lucas Janson. A power analysis of the conditional randomization test and knockoffs. *arXiv preprint arXiv:2010.02304*, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

Asaf Weinstein, Rina Barber, and Emmanuel Candès. A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*, 2017.

Lu Zhang and Lucas Janson. Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*, 2020.

A Knockoff details

A.1 Deferred proofs

Here, for the sake of completeness, we prove several results that appeared first in Barber and Candès (2015) and other works.

Theorem A.1. *Suppose the feature statistics $\boldsymbol{W} = (W_1, W_2, \ldots, W_m)$ satisfy the sufficiency condition and the anti-symmetry condition. Then*

$$\operatorname{sgn}(W_j) \mid |W_j|, \boldsymbol{W}_{-j} \overset{H_j}{\sim} \operatorname{Unif}\{-1, +1\}$$

for any $j \in \mathcal{H}_0$.

Proof. It suffices to show for a given arbitrary $j \in [m]$,

$$\boldsymbol{W} \overset{d}{=} (W_1, \ldots, W_{j-1}, -W_j, W_{j+1}, \ldots, W_m),$$

where we denote the right-hand-side vector as $\boldsymbol{W}^-$ for simplicity.

The sufficiency condition allow us to write

$$\boldsymbol{W} = g(\boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}_+, \boldsymbol{X}_+^\mathsf{T}\boldsymbol{y})$$

for some function g. And the anti-symmetry condition gives

$$\boldsymbol{W}^- = g((\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}_+^{\text{swap}}, (\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{y}),$$

where $\boldsymbol{X}^{\text{swap}}$ is modified from matrix $\boldsymbol{X}_+$ by swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$, i.e.

$$(\boldsymbol{X}_+^{\text{swap}})_j = (\boldsymbol{X}_+)_{j+m}, \quad (\boldsymbol{X}_+^{\text{swap}})_{j+m} = (\boldsymbol{X}_+)_j, \quad (\boldsymbol{X}_+^{\text{swap}})_i = (\boldsymbol{X}_+)_i \text{ for } i \neq j,\, j+m.$$

Notice

$$(\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}_+^{\text{swap}} = \boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}_+, \quad (\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}\beta = \boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}$$

due to $\boldsymbol{X}_j^\mathsf{T}\boldsymbol{X}_i = \tilde{\boldsymbol{X}}_j^\mathsf{T}\boldsymbol{X}_i$ for all $i \neq j$ and $\beta_j = 0$. We further have

$$(\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{y} \overset{d}{=} \mathcal{N}((\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}_+^{\text{swap}}) \overset{d}{=} \mathcal{N}(\boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}_+) \overset{d}{=} \boldsymbol{X}_+^\mathsf{T}\boldsymbol{y}.$$

Therefore

$$\boldsymbol{W}^- = g((\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{X}_+^{\text{swap}}, (\boldsymbol{X}_+^{\text{swap}})^\mathsf{T}\boldsymbol{y}) \overset{d}{=} g(\boldsymbol{X}_+^\mathsf{T}\boldsymbol{X}_+, \boldsymbol{X}_+^\mathsf{T}\boldsymbol{y}) = \boldsymbol{W}.$$

□

Theorem A.2. *Suppose the feature statistics satisfy the sufficiency condition and the anti-symmetry condition. Then*

$$M_t = \frac{|\mathcal{C}(w_t) \cap \mathcal{H}_0|}{1 + |\mathcal{A}(w_t) \cap \mathcal{H}_0|}$$

is a supermartingale with respect to the filtration

$$\mathcal{F}_t = \sigma\Big(|W|,\ (W_j : j \in \mathcal{H}_0^{\mathsf{c}} \text{ or } |W_j| < w_t),\ |\mathcal{C}(w_t)|\Big).$$

Moreover, we have $\mathbb{E}M_1 \leqslant 1$.

Proof. Without loss of generality, assume $|W_1| < \cdots < |W_m|$ for easier notations. So $w_t = W_t$ and $|W_j| < w_t$ is equivalent to $j < t$. Let $V_t^+ \coloneqq |\mathcal{C}(w_t) \cap \mathcal{H}_0|$ and $V_t^- \coloneqq |\mathcal{A}(w_t) \cap \mathcal{H}_0|$ for short. Since the non-null feature statistics are known given $\mathcal{F}_t$, it's easy to see V_t^+ and V_t^- are measurable with respect to $\mathcal{F}_t$. Hence $M_t \in \mathcal{F}_t$.

It remains to show $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] \leqslant M_t$. By construction, $M_{t+1} = M_t$ if $t \in \mathcal{H}_0^{\mathsf{c}}$. Otherwise

$$M_{t+1} = \frac{V_t^+ - \mathbf{1}\{W_t > 0\}}{1 + V_t^- - (1 - \mathbf{1}\{W_t > 0\})} = \frac{V_t^+ - \mathbf{1}\{W_t > 0\}}{(V_t^- + \mathbf{1}\{W_t > 0\}) \vee 1}.$$

Theorem A.1 implies that

$$\operatorname{sgn}(W_j) \mid |\boldsymbol{W}|, (W_i : i \in \mathcal{H}_0^{\mathsf{c}} \text{ or } i < t) \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1, +1\}, \quad \text{for } j \in \mathcal{H}_0 \text{ and } j \geqslant t.$$

Hence

$$\mathbb{P}(W_t > 0 \mid \mathcal{F}_t) = \frac{V_t^+}{V_t^+ + V_t^-}.$$

Therefore

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] \;=\; \frac{1}{V_t^+ + V_t^-}\left[V_t^+ \frac{V_t^+ - 1}{V_t^- + 1} + V_t^- \frac{V_t^+}{V_t^- \vee 1}\right] \;=\; \begin{cases} \frac{V_t^+}{1+V_t^-} = M_t, & V_t^- > 0 \\ V_t^+ - 1 = M_t - 1, & V_t^- = 0 \end{cases}.$$

So M_t is s supermartingale.

To show $\mathbb{E}M_1 \leqslant 1$, note $V_1^+ \mid |\boldsymbol{W}| \sim \text{Binomial}(m_0, 1/2)$. We have

$$\begin{aligned} \mathbb{E}\left[M_1 \mid |\boldsymbol{W}|\right] \;&=\; \mathbb{E}\left[\frac{V_1^+}{1 + m_0 - V_1^+} \mid |\boldsymbol{W}|\right] \;=\; \sum_{i=1}^{m_0} \mathbb{P}(V_1^+ = i) \cdot \frac{i}{1 + m_0 - i} \\ &=\; \sum_{i=1}^{m_0} \frac{1}{2^{m_0}} \frac{m_0!}{i!(m_0 - i)!} \cdot \frac{i}{1 + m_0 - i} \;=\; \sum_{i=1}^{m_0} \frac{1}{2^{m_0}} \frac{m_0!}{(i-1)!(m_0 - i + 1)!} \\ &=\; \sum_{i=1}^{m_0} \mathbb{P}(V_1^+ = i - 1) \leqslant 1. \end{aligned}$$

Then $\mathbb{E}M_1 = \mathbb{E}\left[\mathbb{E}\left[M_1 \mid |\boldsymbol{W}|\right]\right] \leqslant 1$.

□

A.2 A faster version of the LSM statistic

The LSM statistic defined in (5) is computationally burdensome because it requires calculating the entire lasso path. Even if the great majority of variables are null, they will eventually enter and leave the model in a chaotic process once λ becomes small enough. If we stop too early, most variables will never enter, so their feature statistics will be zero and there will be no chance to discover them, but if we stop too late, we will consume most of our computational resources fitting null variables at the end of the path. In practice, the path is also calculated only for a fine grid of λ values, which has the added undesirable effect of introducing artificial ties between variables.

This section introduces a more computationally efficient alternative that uses a coarser grid of λ values and also stops the path early, but breaks ties by using variables' correlations using the residuals at each stage. Assume we calculate the lasso fit $\hat{\boldsymbol{\beta}}^{\lambda(\ell)}$ for $\ell = 0, \ldots, L$ on a coarse grid defined by:

$$\max_{1 \leqslant j \leqslant 2m} \lambda_j^* = \lambda_{\max} = \lambda(0) > \lambda(1) > \cdots > \lambda(L) = \lambda_{\min} = 2\tilde{\sigma} \wedge \frac{\lambda_{\max}}{2}.$$

We stop the path at $2\tilde{\sigma}$ because we find it tends to set most null variables' coefficients to zero, we take $\lambda(0), \ldots, \lambda(L)$ to be a decreasing geometric sequence:

$$\lambda(\ell) = \lambda(0) \cdot \zeta^\ell, \quad \text{for } \zeta = \left(\frac{\lambda(L)}{\lambda(0)}\right)^{1/L}.$$

Then for variable $j = 1, \ldots, 2m$, define its (discrete) time of entry as

$$\ell_j^* = \min\left\{\ell \in \{1, \ldots, L\} : \; \hat{\beta}_j^{\lambda(\ell)} \neq 0\right\},$$

with $\ell_j^* = L+1$ if the set is empty. To break ties between these discrete values, we use each variable's correlation with the lasso residuals at $\lambda_j^+ = \lambda(\ell_j^* - 1)$, the last fit in the discrete path before variable j enters:

$$\rho_j = \left|(\boldsymbol{X}_+)_j^\mathsf{T} \boldsymbol{r}^{\lambda_j^+}\right| \leqslant \lambda_j^+, \quad \text{where } \boldsymbol{r}^\lambda = \boldsymbol{y} - \boldsymbol{X}_+ \hat{\boldsymbol{\beta}}^\lambda.$$

ρ_j can be considered an estimate of λ_j^*, since we have

$$\lambda_j^* = \left|(\boldsymbol{X}_+)_j^\mathsf{T} r^{\lambda_j^*}\right| \approx \left|(\boldsymbol{X}_+)_j^\mathsf{T} r^{\lambda_j^+}\right|.$$

To combine these, we can use any transform that orders variables first by λ_j^+ and then by ρ_j. We use a transform that also aids the accuracy of the local linear regression approximation of $W_j(\boldsymbol{z})$ on $\boldsymbol{X}_j^\mathsf{T}\boldsymbol{z}$. Let ι denote a small positive quantity and $\lambda(L+1) = 0$, and define

$$\hat{\lambda}_j \;=\; \max\{\rho_j - \iota,\, \lambda(\ell_j^*)\} + \iota\rho_j/\lambda_j^+ \;\approx\; \max\{\rho_j,\, \lambda(\ell_j^*)\}.$$

Finally, we define the *coarse LSM* (C-LSM) feature statistics by substituting $\hat{\lambda}_j$ for λ_j^* in (5):

$$W_j^{\text{C-LSM}} = (\hat{\lambda}_j \vee \hat{\lambda}_{j+m}) \cdot \operatorname{sgn}(\hat{\lambda}_j - \hat{\lambda}_{j+m}).$$

A.3 numerical comparisons

Here we compare the performance of the vanilla LSM feature statistics with our LCD-T and C-LSM. The settings are the same as in Section 5.1. Figure 10 shows the results. We see all three feature statistics perform equally well.

B Implementation: calculations to check if $\tilde{E}_j \leqslant 0$

In this section we continue Section 4, regarding $S_j(\boldsymbol{y})$, $c = T_j(\boldsymbol{y})$, and $Q_j(\cdot \mid S_j(\boldsymbol{y}))$ as fixed and use $\boldsymbol{z}$ to denote a generic response vector drawn from the conditional null distribution Q_j. We will explain the sampling scheme, the construction of $\widetilde{\Omega}_j$, and the way to control numerical error when checking if $\tilde{E}_j \leqslant 0$.

B.1 The sampling scheme

We first fix a basis to make our calculations convenient. Let $\boldsymbol{V}_{-j} \in \mathbb{R}^{n\times(m-1)}$ denote an orthonormal basis for the column span of $\boldsymbol{X}_{-j}$, so that $\boldsymbol{\Pi}_{-j} = \boldsymbol{V}_{-j}\boldsymbol{V}_{-j}^\mathsf{T}$. Next, for the projection of $\boldsymbol{X}_j$ orthogonal to the span of $\boldsymbol{X}_{-j}$, define the unit vector in that direction:

$$\boldsymbol{v}_j = \frac{\boldsymbol{\Pi}_{-j}^\perp \boldsymbol{X}_j}{\|\boldsymbol{\Pi}_{-j}^\perp \boldsymbol{X}_j\|}, \quad \text{where } \boldsymbol{\Pi}_{-j}^\perp = \boldsymbol{I} - \boldsymbol{\Pi}_{-j}$$

Finally, let $\boldsymbol{V}_{\text{res}} \in \mathbb{R}^{n\times(n-m)}$ denote an orthonormal basis for the subspace orthogonal to the span of $\boldsymbol{X}$. Then we can decompose $\boldsymbol{z}$ as

$$\boldsymbol{z} = \boldsymbol{V}_{-j}\boldsymbol{V}_{-j}^\mathsf{T}\boldsymbol{z} + \boldsymbol{v}_j \cdot \eta + \boldsymbol{V}_{\text{res}} \cdot \boldsymbol{r},$$

where $\eta = \boldsymbol{v}_j^\mathsf{T}\boldsymbol{z} \in \mathbb{R}$ is the component of $\boldsymbol{z}$ in the direction $\boldsymbol{v}_j$, and $\boldsymbol{r} = \boldsymbol{V}_{\text{res}}^\mathsf{T}\boldsymbol{z} \in \mathbb{R}^{n-m}$ is the residual component.

Recall that Q_j is uniform on its support $\{\boldsymbol{z} :\ S_j(\boldsymbol{z}) = S_j(\boldsymbol{y})\}$. Fixing $S_j(\boldsymbol{z}) = S_j(\boldsymbol{y})$ is equivalent to constraining

$$\boldsymbol{V}_{-j}^\mathsf{T}\boldsymbol{z} = \boldsymbol{V}_{-j}^\mathsf{T}\boldsymbol{y}, \quad \text{and } \ \eta^2 + \|\boldsymbol{r}\|^2 = \|\boldsymbol{\Pi}_{-j}^\perp \boldsymbol{z}\|^2 = \|\boldsymbol{\Pi}_{-j}^\perp \boldsymbol{y}\|^2.$$

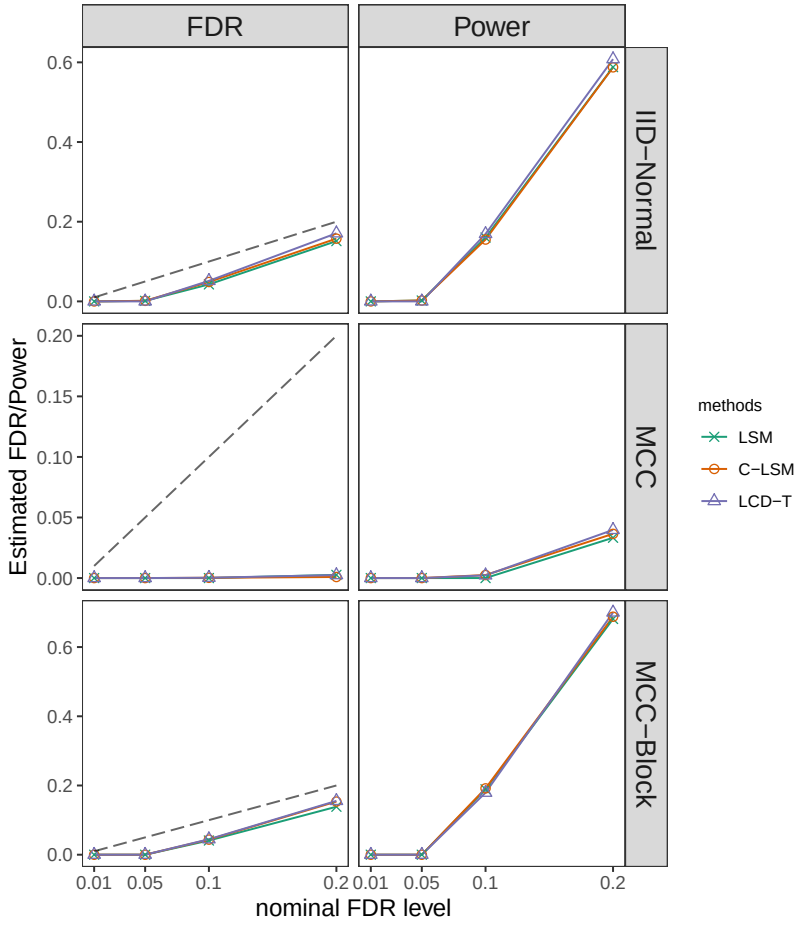

Figure 10: Estimated FDR and TPR of knockoffs employing LSM, LCD-T or C-LSM as feature statistics.

Let $\rho^2 = \|\mathbf{\Pi}_{-j}^{\perp}\boldsymbol{y}\|^2$, which depends only on $S_j(\boldsymbol{y})$. We can sample $\boldsymbol{z} \sim Q_j$ by first sampling

$$(\eta, \boldsymbol{r}) \sim \text{Unif}\left(\rho \cdot \mathbb{S}^{n-m}\right),$$

where $\mathbb{S}^{n-m} \subseteq \mathbb{R}^{n-m+1}$ is the unit sphere of dimension $n-m$, and then reconstructing $\boldsymbol{z}$ using equation (29).

To sample from $\widetilde{\Omega}_j$, we add a further constraint on η, that it lies in some union of intervals $U_j \subseteq [-\rho, \rho]$. See Appendix B.2 for details. In order to sample $\boldsymbol{z}$ satisfying this constraint efficiently, we first sample η marginally obeying the constraint and then $\boldsymbol{r}$ conditional on η. Standard calculations show the cumulative distribution function (CDF) of η is

$$F_\eta(x; S_j) := \mathbb{P}(\eta \leqslant x) = F_{t_{n-m}}\left(\frac{x\sqrt{n-m}}{\sqrt{(\rho^2 - x^2)}}\right) \quad \text{for } |x| < \rho, \tag{28}$$

where $F_{t_{n-m}}$ is the CDF of the t-distribution with degree of freedom $n-m$. Now write $\boldsymbol{r} = \|\boldsymbol{r}\| \cdot \boldsymbol{u}$ with $\boldsymbol{u}$ being the direction of $\boldsymbol{r}$. Given η, we have $\|\boldsymbol{r}\| = \sqrt{\rho^2 - \eta^2}$ and $\boldsymbol{u} \sim \text{Unif}(\mathbb{S}^{n-m-1})$ is independent of η.

To conclude, we sample $\eta \sim F_\eta$ in the desired union of intervals and $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-m-1})$ independently. The $\boldsymbol{z}$ is given by

$$\boldsymbol{z}(\eta, \boldsymbol{u}) = \boldsymbol{V}_{-j}\boldsymbol{V}_{-j}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{v}_j \cdot \eta + \sqrt{\rho^2 - \eta^2} \cdot \boldsymbol{V}_{\mathrm{res}}\boldsymbol{u}. \tag{29}$$

B.2 The construction of $\widetilde{\Omega}_j$

Section 4 describes $\widetilde{\Omega}_j$ in terms of $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z}$. To be consistent with the sampling scheme, we first rephrase $\widetilde{\Omega}_j$ in terms of η as used in (29).

We decompose

$$\boldsymbol{X}_j^T\boldsymbol{z} = \boldsymbol{X}_j^T\boldsymbol{\Pi}_{-j}\boldsymbol{z} + (\boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}_j) \cdot \eta, \tag{30}$$

which establishes a linear relationship equation $\boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{z}$ and η (recall $\boldsymbol{\Pi}_{-j}\boldsymbol{z} = \boldsymbol{\Pi}_{-j}\boldsymbol{y}$ is fixed). Thus we can write

$$\widetilde{\Omega}_j \;=\; \{\boldsymbol{z}(\eta, \boldsymbol{u}) \in \mathrm{Supp}(Q_j) :\; \eta \in A_j\}, \qquad \text{for } A_j = A_j^{(2)} \cup (a_j^{(1)}, a_j^{(2)})^{\mathsf{c}},$$

where $a_j^{(1)}$ and $a_j^{(2)}$ are solved from

$$\Omega_j^{(1)}(\boldsymbol{y}) \;=\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; T_j(\boldsymbol{z}) \geqslant c\} \;=\; \{\boldsymbol{z}(\eta, \boldsymbol{u}) \in \mathrm{Supp}(Q_j) :\; \eta \in (a_j^{(1)}, a_j^{(2)})^{\mathsf{c}}\} \tag{31}$$

and $A_j^{(2)}$ is a union of intervals in order to have

$$\Omega_j^{(2)}(\boldsymbol{y}) \approx \{\boldsymbol{z}(\eta, \boldsymbol{u}) \in \mathrm{Supp}(Q_j) :\; \eta \in A_j^{(2)}\}.$$

Note we reuse the notation A_j, $A_j^{(2)}$, $a_j^{(1)}$, and $a_j^{(2)}$ as the constraint sets or boundaries for η and they shouldn't be confused with those in Section 4.

In the rest of this section, we give explicit way to obtain $a_j^{(1)}$, $a_j^{(2)}$, and $A_j^{(2)}$.

For $a_j^{(1)}$ and $a_j^{(2)}$, note

$$\Omega_j^{(1)}(\boldsymbol{y}) \;=\; \left\{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; \left|\boldsymbol{X}_j^{\mathsf{T}}\left(\boldsymbol{z} - \hat{\boldsymbol{y}}^{(j)}(S_j)\right)\right| \geqslant c\right\}.$$

Using (30), direct calculation shows

$$a_j^{(1)} = \frac{\hat{\boldsymbol{y}}^{(j)}(S_j) - \boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{\Pi}_{-j}\boldsymbol{y} - c}{\boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}_j}, \quad a_j^{(2)} = \frac{\hat{\boldsymbol{y}}^{(j)}(S_j) - \boldsymbol{X}_j^{\mathsf{T}}\boldsymbol{\Pi}_{-j}\boldsymbol{y} + c}{\boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}_j}.$$

For $A_j^{(2)}$, note

$$tOmega_j^{(2)} = \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; j \in \mathcal{C}(w_{\tau_1})\} \;\subseteq\; \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; |W_j| \geqslant w_{\tau_1}\}.$$

Hence we over-estimate $\Omega_j^{(2)}$ by $\{\boldsymbol{z} \in \mathrm{Supp}(Q_j) :\; |W_j| \geqslant w_{\tau_1}\}$, which is approximately identified by local linear regression.

Specifically, we treat $\boldsymbol{z} = \boldsymbol{z}(\eta, \boldsymbol{u})$ as a one-dimensional random function of η with $\boldsymbol{u}$ being an independent random noise. Our local linear regression scheme then regresses $|W_j|$ and w_{τ_1} on η, and solve for the region where $|W_j| \geqslant w_{\tau_1}$ numerically.

The effectiveness of this method is based on the following observations we see in simulation studies.

1. Given a fixed $\boldsymbol{u}$, for the lasso-based *LSM* (*C-LSM*) and *LCD* (*LCD-T*) feature statistics we consider, $|W_j(\boldsymbol{z})|$ is roughly a piecewise linear function of η;
2. $|W_j| \approx \mathbb{E}\left[|W_j| \mid \eta\right]$ if $\mathbb{E}\left[|W_j| \mid \eta\right]$ is large. Specifically, the standard deviation of $|W_j|$ conditional on η is typically at most 10% of the conditional mean when $\mathbb{E}\left[|W_j| \mid \eta\right] \geqslant \mathbb{E}\left[w_{\tau_1} \mid \eta\right]$.

Heuristically, the first observation is because that the KKT conditions of Lasso is a piecewise linear system. And the second is that the randomness in $\boldsymbol{u}$ contributes to the correlation between $\boldsymbol{y}$ and the knockoff variables, which is considered noise. When $|W_j|$ is large, such a noise is expected to be dominated by the signal from the original variable. w_{τ_1} shares a similar behavior, though its value is determined by a more complicated mechanism.

These observations then allow us to approximate

$$\{\boldsymbol{z} \in \mathrm{Supp}(Q_j) : \ |W_j| \geqslant w_{\tau_1}\} \approx \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) : \ \mathbb{E}\left[|W_j| \mid \eta\right] \geqslant \mathbb{E}\left[w_{\tau_1} \mid \eta\right]\},$$

with the conditional expectation estimated by the local linear regression.

To be specific, the construction of $\widetilde{\Omega}_j^{(2)}$ is done as follows.

1. sample $\boldsymbol{z}_1(\eta_1, u_1), \ldots, \boldsymbol{z}_k(\eta_k, u_k)$ such that $\eta_1, \ldots, \eta_k$ are equi-spaced nodes in $[-\rho, \rho]$.
2. compute $|W_j(\boldsymbol{z}_i)|$ and $w_{\tau_1}(\boldsymbol{z}_i)$ for each i.
3. estimate $\mathbb{E}[|W_j(\boldsymbol{z})| \mid \eta]$, denoted as $\widehat{W}_j(\eta)$, by a local linear regression on $|W_j(\boldsymbol{z}_i)|$ for $i = 1, \ldots, k$. We use the Gaussian kernel and set the bandwidth as the distance between consecutive nodes $\eta_i - \eta_{i-1}$. Similarly, estimate $\mathbb{E}[w_{\tau_1}(\boldsymbol{z}) \mid \eta]$ as $\widehat{w}(\eta)$.
4. $\widetilde{\Omega}_j^{(2)}$ is then

$$\widetilde{\Omega}_j^{(2)} \ = \ \{\boldsymbol{z} \in \mathrm{Supp}(Q_j) : \ \eta \in A_j^{(2)}\}, \quad A_j^{(2)} := \left\{\eta : \widehat{W}_j(\eta) \geqslant \widehat{w}(\eta)\right\},$$

   where the inequality $\widehat{W}_j(\eta) \geqslant \widehat{w}(\eta)$ is solved numerically.

B.3 Numerical error control

With enough computational budget, we can compute $\widetilde{E}_j$ at arbitrary precision. While in practice, we should tolerate some level of numerical error introduced by Monte-Carlo. The key idea for such error control is to upper bound the probability of mistakenly claiming the sign of $\widetilde{E}_j$ at some α_c.

Our process of deciding $\mathrm{sgn}(\widetilde{E}_j)$ can be formalized as constructing a confidence interval for $\widetilde{E}_j$ from a sequence of i.i.d. samples $f_j(\boldsymbol{z}_1)$, $f_j(\boldsymbol{z}_2)$, $\ldots$, $f_j(\boldsymbol{z}_k)$, with a common mean $\widetilde{E}_j/Q_j(\widetilde{\Omega}_j)$. For each $k = 1, 2, \ldots$, such a confidence interval at level α_c, $C_k(f_j(\boldsymbol{z}_1), \ldots, f_j(\boldsymbol{z}_k); \alpha_c)$, is computed and once we see it excludes 0, we stop the calculation and decide if $\widetilde{E}_j \leqslant 0$ accordingly.

To control the probability of deciding $\mathrm{sgn}(\widetilde{E}_j)$ wrongly, it suffices to have the sequence of confidence intervals C_k hold for all (infinite many) k simultaneously,

$$\mathbb{P}(\exists k : \ \widetilde{E}_j \notin C_k(f_j(\boldsymbol{z}_1), \ldots, f_j(\boldsymbol{z}_k)) \leqslant \alpha_c.$$

We can further control the inflation of the FDR due to the Monte-Carlo error, as shown next. This bound is rather loose and we found the inflation is ignorable in practice.

Proposition B.1. *Let $\alpha_c(\boldsymbol{y}) = \mathcal{R}^{\mathrm{Kn}(\alpha)}(\boldsymbol{y}) \cdot \alpha_0/m$ for some constant α_0 and reject*

$$\mathcal{R} = \mathcal{R}^{\mathrm{Kn}(\alpha)} \cup \{j : \ \exists k, \ s.t. \ x \leqslant 0 \text{ for all } x \in C_k(f_j(\boldsymbol{z}_1), \ldots, f_j(\boldsymbol{z}_k); \alpha_c)\}.$$

Note $\mathcal{R}$ is the cKnockoff rejection set if we compute E_j and claim its sign with confidence as described above. Then

$$\mathrm{FDR}(\mathcal{R}) \leqslant \alpha + \alpha_0.$$

Proof. Denote event $D_j \coloneqq \{\exists k, \text{ s.t. } x \leqslant 0 \text{ for all } x \in C_k(f_j(\boldsymbol{z}_1), \ldots, f_j(\boldsymbol{z}_k); \alpha_c)\}$ and

$$D_j^{\text{in}} \coloneqq D \cap \left\{\forall k,\ \widetilde{E}_j \in C_k\right\}, \quad D_j^{\text{out}} \coloneqq D \cap \left\{\exists k,\ \widetilde{E}_j \notin C_k\right\}.$$

Recall D_j^{in} implies $T_j \geqslant \hat{c}_j$. And we have $\mathbb{P}(D_j^{\text{out}} \mid \boldsymbol{y}) \leqslant \alpha_c$ by the construction of C_k. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}) \mid S_j] \ &= \ \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\text{Kn}}\right\} \vee \mathbf{1}\left\{D_j\right\}}{|\mathcal{R} \cup \{j\}|} \,\middle|\, S_j\right] \\
&\leqslant \ \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\text{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\} \vee \mathbf{1}\left\{D_j^{\text{out}}\right\}}{|\mathcal{R}^{\text{Kn}} \cup \{j\}|} \,\middle|\, S_j\right] \\
&\leqslant \ \mathbb{E}_{H_j}[b_j \mid S_j] + \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{D_j^{\text{out}}\right\}}{|\mathcal{R}^{\text{Kn}} \cup \{j\}|} \,\middle|\, S_j\right].
\end{aligned}$$

Marginalizing over S_j, we have

$$\mathrm{FDR}(\mathcal{R}) = \sum_{j \in \mathcal{H}_0} \mathbb{E}[\,\mathrm{DP}_j(\mathcal{R})\,] \leqslant \sum_{j \in \mathcal{H}_0} \mathbb{E}[b_j] + \sum_{j \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{D_j^{\text{out}}\right\}}{|\mathcal{R}^{\text{Kn}}(\boldsymbol{y}) \cup \{j\}|} \,\middle|\, \boldsymbol{y}\right]\right] \leqslant \alpha + \alpha_0.$$

□

Remark B.1. *The denominator m in our choice of α_c can be replaced by the number of hypotheses that survived after filtering (see Appendix C), which would increase α_c and save some computation time while keeping the same FDR control.*

Remark B.2. *In our implementation, we truncate the Monte-Carlo sample sequence and use their sample mean to decide if $\widetilde{E}_j \leqslant 0$ once we already have 500 samples but still $0 \in C_k$ for all $k = 1, \ldots, 500$.*

In particular, since the values of $f_j(z_k)$ are bounded, we apply Theorem 3 in Waudby-Smith and Ramdas (2020) to construct such a confidence sequence C_k, which is adaptive to the sample variance and achieves state-of-the-art performance for bounded random variables.

C Implementation: filtering the rejection set beforehand

As suggested in section 3.4, we reject the *filtered cKnockoff rejection set*

$$\mathcal{R} = \mathcal{R}^{\text{Kn}} \cup \{j \in \mathcal{S} :\ E_j(T_j; S_j) \leqslant 0\} \ \subseteq \ \mathcal{R}^{\text{cKn}}$$

for some set $\mathcal{S}$ that only contains the hypotheses likely to be non-null. In practice, we find it is a good choice to set

$$\mathcal{S}(s; \boldsymbol{y}) = (\mathcal{S}^{\text{BH}}(s; \boldsymbol{y}) \cap \mathcal{S}^{\text{p}}(s; \boldsymbol{y})) \cup \mathcal{S}^{\text{Kn}}(s; \boldsymbol{y}) \backslash \mathcal{R}^{\text{Kn}}(\boldsymbol{y}), \tag{32}$$

where

$$\mathcal{S}^{\text{BH}} = \mathcal{R}^{\text{BH}(s \cdot 4\alpha)}, \quad \mathcal{S}^{\text{p}} = \{j :\ p_j \leqslant s \cdot \alpha/2\}, \quad \mathcal{S}^{\text{Kn}} = \left\{j :\ |W_j| \geqslant w_{m - |\mathcal{S}^{\text{BH}} \cap \mathcal{S}^{\text{p}}|} \wedge w_\tau\right\}$$

with $s = 1$. It almost doesn't exclude any rejections in the vanilla cKnockoff among our simulations and achieves $|S| \ll m$. That is to say, it preserves the power of cKnockoff when accelerating it a lot by filtering out many non-promising hypotheses.

Moreover, such filtering can be done in an online manner. Let

$$s_j^+(\boldsymbol{y}) = \min\{s:\ j \in \mathcal{S}(s;\boldsymbol{y})\}. \tag{33}$$

Then s_j^+, which we call *promising score*, roughly measures how likely H_j will be rejected. Then we can check $E_j(T_j;S_j) \leqslant 0$ sequentially on the hypotheses ranked by their promising scores. Theorem 3.4 allows us to stop at any time we like and report the rejection set with a valid FDR control. This feature is available in our `R` package but not employed in the numerical experiments shown in this paper.

To apply filtering to cKnockoff*, we reject the *filtered cKnockoff* rejection set*

$$\mathcal{R} = \mathcal{R}^{\mathrm{Kn}} \cup \{j \in \mathcal{S}:\ E_j^*(T_j;S_j) \leqslant 0\} \ \subseteq\ \mathcal{R}^{\mathrm{cKn}*}$$

where $\mathcal{S}$ is further required to satisfy $\mathcal{R}^{\mathrm{Kn}} \cup \mathcal{S} \supseteq \mathcal{R}^*$. Like in cKnockoff, Theorem C.1 shows such filtering doesn't hurt the FDR control.

Theorem C.1. *For any rejection rule $\mathcal{R}$ with $\mathcal{R}^* \subseteq \mathcal{R} \subseteq \mathcal{R}^{\mathrm{cKn}*}$ almost surely, we have* $\mathrm{FDR}(\mathcal{R}) \leqslant \alpha$.

Proof. Recall

$$\mathcal{R}^{\mathrm{cKn}*} \ = \ \mathcal{R}^{\mathrm{Kn}} \cup \{j:\ T_j \geqslant \hat{c}_j^*\}.$$

Hence $\mathcal{R}^* \subseteq \mathcal{R} \subseteq \mathcal{R}^{\mathrm{cKn}*}$ implies

$$\mathbb{E}_{H_j}[\mathrm{DP}_j(\mathcal{R}) \mid S_j] \ \leqslant\ \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j^*\right\}}{|\mathcal{R}^* \cup \{j\}|}\ \Big|\ S_j\right] \ \leqslant\ \mathbb{E}_{H_j}[b_j \mid S_j],$$

so that $\mathrm{FDR}(\mathcal{R}) \leqslant \sum_{j\in\mathcal{H}_0}\mathbb{E}[b_j] \leqslant \alpha$. □

For cKnockoff*, $\mathcal{S}$ is set again by (32). The condition $\mathcal{R}^{\mathrm{Kn}} \cup \mathcal{S} \supseteq \mathcal{R}^*$ is made true by the construction of $\mathcal{R}^*$ in Appendix D.

D Implementation: efficient cKnockoff*

The core in implementing cKnockoff* is the calculation related to $\mathcal{R}^*$. In this section, we introduce how we construct, compute, and apply $\mathcal{R}^*$ in cKnockoff*.

D.1 Construct $\mathcal{R}^*$

To make $\mathcal{R}^*$ easy to compute while bringing in additional power, we need a delicate construction. Let

$$\mathcal{R}^* = \mathcal{R}^{\mathrm{Kn}} \cup \{j \in \mathcal{S}^*:\ E_j(T_j;S_j) \leqslant 0\},$$

with

$$\mathcal{S}^* = \left\{j \in \mathcal{S}:\ p_j \leqslant \frac{\alpha}{m} \wedge \frac{0.01\alpha}{\lceil 1/\alpha - 1\rceil},\ s_j^+ \text{ has rank no larger than } K^{\mathrm{cand}}\right\},$$

where $\mathcal{S}$ is the filtering set in (32), s_j^+ is the promising score in (33), and K^{cand} is a pre-specified constant to restrict the size of $\mathcal{S}^*$. In words, we construct $\mathcal{S}^*$ by picking the K^{cand}-most promising hypotheses in $\mathcal{S}$ who have p-values below certain threshold. We will explain the rationale behind this p-value cutoff in the next two sections. By construction, we have $\mathcal{R}^{\mathrm{Kn}} \subseteq \mathcal{R}^* \subseteq \mathcal{R}^{\mathrm{cKn}}$ and $\mathcal{R}^{\mathrm{Kn}} \cup \mathcal{S} \supseteq \mathcal{R}^*$.

D.2 Compute $\mathcal{R}^*$

Computing $\mathcal{R}^*$ shares the same goal as computing $\mathcal{R}^{\mathrm{cKn}}$. That is, we want to calculate

$$E_j(c; S_j) \;=\; \int f_j(\boldsymbol{z}; c)\, dQ_j(\boldsymbol{z})$$

with $c = T_j(\boldsymbol{y})$. Let's adopt the same narrative as in Section 4, regarding $S_j(\boldsymbol{y})$, $c = T_j(\boldsymbol{y})$, and $Q_j(\cdot \mid S_j(\boldsymbol{y}))$ as fixed and use $\boldsymbol{z}$ to denote a generic response vector drawn from the conditional null distribution Q_j. But this time, we will use numerical integration instead of Monte-Carlo to compute E_j approximately.

D.2.1 Formulation of the numerical scheme

To formulate the calculation as a numerical integration, recall (29) that $\boldsymbol{z}$ can be written as

$$\boldsymbol{z} = \boldsymbol{z}(\eta, \boldsymbol{u}) = \boldsymbol{V}_{-j}\boldsymbol{V}_{-j}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{v}_j \cdot \eta + \sqrt{\rho^2 - \eta^2} \cdot \boldsymbol{V}_{\mathrm{res}}\boldsymbol{u}.$$

given S_j, where the two random variables $\eta = \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{z}$ and $\boldsymbol{u} = \boldsymbol{V}_{\mathrm{res}}^{\mathsf{T}}\boldsymbol{z} / \left\|\boldsymbol{V}_{\mathrm{res}}^{\mathsf{T}}\boldsymbol{z}\right\|$ are independent.

Following the idea in Appendix B.2, we treat $f_j(\boldsymbol{z}(\eta, \boldsymbol{u}))$ as a random function of η with an independent random noise $\boldsymbol{u}$. Then

$$E_j \;=\; \int \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-m-1})}\left[f_j(\boldsymbol{z}(\eta, \boldsymbol{u}))\right]\, d\, F_\eta(\eta),$$

where $F_\eta(\cdot; S_j)$ is the CDF of η given in (28).

This motivates computing E_j approximately by

$$E_j \approx \sum_{i=1}^{1/h} f_j(\boldsymbol{z}(\eta_i, \boldsymbol{u}_i)) \cdot h,$$

where

$$\eta_i = F_\eta^{-1}(ih), \quad \boldsymbol{u}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{n-m-1})$$

In words, we numerically integrate over the variable η and use Monte-Carlo for the leftover noise variable $\boldsymbol{u}$. When h is tiny, we have fine grid for the numerical integration and many samples for Monte-Carlo and hence the calculation would be accurate.

D.2.2 Check if $E_j \leqslant 0$ efficiently

Since $f_j \leqslant 1$, we have

$$\int_{\Omega_j^{(1)}} f_j\, dQ_j \;\leqslant\; Q_j(\Omega_j^{(1)}) \coloneqq B_j^+.$$

The bound B_j^+ is easy to compute, referring to Appendix B.2. Therefore, to conclude $E_j \leqslant 0$, it suffices to show

$$B_j^+ + \int_{\left(\Omega_j^{(1)}\right)^{\mathsf{c}}} f_j\, dQ_j \leqslant 0.$$

We use our numerical scheme derived in the previous section to compute the integral. Recall (31),

$$\left(\Omega_j^{(1)}\right)^{\mathsf{c}} = \left\{\boldsymbol{z}(\eta, \boldsymbol{u}) : \eta \in \left(a_j^{(1)}, a_j^{(2)}\right)\right\}.$$

It's easy to see that $\boldsymbol{v}_j^\mathsf{T}\boldsymbol{y}$ is equal to either $a_j^{(1)}$ or $a_j^{(2)}$ since we have $T(\boldsymbol{z}) = T(\boldsymbol{y})$ on the boundary of $\Omega_j^{(1)}$. Assuming $\boldsymbol{v}_j^\mathsf{T}\boldsymbol{y} = a_j^{(1)}$ without loss of generality, we have

$$\int_{\left(\Omega_j^{(1)}\right)^\mathsf{c}} f_j \, dQ_j \approx \sum_{i=1}^{(1-B_j^+)/h} f_j(\boldsymbol{z}(\eta_i, \boldsymbol{u}_i)) \cdot h,$$

where

$$\eta_i = F_\eta^{-1}(F_\eta(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{y}) + ih), \quad \boldsymbol{u}_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{n-m-1}). \tag{34}$$

The key of our fast algorithm is to recall that $f_j(\boldsymbol{z}(\eta, \boldsymbol{u})) \leqslant 0$ on $\left(\Omega_j^{(1)}\right)^\mathsf{c}$. That is to say, $f_j(\boldsymbol{z}(\eta_i, \boldsymbol{u}_i)) \leqslant 0$ for all η_i and $\boldsymbol{u}_i$ considered in (34). Therefore, once we see

$$\sum_{i=1}^{k} f_j(\boldsymbol{z}(\eta_i, \boldsymbol{u}_i)) \cdot h \leqslant -B_j^+$$

for some $k \leqslant (1 - B_j^+)/h$, we can claim $E_j \leqslant 0$ without further calculation. Setting the starting point of the numerical integration nodes as $\eta_0 = \boldsymbol{v}_j^\mathsf{T}\boldsymbol{y}$ rather than the other endpoint $a_j^{(2)}$ makes f_j at $\boldsymbol{z}(\eta_1, \boldsymbol{u}_1)$, $\boldsymbol{z}(\eta_2, \boldsymbol{u}_2), \ldots$ more likely to be nonzero, hence earlier to conclude $E_j \leqslant 0$, when H_j is nonnull.

In particular, we try $k = 1, 2, \ldots, K^{\text{step}}$ sequentially, once we see $\sum_{i=1}^k f_j(\boldsymbol{z}(\eta_i, u_i)) \cdot h \leqslant -B_j^+$, we add j into $\mathcal{R}^*$ and stop the calculation; otherwise we still stop but don't include j.

The overall running time of $\mathcal{R}^*$ is upper bounded by $K^{\text{cand}} \cdot K^{\text{step}}$ times the time of evaluating the knockoff feature statistics. In our implementation, we set $K^{\text{cand}} = K^{\text{step}} = 3$.

D.2.3 Choose a proper h

Now the only question that remains is how to choose a proper h. The choice of h is driven by a tradeoff — a smaller h yields a more accurate estimate of E_j and a larger h allows us to determine if $E_j \leqslant 0$ within a few steps.

There are only two cases for $\boldsymbol{z} \in \left(\Omega_j^{(1)}\right)^\mathsf{c}$ where $f_j(\boldsymbol{z}) \neq 0$: when $j \in \mathcal{R}^{\text{Kn}}(\boldsymbol{z})$ or $j \in \mathcal{C}(w_{\tau_1})$. As we will see, the computational tricks introduced in Appendix D.3 implies the first case is uncommon when we need to compute $\mathcal{R}^*$. So considering $j \in \mathcal{C}(w_{\tau_1})$ but $j \notin \mathcal{R}^{\text{Kn}}(\boldsymbol{z})$, we have

$$f_j = -b_j(\alpha; \boldsymbol{z}) \approx -\frac{\alpha}{\lceil 1/\alpha - 1 \rceil}$$

since $|\mathcal{C}(w_{\tau_1})| = \lceil 1/\alpha - 1 \rceil$ and $\widehat{\text{FDP}}(w_{\tau_1}) \lessapprox 1$. Therefore, if we set

$$h = B_j^+ / \frac{\alpha}{\lceil 1/\alpha - 1 \rceil},$$

we expect to see $\sum_{i=1}^k f_j(\boldsymbol{z}(\eta_i, \boldsymbol{u}_i)) \cdot h \leqslant -B_j^+$ with $k = 1$ or 2, if j is to be rejected.

Moreover, by noticing a monotone bijection between $\eta = \boldsymbol{v}_j^\mathsf{T}\boldsymbol{z}$ and the t-statistic t_j,

$$t_j = \frac{\boldsymbol{v}_j^\mathsf{T}\boldsymbol{z}}{\sqrt{(\|\boldsymbol{y}\|^2 - \|\boldsymbol{\Pi}_{-j}\boldsymbol{y}\|^2 - (\boldsymbol{v}_j^\mathsf{T}\boldsymbol{z})^2)/(n-m)}}, \quad \boldsymbol{v}_j^\mathsf{T}\boldsymbol{z} = t_j \cdot \sqrt{\frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{\Pi}_{-j}\boldsymbol{y}\|^2}{t_j^2 + n - m}},$$

simple calculation shows

$$Q_j\left(\left\{\boldsymbol{v}_j^\mathsf{T}\boldsymbol{z} \notin \left(a_j^{(1)}, a_j^{(2)}\right)\right\}\right) = B_j^+ \ \leqslant \ p_j = 2F_{t_{n-m}}(-|t_j|).$$

Recall we have

$$p_j \leqslant \frac{\alpha}{m} \wedge \frac{0.01\alpha}{\lceil 1/\alpha - 1 \rceil}, \quad \forall j \in \mathcal{S}^*.$$

This guarantees

$$h = B_j^+ / \frac{\alpha}{\lceil 1/\alpha - 1 \rceil} \leqslant 0.01,$$

which is reasonably small.

It is worth pointing out here that with calculating $\mathcal{R}^*$ in this way, we don't have an estimation of the error in computing E_j, unlike in cKnockoff where we have a confidence sequence for this purpose. Although our current implementation of cKnockoff* works very well in simulations, analysts can choose a smaller h or just stick to cKnockoff if they don't want to rely on a numerical integration without an error estimation.

D.3 Computational tricks in applying $\mathcal{R}^*$

Now we have an algorithm to compute $\mathcal{R}^*$ efficiently. When computing $\mathcal{R}^{\mathrm{cKn}^*}$, we need to calculate $E_j^*(T_j; S_j)$. This is done in the same way as calculating $E_j(T_j; S_j)$ in cKnockoff, but replacing

$$f_j(\boldsymbol{z}_i) \;\coloneqq\; \frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i)\right\} \;\vee\; \mathbf{1}\left\{T_j(\boldsymbol{z}_i) \geqslant T_j(\boldsymbol{y})\right\}}{|\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i) \cup \{j\}|} \;-\; b_j(\boldsymbol{z}_i)$$

by a smaller value

$$f_j^*(\boldsymbol{z}_i) \;\coloneqq\; \frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i)\right\} \;\vee\; \mathbf{1}\left\{T_j(\boldsymbol{z}_i) \geqslant T_j(\boldsymbol{y})\right\}}{|\mathcal{R}^*(\boldsymbol{z}_i) \cup \{j\}|} \;-\; b_j(\boldsymbol{z}_i)$$

for all Monte-Carlo samples $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots$.

Although computing $\mathcal{R}^*$ is now affordable, it's still heavier comparing to $\mathcal{R}^{\mathrm{Kn}}$. Hence we want to spend our computational budget where such a replacement is most likely to gives us additional rejections. In particular, suppose

$$\boldsymbol{z}_1,\, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k \in \widetilde{\Omega}_j$$

is the sequence of Monte-Carlo samples we use to compute $\widetilde{E}_j^*$, where $\widetilde{E}_j^*$ is the conservative approximation of E_j^*, defined in the same way as in (27). Our computational trick is to replace $f_j(\boldsymbol{z}_i)$ by $f_j^*(\boldsymbol{z}_i)$ only for those $i \in I$, where $I \subseteq [k]$ is a subset of the index set of all the Monte-Carlo samples. So if $|I|$ is small, we only need to compute $\mathcal{R}^*$ for a few Monte-Carlo samples. Formally, let

$$\widehat{E}_j(I) \coloneqq \frac{1}{k}\left(\sum_{i \in I} f_j^*(\boldsymbol{z}_i) + \sum_{i \notin I} f_j(\boldsymbol{z}_i)\right).$$

Then $\widehat{E}_j(I)$ for $I = \varnothing$ is our Monte-Carlo calculation of $\widetilde{E}_j$ and $\widehat{E}_j(I)$ for $I = [k]$ is the Monte-Carlo calculation of $\widetilde{E}_j^*$. The trick is to use $\widehat{E}_j(I)$ for some $I \subseteq [k]$, instead of $\widehat{E}_j([k])$, as the computed approximation of $\widetilde{E}_j^*$.

In principle, we want a small set I that makes $\widehat{E}_j(\varnothing) - \widehat{E}_j(I)$ large. Then $\mathcal{R}^*$ needs being evaluated only on a small set of Monte-Carlo samples but still it is very likely $\widehat{E}_j(I) \leqslant 0$ even if $\widehat{E}_j(\varnothing) > 0$. Specifically, the choice of set I follows the steps below.

1. Note if we want to compute $\widehat{E}_j(I)$, the value of $\widehat{E}_j(\varnothing)$ is an inevitable by-product. Hence it doesn't hurt to compute $\widehat{E}_j(\varnothing)$ before we decide I. Therefore, we set $I = \varnothing$ if $\widehat{E}_j(\varnothing) \leqslant 0$, since it already implies $\widehat{E}_j(I) \leqslant 0$.

2. If the previous step doesn't set $I = \varnothing$, we propose to set $I = \bar{I}$ and trim it in the next step, where
$$\bar{I} := \left\{ i : \ \frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i)\right\} \ \vee \ \mathbf{1}\left\{T_j(\boldsymbol{z}_i) \geqslant c\right\}}{|\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i) \cup \{j\}|} = 1 \right\}.$$
In words, we include i in I only if $\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i) = \varnothing$ and $T_j(\boldsymbol{z}_i) \geqslant c$ for $c = T_j(\boldsymbol{y})$. The idea behinds is that, in this case, including i in I may introduce the largest decline in the value of $\widehat{E}_j(I)$.
Since $\mathcal{R}^{\mathrm{Kn}}(\boldsymbol{z}_i) = \varnothing$ implies that Knockoff is having a hard time to make any rejections, it's uncommon to see $j \in \mathcal{R}^{\mathrm{Kn}}$ in computing $\mathcal{R}^*$, as mentioned in Appendix D.2.3.

3. Following step 2, if the denominators $|\mathcal{R}^*(\boldsymbol{z}_i) \cup \{j\}|$ take the same value for all i, then we can easily solve for the desired smallest value of them, denoted as R^*, so as to make $\widehat{E}_j(\bar{I})$ below zero given $\widehat{E}_j(\varnothing) > 0$. We set $R^* = \infty$ if we can't have $\widehat{E}_j(\bar{I}) \leqslant 0$.
If $R^* > K^{\mathrm{cand}} + 1$, it's impossible that the computed $\mathcal{R}^*$ can be large enough to make H_j rejected, since the computed $|\mathcal{R}^*(\boldsymbol{z}_i) \cup \{j\}| \leqslant K^{\mathrm{cand}} + 1$ by construction. If this is the case, we trim all elements in the proposed $I = \bar{I}$. That is, we set $I = \varnothing$.
If $R^* \leqslant K^{\mathrm{cand}} + 1$, we trim I in an online manner. That is, we compute $\mathcal{R}^*(\boldsymbol{z}_i)$ one by one for $i \in I$. Once we see most $\mathcal{R}^*(\boldsymbol{z}_i)$ computed so far have $|\mathcal{R}^*(\boldsymbol{z}_i) \cup \{j\}| < R^*$, we trim the rest elements in I.

Details of the implementation of this trick can be found in `https://github.com/yixiangLuo/cknockoff`.

Note this computational trick doesn't hurt the FDR control of cKnockoff* no matter how we choose set I. Since $\widehat{E}_j(I)$ is in between the Monte-Carlo calculation of $\widetilde{E}_j$ and $\widetilde{E}_j^*$, the resulting rejection set is in between $\mathcal{R}^{\mathrm{cKn}}$ and the vanilla $\mathcal{R}^{\mathrm{cKn}*}$ without this trick. Theorem C.1 then ensures its FDR control.

E Deferred proofs

E.1 Formulating MCC as a linear model

In this section we formulate Example 1.1 as a linear model. Since all blocks in the experiment are mutually independent, it suffices to show this formulation under $K = 1$, the classical MCC problem. Hence we suppress the subscript k for simplicity of notation. For example, we denote $z_{g,k,i}$ as $z_{g,i}$ in this section.

Denote $\boldsymbol{z}_i = (z_{0,i}, z_{1,i}, \ldots, z_{G,i})^{\mathsf{T}} \in \mathbb{R}^{G+1}$ as the vector of observed outcome in the ith replicate from the control and treatment groups and denote $\boldsymbol{\varepsilon}_i = (\varepsilon_{0,i}, \varepsilon_{1,i}, \ldots, \varepsilon_{G,i})^{\mathsf{T}} \in \mathbb{R}^{G+1}$ as the corresponding noise vector. Define
$$\check{\boldsymbol{I}}_G = \begin{pmatrix} \mathbf{0}^{\mathsf{T}} \\ \boldsymbol{I}_G \end{pmatrix} = \begin{pmatrix} 0 & \ldots & 0 \\ 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \in \mathbb{R}^{(G+1)\times G}$$
as the matrix obtained by padding the identity matrix of dimension G on top with a row vector of all zeros. Then let
$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_r \end{pmatrix} \in \mathbb{R}^{r(G+1)}, \quad \check{\boldsymbol{X}} = \begin{pmatrix} \check{\boldsymbol{I}}_G \\ \check{\boldsymbol{I}}_G \\ \vdots \\ \check{\boldsymbol{I}}_G \end{pmatrix} \in \mathbb{R}^{r(G+1)\times G}, \quad \boldsymbol{\beta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_G \end{pmatrix} \in \mathbb{R}^{G}, \quad \check{\boldsymbol{\varepsilon}} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_r \end{pmatrix} \in \mathbb{R}^{r(G+1)}.$$
We have
$$\boldsymbol{z} = \check{\boldsymbol{X}}\boldsymbol{\beta} + \mathbf{1}\mu + \check{\boldsymbol{\varepsilon}},$$

where $\mathbf{1} = (1, 1, \ldots, 1)^\mathsf{T} \in \mathbb{R}^{r(G+1)}$ is a vector of all ones. This is a linear model with an intercept term. So we can project everything onto the subspace orthogonal to $\mathbf{1}$ to get rid of the intercept. Specifically, let $\boldsymbol{V}_{1,\text{res}} \in \mathbb{R}^{r(G+1)\times(r(G+1)-1)}$ be an orthonormal basis for the subspace orthogonal to $\mathbf{1}$. Define

$$\boldsymbol{y} \coloneqq \boldsymbol{V}_{1,\text{res}}^\mathsf{T}\boldsymbol{z}, \quad \boldsymbol{X} \coloneqq \boldsymbol{V}_{1,\text{res}}^\mathsf{T}\check{\boldsymbol{X}}, \quad \boldsymbol{\varepsilon} \coloneqq \boldsymbol{V}_{1,\text{res}}^\mathsf{T}\check{\boldsymbol{\varepsilon}}.$$

We have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

follows the Gaussian linear model (1) with $m = G$ and $n = r(G+1) - 1$.

More generally, if $K > 1$, we do the same thing for each experiment block and concatenate the linear models derived from them. Specifically, the resulting $\boldsymbol{X} \in \mathbb{R}^{K(r(G+1)-1)\times KG}$ for $K > 1$ is block-diagonal with K blocks of dimension $(r(G+1)-1)$-by-G. Each block of $\boldsymbol{X}$ comes from the same procedure as above applying to each block of the experiments.

In our simulations of Section 5.1, to maintain consistent dimensions $n = 3000, m = 1000$ as we used for the i.i.d. Gaussian case, we slightly modify the linear model construction by removing a few extra residual degrees of freedom, while maintaining the same correlation structure $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ for the explanatory variables. For example, in the MCC problem with $m = G = 1000$, $K = 1$, and $r = 3$, the canonical construction above would give $n = r(G+1) - 1 = 3002$. In our simulation, we remove the extra 2 residual degrees of freedom, while maintaining the correlation structure of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$, thus ensuring that $\hat{\boldsymbol{\beta}}$ has the same positively equicorrelated covariance structure, and $[\boldsymbol{X}\tilde{\boldsymbol{X}}]^\mathsf{T}\boldsymbol{y}$ has the same distribution when we make an equivalent choice of $\tilde{\boldsymbol{X}}$. However, the distribution of the residual variance estimators $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ both change slightly because they are based on 2002 and 1002 residual degrees of freedom, respectively, instead of 2000 and 1000 residual degrees of freedom. A direct comparison confirms the results have no observable difference. For consistency, we also use the construction with $n = 3000$ for the simulation in Figure 2.

E.2 Null distribution of $\boldsymbol{y}$ conditional on S_j

Recall we define

$$\boldsymbol{\Pi}_{-j} = \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\mathsf{T}, \quad \boldsymbol{\Pi}_{-j}^\perp = \boldsymbol{I} - \boldsymbol{\Pi}_{-j}$$

as the projection onto the column span of $\boldsymbol{X}_{-j}$ and its orthogonal projection, respectively. The bijection between $S_j = (\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{y}, \|\boldsymbol{y}\|^2)$ and $(\boldsymbol{\Pi}_{-j}\boldsymbol{y}, \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2)$

$$\boldsymbol{\Pi}_{-j}\boldsymbol{y} \;=\; (\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{X}_{-j})^{-1}) \cdot (\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{y}), \quad \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2 \;=\; \|\boldsymbol{y}\|^2 - \|\boldsymbol{\Pi}_{-j}\boldsymbol{y}\|^2$$

$$\boldsymbol{X}_{-j}^\mathsf{T}\boldsymbol{y} \;=\; \boldsymbol{X}_{-j}^\mathsf{T}(\boldsymbol{\Pi}_{-j}\boldsymbol{y}), \quad \|\boldsymbol{y}\|^2 \;=\; \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2 + \|\boldsymbol{\Pi}_{-j}\boldsymbol{y}\|^2$$

shows that conditioning on S_j is equivalent to conditioning on $(\boldsymbol{\Pi}_{-j}\boldsymbol{y}, \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2)$. Then we have the following null conditional distribution.

Proposition E.1. *Assume the linear model* (1) *and that H_j is true. Then*

$$\boldsymbol{y} \mid \boldsymbol{\Pi}_{-j}\boldsymbol{y}, \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2 \;\overset{d}{=}\; \boldsymbol{\Pi}_{-j}\boldsymbol{y} + \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\| \cdot \boldsymbol{V}_{-j,res}\boldsymbol{u},$$

where $\boldsymbol{V}_{-j,res} \in \mathbb{R}^{n\times(n-m+1)}$ is an orthonormal basis for the subspace orthogonal to the span of $\boldsymbol{X}_{-j}$, and $\boldsymbol{u} \sim$ Unif$(\mathbb{S}^{n-m})$ is a vector uniformly distributed on the unit sphere of dimension $n - m$.

Proof. Since $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ is isotropic Gaussian, $\boldsymbol{\Pi}_{-j}\boldsymbol{y}$ and $\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}$ are independent. So it suffices to show

$$\boldsymbol{V}_{-j,\text{res}}^\mathsf{T} \cdot \boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y} \mid \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\|^2 \;\overset{d}{=}\; \left\|\boldsymbol{\Pi}_{-j}^\perp\boldsymbol{y}\right\| \cdot \text{Unif}(\mathbb{S}^{n-m}).$$

This is true since

$$\left\|\boldsymbol{V}_{-j,\mathrm{res}}^{\mathsf{T}} \cdot \boldsymbol{\Pi}_{-j}^{\perp} \boldsymbol{y}\right\| = \left\|\boldsymbol{\Pi}_{-j}^{\perp} \boldsymbol{y}\right\|$$

and

$$\boldsymbol{V}_{-j,\mathrm{res}}^{\mathsf{T}} \cdot \boldsymbol{\Pi}_{-j}^{\perp} \boldsymbol{y} \;\sim\; \mathcal{N}(\mathbf{0},\, \sigma^2 \boldsymbol{I}_{n-m+1}).$$

The second claim is by

$$\boldsymbol{\Pi}_{-j}^{\perp} \boldsymbol{y} \;\sim\; \mathcal{N}(\boldsymbol{\Pi}_{-j}^{\perp} X\beta,\, \sigma^2 \boldsymbol{\Pi}_{-j}^{\perp} \boldsymbol{I}_n) \;\overset{H_j}{\sim}\; \mathcal{N}(\mathbf{0},\, \sigma^2 \boldsymbol{\Pi}_{-j}^{\perp})$$

and noticing

$$\boldsymbol{\Pi}_{-j}^{\perp} = \boldsymbol{V}_{-j,\mathrm{res}} \boldsymbol{V}_{-j,\mathrm{res}}^{\mathsf{T}}.$$

□

E.3 Proof of Theorem 3.2

The proof of Theorem 3.2 first needs a technical lemma.

Lemma E.1. *Let the budget be defined as in* (19). *Suppose* $\mathbb{P}_{H_j}(b_j > \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}}) \mid S_j) > 0$ *and* $\mathbb{P}_{H_j}(j \notin \mathcal{R}^{\mathrm{Kn}} \mid S_j) > 0$, *then*

$$\mathbb{E}_{H_j}\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \;\middle|\; S_j\right] \;<\; \mathbb{E}_{H_j}\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\} \;\middle|\; S_j\right].$$

Proof. Recall

$$E_j(c\,; S_j) \;\coloneqq\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant c\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;-\; b_j \;\middle|\; S_j\right].$$

Since $b_j \geqslant \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}})$ almost surely and $\mathbb{P}_{H_j}(b_j > \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}}) \mid S_j) > 0$, we have

$$E_j(\infty\,; S_j) < 0.$$

Moreover, $\mathbb{P}_{H_j}(j \notin \mathcal{R}^{\mathrm{Kn}} \mid S_j) > 0$ gives

$$E_j(\infty\,; S_j) < E_j(0\,; S_j).$$

Recall $E_j(c\,; S_j)$ is a continuous, non-increasing function of c and

$$\hat{c}_j = \min_{c \geqslant 0} \left\{E_j(c\,; S_j) \leqslant 0\right\}.$$

We have

$$E_j(\infty\,; S_j) < E_j(\hat{c}_j\,; S_j).$$

That is

$$\mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right] \;<\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right].$$

Now we prove our claim by contradiction. Note

$$\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \leqslant \mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}, \quad \text{almost surely.}$$

So the opposite of our proposition implies

$$\mathbb{P}_{H_j}\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} = \mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\} \;\middle|\; S_j\right] = 1,$$

which further yields

$$\mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right] \;=\; \mathbb{E}_{H_j}\left[\frac{\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}}{|\mathcal{R}^{\mathrm{Kn}} \cup \{j\}|} \;\middle|\; S_j\right]$$

since the denominators of both sides are the same. This contradicts with what we have derived. □

Then we can prove Theorem 3.2.

Proof of Theorem 3.2. For a general testing procedure that produces rejection set $\mathcal{R}(\boldsymbol{y})$, define its *$H_j$-rejection region* as

$$\mathcal{G}_j = \{\boldsymbol{y} : \ j \in \mathcal{R}(\boldsymbol{y})\}.$$

In words, $\mathcal{G}_j$ is the set of observed data $\boldsymbol{y}$ such that H_j is rejected. Note G_j is determined by the testing procedure $\mathcal{R}$ itself and is not affected by the unknown value of $\boldsymbol{\beta}$ or σ.

By construction, the cKnockoff rejection region is always no smaller than the knockoffs, i.e. $\mathcal{G}_j^{\mathrm{cKn}} \supseteq \mathcal{G}_j^{\mathrm{Kn}}$ for all j. So

$$\mathrm{TPR}(\mathcal{R}^{\mathrm{cKn}}) \ = \ \frac{1}{m_1} \sum_{j \in \mathcal{H}_0^{\mathsf{c}}} \mathbb{P}(j \in \mathcal{G}_j^{\mathrm{cKn}}) \ \geqslant \ \frac{1}{m_1} \sum_{j \in \mathcal{H}_0^{\mathsf{c}}} \mathbb{P}(j \in \mathcal{G}_j^{\mathrm{Kn}}) \ = \ \mathrm{TPR}(\mathcal{R}^{\mathrm{Kn}}).$$

To show the inequality is strict, it suffices to prove

$$\mathcal{G}_j^{\mathrm{Kn}} \subsetneq \mathcal{G}_j^{\mathrm{cKn}}$$

for some $j \in \mathcal{H}_0^{\mathsf{c}}$, because the support of the density of $\boldsymbol{y}$ is the whole $\mathbb{R}^n$ space no matter what β is. By the same reason, this is equivalent to show

$$\mathbb{P}_0(j \in \mathcal{R}^{\mathrm{Kn}}) = \mathbb{P}_0(\mathcal{G}_j^{\mathrm{Kn}}) \ < \ \mathbb{P}_0(\mathcal{G}_j^{\mathrm{cKn}}) = \mathbb{P}_0(j \in \mathcal{R}^{\mathrm{cKn}}),$$

where $\mathbb{P}_0$ is the probability measure under the global null model $\mathcal{H}_0 = [m]$.

Recall that under the global null and conditional on $|\boldsymbol{W}|$, $\mathrm{sgn}(W_i)$ are independent Bernoulli for all i. Hence $\mathbb{P}_0(A \mid |\boldsymbol{W}|) > 0$, where

$$A = \{W_j > 0 \text{ and } |\{i : \ W_i > 0\}| = (\lceil 1/\alpha \rceil - 1) \wedge m\}.$$

Note A implies $\mathcal{R}^{\mathrm{Kn}} = \varnothing$ (hence $\mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}}) = 0$) and $b_j > 0$. We have

$$\mathbb{P}_0\left(\left\{b_j > \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}})\right\} \cap \left\{j \notin \mathcal{R}^{\mathrm{Kn}}\right\}\right) > 0$$

by the tower property.

As a consequence, there exist a set $\mathcal{C} \subseteq \mathbb{R}^{m-1} \times \mathbb{R}$ such that $\mathbb{P}_0(S_j(\boldsymbol{y}) \in \mathcal{C}) > 0$ and

$$\mathbb{P}_0\left(\left\{b_j > \mathrm{DP}_j(\mathcal{R}^{\mathrm{Kn}})\right\} \cap \left\{j \notin \mathcal{R}^{\mathrm{Kn}}\right\} \mid S_j\right) > 0$$

for all $S_j \in \mathcal{C}$. Then by Lemma E.1, for any $S_j \in \mathcal{C}$,

$$\mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \ \middle| \ S_j\right] \ < \ \mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\} \ \middle| \ S_j\right],$$

where $\mathbb{E}_0$ is taking expectation over the global null distribution. So

$$\mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \ \middle| \ S_j \in \mathcal{C}\right] \ < \ \mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\} \ \middle| \ S_j \in \mathcal{C}\right].$$

Note

$$\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \leqslant \mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}$$

and $\mathbb{P}_0(S_j \in \mathcal{C}) > 0$. We have

$$\mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\}\right] \ < \ \mathbb{E}_0\left[\mathbf{1}\left\{j \in \mathcal{R}^{\mathrm{Kn}}\right\} \vee \mathbf{1}\left\{T_j \geqslant \hat{c}_j\right\}\right].$$

That is

$$\mathbb{P}_0(j \in \mathcal{R}^{\mathrm{Kn}}) < \mathbb{P}_0(j \in \mathcal{R}^{\mathrm{cKn}}).$$

□

F Numerical simulations

The simulation settings in this section are the same as in Section 5.1 if not specified.

F.1 Extensions of simulations in Section 5.1

F.1.1 Additional design matrix settings

Consider the following two design matrix settings.

1. **OLS $\hat{\beta}_j$ positively auto-regression (Coef-AR)**: Set $\boldsymbol{X}$ such that the OLS fitted $\hat{\beta}_j$ is AR(1) with $\text{cov}(\hat{\beta}_j, \hat{\beta}_{j+1}) = 0.5$.
2. **$\boldsymbol{X}_j$ positively auto-regression (X-AR)**: Set $\boldsymbol{X}$ such that $\boldsymbol{X}_j$ is AR(1) with $\text{cov}(\boldsymbol{X}_j, \boldsymbol{X}_{j+1}) = 0.5$.

Figure 11 shows the results. They are similar to the ones from the MCC-Block problem.

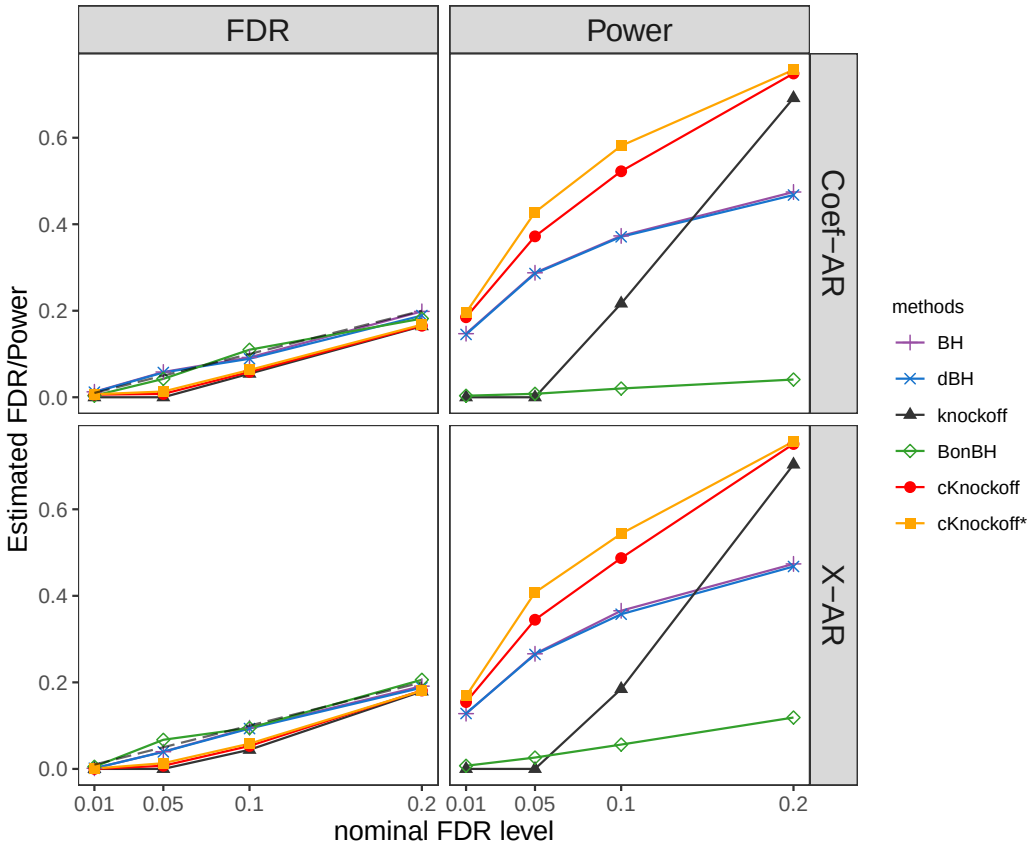

Figure 11: Estimated FDR and TPR for additional design matrix settings.

F.1.2 Cases where $m_1 \gg 1/\alpha$

We show the performance of the procedures in the case where $\boldsymbol{\beta}$ is not too sparse. Figure 11 shows the results with $m_1 = 30$ non-null hypotheses, instead of $m_1 = 10$ in Section 5.1. The general behavior of

the procedures remains the same, but

1. the power-gain of cKnockoff over knockoffs is smaller;
2. the knockoff-like methods are less powerful comparing to the BH-like methods. This is because that the extra power of knockoff-like methods come from sparsity (when lasso-based feature statistics are used).

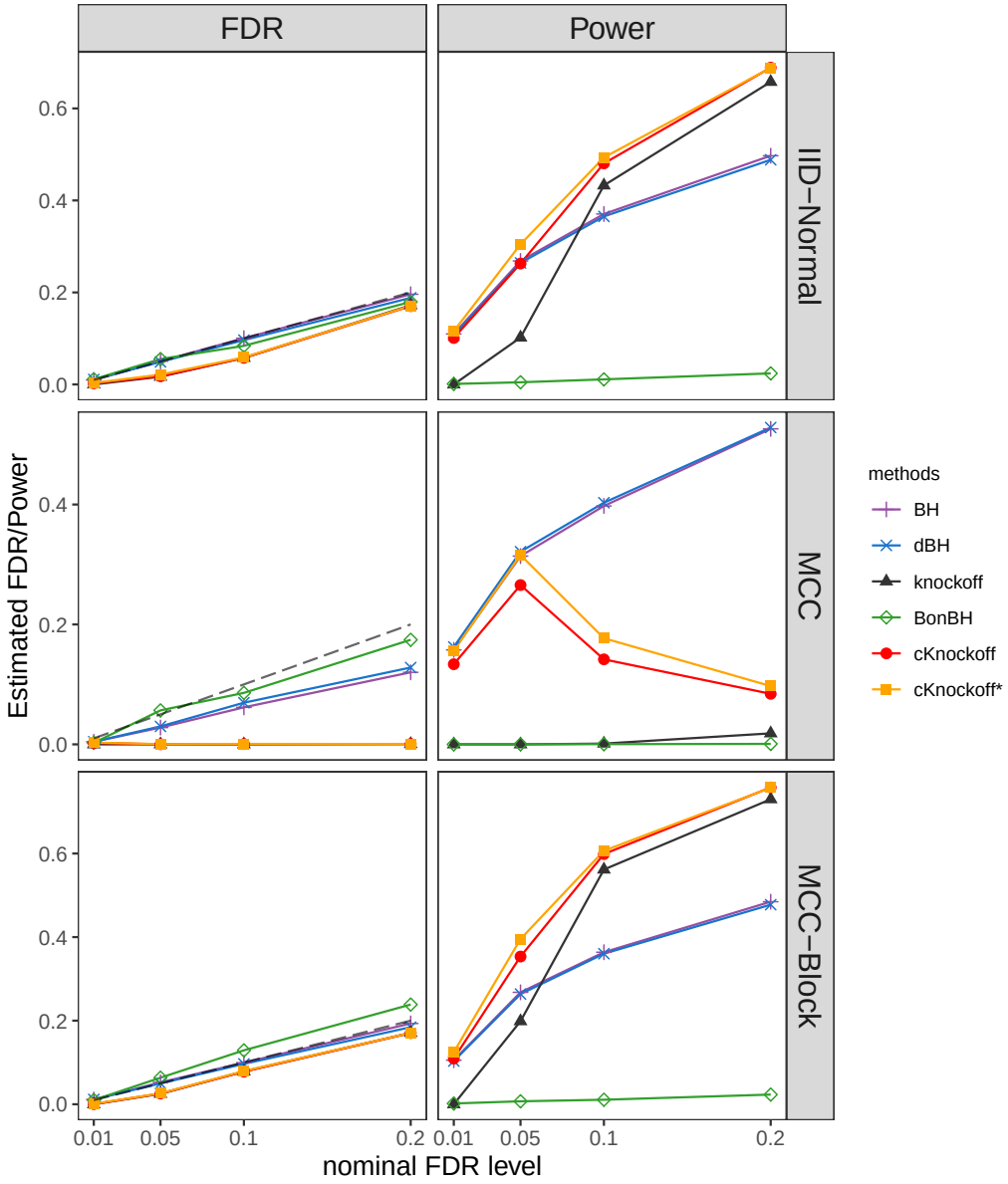

Figure 12: Estimated FDR and TPR with 30 non-null hypotheses.

F.1.3 Comparison with multiple knockoffs

Multiple knockoffs has more implementation choices than the vanilla knockoffs, e.g. the number of knockoff variables to employ and the threshold for deciding whether the original variable "wins" the competition with its knockoffs. There is no known best choice for them, but we implement multiple knockoffs as follows:

1. We employ 5-multiple knockoffs. That is, we generate 5 knockoff matrices $\tilde{\boldsymbol{X}}_{(i)}$ for $i = 1, \dots, 5$ such that $\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \tilde{\boldsymbol{X}}_{(i)}^\mathsf{T}\tilde{\boldsymbol{X}}_{(i)}$ for all i and $\boldsymbol{X}^\mathsf{T}\boldsymbol{X} - \tilde{\boldsymbol{X}}^\mathsf{T}\tilde{\boldsymbol{X}}_{(j)} = \boldsymbol{X}^\mathsf{T}\boldsymbol{X} - \tilde{\boldsymbol{X}}_{(i)}^\mathsf{T}\tilde{\boldsymbol{X}}_{(j)}$ is certain diagonal matrix for all $i \neq j$.

2. For the feature statistics, we run lasso on the augmented model

$$\boldsymbol{y} = [\boldsymbol{X}, \tilde{\boldsymbol{X}}_{(1)}, \dots, \tilde{\boldsymbol{X}}_{(5)}]\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

   with regularity parameter λ determined in the same way as in LCD-T. And let

$$\operatorname{sgn}(W_j) \coloneqq \begin{cases} 1 & \text{if } |\hat{\beta}_{j,(0)}| > |\hat{\beta}_{j,(i)}| \text{ for all } i \in [5] \\ -1 & \text{otherwise} \end{cases}, \qquad |W_j| \coloneqq \max_{i=0,\dots,5} \left\{ |\hat{\beta}_{j,(i)}| \right\},$$

   where $\hat{\beta}_{j,(i)}$ is the fitted lasso coefficient of the ith knockoffs of the variable $\boldsymbol{X}_j$ and $\hat{\beta}_{j,(0)}$ is the fitted lasso coefficient of the original variable $\boldsymbol{X}_j$.

To make multiple knockoffs applicable, we set $n = 7m$ with $m = 300$. The number of hypotheses is smaller than our usual setting, so as to save memory space in our laptop. The number of non-null hypotheses is set to be $m_1 = 30$ so as to show the performance of multiple knockoffs in the region where $m \gg 1/\alpha$.

Figure 13 shows the results. When $m_1 < 1/\alpha$, multiple knockoffs relieves the threshold phenomenon but performs worse than cKnockoff/cKnockoff*; when $m_1 \gg 1/\alpha$, multiple knockoffs is even worse than the vanilla knockoffs. Moreover, multiple knockoffs doesn't help the whiteout phenomenon.

In addition, we see the knockoff-like methods are even less powerful, comparing to the BH-like methods, in Figure 13 than in Figure 12. This is because $\pi_1 = m_1/m = 0.1$ in this setting. That is, $\boldsymbol{\beta}$ is less sparse.

F.2 Variations of cKnockoff

Figure 14 shows the estimated FDR and power when we use C-LSM rather than LCD-T as the feature statistics. The behavior of cKnockoff/cKnockoff* is almost the same as in Figure 3.

F.3 Robustness tests

In this section, we test the robustness of our methods, especially if the FDR is still controlled when certain model assumptions don't hold.

F.3.1 Noisy signal

First consider the case where $\boldsymbol{\beta}$ is noisy. That is, the null hypotheses now have

$$\beta_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(-a, a), \ j \in \mathcal{H}_0$$

for some $a > 0$. And we set $\alpha = 0.2$ so that knockoffs can make a decent number of rejections.

Figure 15 shows the results as a/β^* increases. The behavior of the estimated FDR are the same for knockoff, cKnockoff, and cKnockoff*. They all roughly controls FDR when $a/\beta^* \leqslant 0.1$.

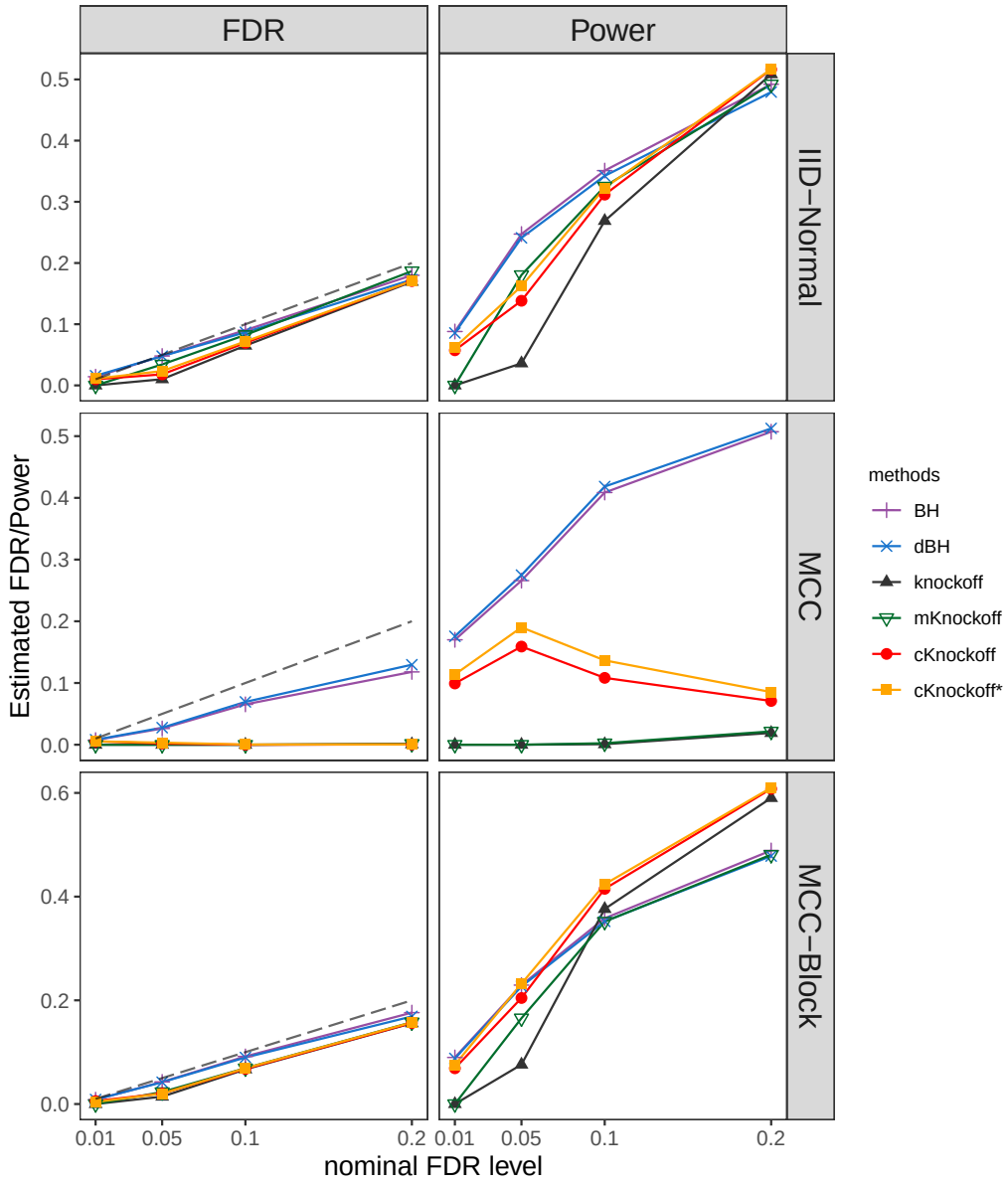

Figure 13: Estimated FDR and TPR with multiple knockoffs included.

F.3.2 Heavy-tailed noise

Consider a t-noise instead of a Gaussian noise in the linear model (1), i.e.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} t_k$$

with certain degree of freedom k. We set $\alpha = 0.2$ as before and apply the procedures on an additional design matrix setting:

Sparse: Set $\boldsymbol{X}$ to have diagonal $X_{ii} = 1$ and off-diagonal entries $X_{ij} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.01) for $i \neq j$.

Figure 16 shows the results as the degree of freedom k decreases (noise more heavy-tailed). For the dense design matrices like IID-Normal, MCC, and MCC-Block, cKnockoff/cKnockoff* is robust

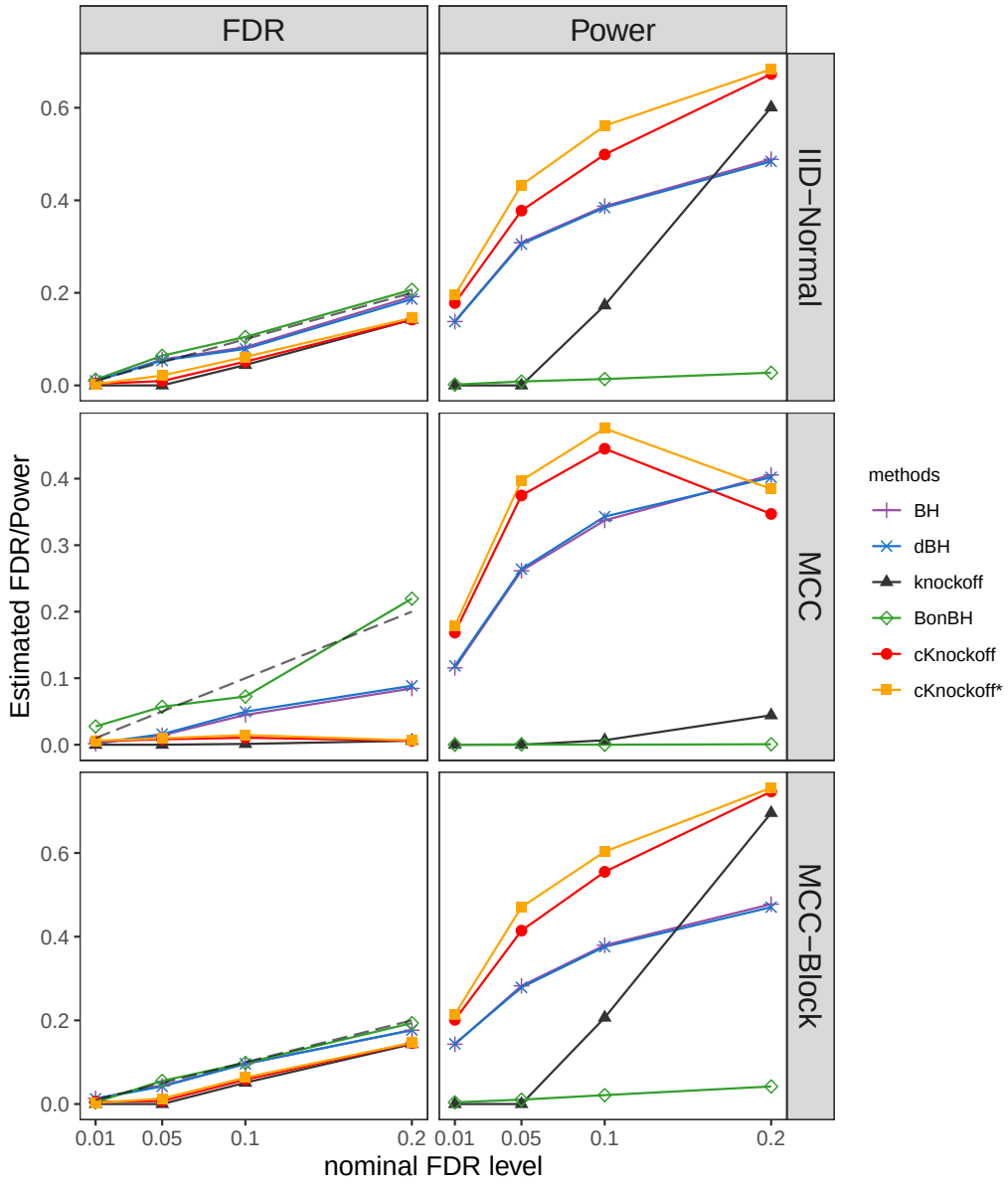

Figure 14: Estimated FDR and TPR under settings using C-LSM as the feature statistics.

in the same way as knockoffs. But for the sparse design matrix, cKnockoff/cKnockoff* doesn't control FDR when k is small, in a way similar to the p-value based methods BH and dBH. This is because $\boldsymbol{X}^\top \boldsymbol{y}$ is now heavy-tailed. So some null hypotheses have p-values stochastically smaller than uniform. Recall our calibration statistic T_j is roughly a proxy of the one-sided p-value. Such a deviation in the p-value distribution is expected to affect our methods more than knockoffs.

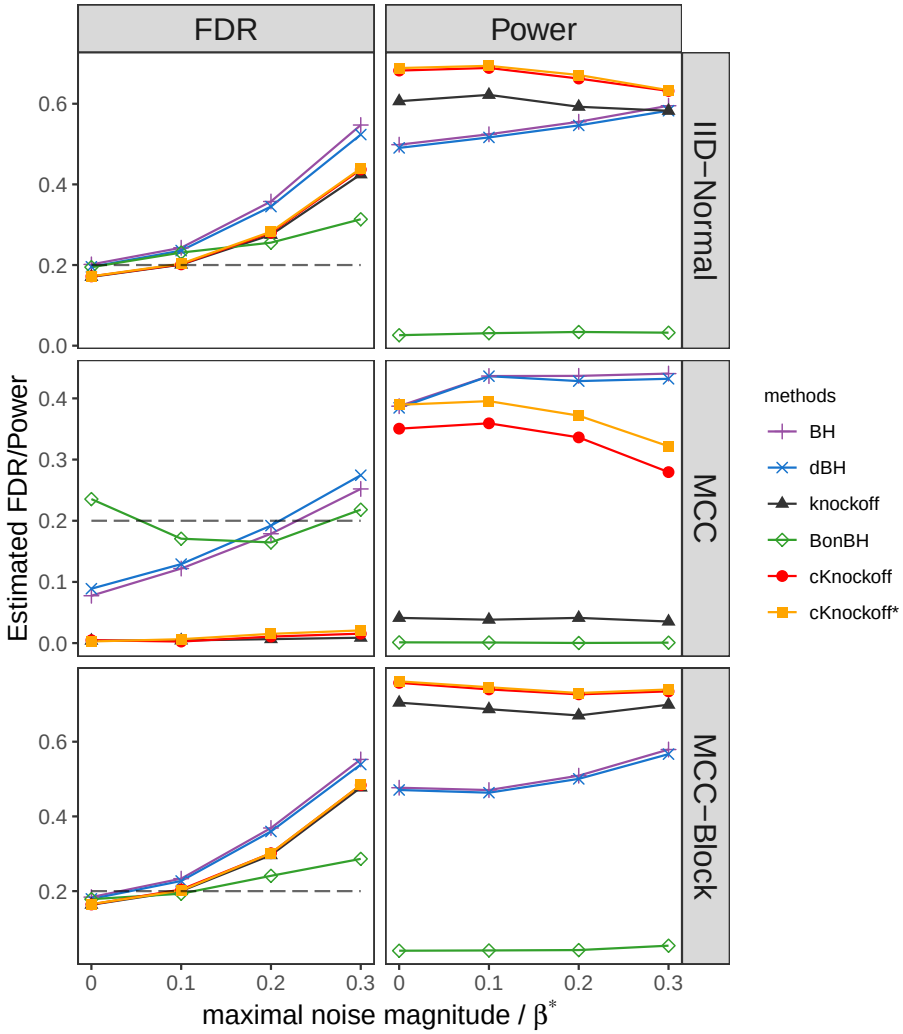

Figure 15: Estimated FDR and TPR as the signal becomes noisy.

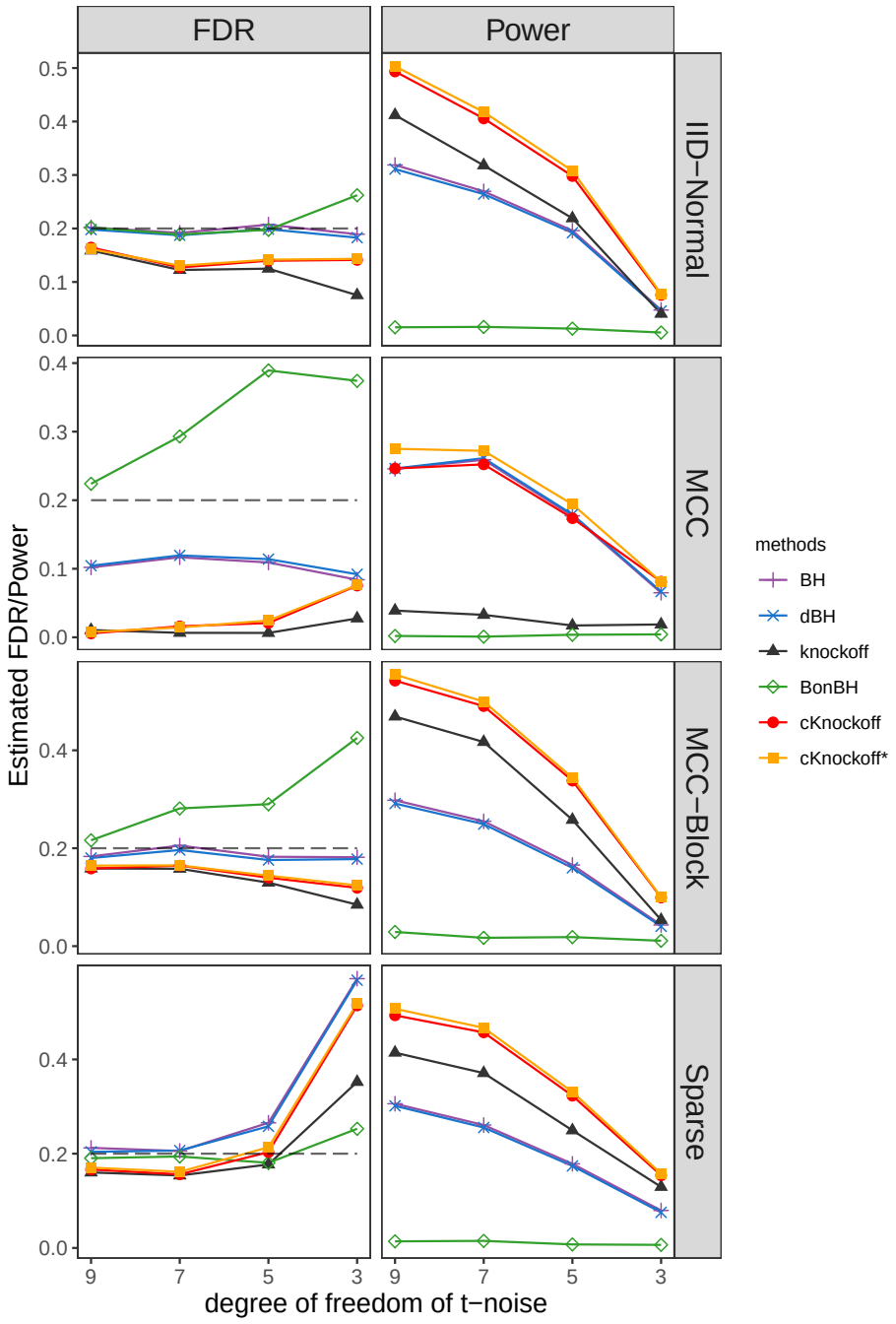

Figure 16: Estimated FDR and TPR when the noise is a heavy-tailed t-distribution.