Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization

Jaime Roquero Gimenez
Department of Statistics
Stanford University
Stanford, CA 94305
roquero@stanford.edu

James Zou
Department of Biomedical
Data Science
Stanford University
Stanford, CA 94305
jamesz@stanford.edu

Abstract

The *Model-X knockoff* procedure has recently emerged as a powerful approach for feature selection with statistical guarantees. The advantage of knockoffs is that if we have a good model of the features X, then we can identify salient features without knowing anything about how the outcome Y depends on X. An important drawback of knockoffs is its instability: running the procedure twice can result in very different selected features, potentially leading to different conclusions. Addressing this instability is critical for obtaining reproducible and robust results. Here we present a generalization of the knockoff procedure that we call simultaneous multi-knockoffs. We show that multi-knockoffs guarantee false discovery rate (FDR) control, and are substantially more stable and powerful compared to the standard (single) knockoffs. Moreover we propose a new algorithm based on entropy maximization for generating Gaussian multi-knockoffs. We validate the improved stability and power of multi-knockoffs in systematic experiments. We also illustrate how multi-knockoffs can improve the accuracy of detecting genetic mutations that are causally linked to phenotypes.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 INTRODUCTION

In many machine learning and statistics settings, we have a supervised learning problem where the outcome Y depends on a subset of the features X, potentially in complex ways, and we would like to identify these salient features. Take medical genetics as an example, the features $X = (X_1, ..., X_d)$ are the genotypes at d variants in the genome, and Y is a binary indicator for the presence/absence of disease. The true model could be that $Y = f(X_{\mathcal{H}}, \omega)$, where $X_{\mathcal{H}} = \{X_i : i \in \mathcal{H}\}$ is the salient subset of the variants, and ω is some noise/randomness. It is tremendously important to identify which feature/variant is in $\mathcal{H}$.

If we assume that f is simple, say a linear function, then we might hope to use the fitted parameters of a model to select salient features. For example, we might fit a Generalized Linear Models (GLM) on (X, Y) with LASSO penalty to promote sparsity in the coefficients (Tibshirani, 1996), and subsequently select those features with non-zero coefficients. Step-wise procedures where we sequentially modify the model is another way of doing feature selection (Mallows, 1973).

A clear limitation of this parametric approach is the need to to have a good model for f. For the genetics example, there is no great model. Moreover the standard feature selection methods are all plagued by correlations between the features: a feature that is not really relevant for the outcome, i.e. not in $\mathcal{H}$, can be selected by LASSO or Step-wise procedure, because it is correlated with relevant features. In these settings we usually lack statistical guarantees on the validity of the selected features. Finally, even procedures with statistical guarantees usually depend on having *valid p-values*, which are based on a correct modeling of $Y|X$ and (sometimes) assume some asymptotic regime. However there are many common settings where these assumptions fail and we cannot perform inference based

on those p-values (Sur and Candès, 2018).

A powerful new approach called *Model-X knockoff procedure* (Candès et al., 2018) has recently emerged to deal with these issues. This method introduces a new paradigm: we no longer assume any model for the distribution of $Y|X$ in order to do feature selection (and therefore do not compute p-values), but we assume that we have full knowledge of the feature distribution P^X – or at least we can accurately model it, though there are some robustness results (Barber et al., 2018). This knowledge of the ground truth P^X allows us to sample new *knockoff* variables $\tilde{X}$ satisfying some precise distributional conditions. Although we make no assumption on $Y|X$, we can use the knockoff procedure to select features while controlling the False Discovery Rate (FDR), which is the average proportion of the selected features that are not in $\mathcal{H}$.

One major obstacle for the widespread application of knockoffs is its instability. The entire procedure sensitively depends on the knockoff sample $\tilde{X}$, which is random. Therefore, running the knockoff procedure twice may lead to very different selected sets of features. Our analysis in Section 4 shows that instability is especially severe when the number of salient features (i.e. the size of $\mathcal{H}$) is small, as is often the case. Also, whenever the number of features is very large, previous methods for generating knockoffs failed to consistently generate good samples for $\tilde{X}$, leading to inconsistent selection sets if several runs of the procedure were done simultaneously. Power also decreases drastically under those previous methods. This makes it challenging to confirm the selected variants in a replication experiment. Addressing the instability of knockoffs is therefore an important problem.

Our Contributions. We generalize the standard (single) knockoff procedure to simultaneous multiple knockoffs (or multi-knockoffs for short). Our multi-knockoff procedure guarantees FDR control and has better statistical properties than the original knockoff, especially when the number of salient features is small. We propose a new entropy maximization algorithm to sample Gaussian multi-knockoffs. Our systematic experiments demonstrate that multi-knockoffs is more stable and more powerful than the original (single) knockoffs. Moreover we illustrate how multi-knockoffs can improve the ability to select causal variants in Genome Wide Association Studies (GWAS).

2 BACKGROUND ON KNOCKOFFS

We begin by introducing the usual setting of feature selection procedures. We consider the data as a sequence of i.i.d. samples from some unknown joint distribution: $(X_{i1}, \dots, X_{id}, Y_i) \sim P^{XY}$, $i = 1, \dots, N$. We then define the set of null features $\mathcal{H}_0 \subset \{1, \dots, d\}$ by $j \in \mathcal{H}_0$ if and only if $X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j}$ (where the $-j$ subscript indicates all variables except the jth and bold letters indicate vectors). The non-null features, also called alternatives, are important because they capture the truly salient effects and the goal of selection procedures is to identify them. Running the knockoff procedure gives us a selected set $\hat{\mathcal{S}} \subset \{1, \dots, d\}$, while controlling for False Discovery Rate (FDR), which stands for the expected rate of false discoveries: $FDR = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| \vee 1}\right]$. The ratio $\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| \vee 1}$ is also called False Discovery Proportion (FDP).

Assuming we know the ground truth for the distribution P^X, the first step of the standard knockoff procedure is to obtain a *knockoff* sample $\tilde{\boldsymbol{X}}$ that satisfies the following conditions:

Definition 2.1 (Knockoff sample)**.** *A knockoff sample $\tilde{\boldsymbol{X}}$ of a d-dimensional random variable $\boldsymbol{X}$ is a d-dimensional random variable such that two properties are satisfied:*

- *Conditional independence:* $\tilde{\boldsymbol{X}} \perp\!\!\!\perp Y | \boldsymbol{X}$
- *Exchangeability :*

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)} \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}] \qquad \forall S \subset \{1, \dots, d\}$$

where the symbol $\overset{d}{=}$ stands for equality in distribution and $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}$ refers to the vector where the original jth feature and the jth knockoff feature have been swapped whenever $j \in S$.

The first condition is immediately satisfied as long as knockoffs are sampled conditionally on the sample $\boldsymbol{X}$ without considering any information about Y (which will be the case in our sampling methods so we will not mention it again). The second condition ensures that the knockoff of each feature is sufficiently similar to the original feature in order to be a good comparison baseline. We also denote by $\boldsymbol{X}$ the $N \times d$ matrix where we stack the N i.i.d. d-dimensional samples into one matrix (this is acceptable as the i.i.d. assumption allows for all sampling procedures to be done samplewise).

The next step of the knockoff procedure constructs what we call *feature statistics* $\boldsymbol{W} = (W_1, \dots, W_d)$, such that a high value for $W_j \in \mathbb{R}$ is evidence that the jth feature is non-null. Feature statistics described in Candès et al. (2018) depend only on $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \in \mathbb{R}^{N \times 2d}, Y \in \mathbb{R}^N$ such that for each $j \in \{1, \dots, d\}$ we can write $W_j = w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y)$ for some function w_j. The only restriction these statistics must satisfy is the *flip-sign property*: swapping the jth feature and its corresponding knockoff feature should flip the sign of the statistic

W_j while leaving other feature statistics unchanged. More formally, for a subset $S \subset \{1, \dots, d\}$ of features, denoting $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}$ the data matrix where the original jth variable and its corresponding knockoff have been swapped whenever $j \in S$, we have:

$$w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}, Y) = \begin{cases} -w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y), & \text{if } j \in S. \\ w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y), & \text{otherwise.} \end{cases}$$

As suggested in Candès et al. (2018), the choice of feature statistics can be done in two steps: first, find a statistic

$$\bar{\boldsymbol{T}} = \bar{\boldsymbol{T}}([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y) = (T_1, \dots, T_d, \tilde{T}_1, \dots, \tilde{T}_d) \in \mathbb{R}^{2d}$$

where each coordinate corresponds to the "importance" — hence we will call them *importance scores* — of the corresponding feature (either original or knockoff). For example, T_j would be the absolute value of the regression coefficient of the jth feature.

After obtaining the importance score for the original and knockoff feature, we take the difference to compute the feature statistic $W_j = T_j - \tilde{T}_j$. The intuition is that importance scores of knockoffs serve as a control, a larger importance score of the jth feature compared to that of its knockoff implies a larger positive W_j (and therefore is evidence against the null). Given some target FDR level $q \in (0, 1)$ that we fix in advance, we define the selection set $\hat{\mathcal{S}}$ based on the following threshold $\hat{\tau}$:

$$\hat{\tau} = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\} \quad (1)$$

$$\hat{\mathcal{S}} = \{j : W_j \geq \hat{\tau}\} \quad (2)$$

According to Theorem 3.4 in Candès et al. (2018), this procedure controls FDR at level q (actually called Knockoff + procedure). The mechanism behind this procedure is the Selective SeqStep+ introduced in Barber et al. (2015). The intuition is that we try to maximize the number of selections while bounding by q an upwardly biased estimate of the FDP, which is the fraction:

$$\widehat{FDP}_{KN+} = \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \quad (3)$$

The added constant in the numerator is called the "offset", equal to one in our case; a different FDP estimate with offset equal to 0 leads to a slightly different procedure that controls a modified, less stringent version of the FDR.

Instability of Knockoffs If we generate multiple replication datasets —multiple versions of $(\boldsymbol{X}, Y)$, each of which is sampled from the common P^{XY}—then the knockoff procedure guarantees that *on average*, the proportion of false discoveries is less than the desired threshold. However for a particular dataset $(\boldsymbol{X}, Y)$, the selected features could be very different from that of another sampled dataset, as we empirically demonstrate in Section 4. There are settings where in half of the experiments, the knockoff procedure selects a large number of features, and it selects zero features the other half of the times. This instability is a major issue if we want to ensure that the discoveries from data are reproducible.

The instability in the selected features is partially due to the randomness in $(\boldsymbol{X}, Y)$ but also in the knockoff sample $\tilde{\boldsymbol{X}}$. The knockoff procedure, equations 1-3, is sensitive to the sample $\tilde{\boldsymbol{X}}$. The knockoff selection set is based on a conservative estimate of the FDP given by equation (3). The threshold that determines the selected set requires such FDP estimate to be below some target FDR level q set in advance, which in turn requires us to select at least $\lceil \frac{1}{q} \rceil$ features due to the presence of the offset in the numerator. This requirement is a great source of instability of the knockoff procedure: whenever the number of non-nulls is close to that threshold value $\lceil \frac{1}{q} \rceil$, we can end up either selecting a fairly large number of non-nulls, or not selecting any, even when the signal is strong. Our goal is to develop a new knockoff procedure which controls FDR and is more stable. We achieve this by introducing simultaneous multiple knockoffs, called multi-knockoffs for short, which extends the standard knockoff procedure.

3 SIMULTANEOUS MULTIPLE KNOCKOFFS

A Naive Flawed Approach to Multi-knockoffs One approach to improve the stability of the selected features is to run the standard knockoff procedure multiple times in parallel and to take some type of consensus. This approach is flawed and does not control FDR. The reason is that by running knockoff multiple times in parallel, the symmetry between $\boldsymbol{X}$ and the knockoff samples is broken. To maintain symmetry and guarantee FDR, we need to simultaneously sample multiple knockoffs. We will make this more precise now.

3.1 Multi-knockoff Selection Procedure

Fix an positive integer $\kappa \geq 2$, the multi-knockoff parameter (the usual single knockoff case corresponds to $\kappa = 1$, for which all of our results are also valid). The goal is to extend the previous distributional properties of knockoffs of Definition 2.1 to settings where we simultaneously sample κ knockoff copies $(\boldsymbol{X}^k)_{1 \leq k \leq \kappa}$ of the same $\mathbb{R}^d$-valued dataset $\boldsymbol{X}$ (where, again, $\boldsymbol{X}$ denotes either the $\mathbb{R}^d$-valued random variable when making distributional statements or a $\mathbb{R}^{N \times d}$ matrix when referring to the feature set of N i.i.d. samples). As in the single knockoff setting, we can define an equivalence notion and notation for swapping multiple vectors. Instead of defining $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}$ for some subset $S \subset \{1, \dots, d\}$

of indices, we consider a collection $\sigma = (\sigma_i)_{1\le i\le d}$ of d permutations σ_i over the set of integers $\{0, 1, \dots, \kappa\}$, one for each of the d initial dimensions. Whenever we will use multi-knockoffs, we will index the original features $\boldsymbol{X}$ by $\boldsymbol{X}^0$. We define the permuted vector $[\boldsymbol{X}^0, \boldsymbol{X}^1, \dots \boldsymbol{X}^\kappa]_{swap(\sigma)} := [\boldsymbol{U}^0, \boldsymbol{U}^1, \dots, \boldsymbol{U}^\kappa]$, where $U_i^k = X_i^{\sigma_i(k)}$ for all $1 \le i \le d$, $0 \le k \le \kappa$. Each σ_i permutes the $\kappa + 1$ features corresponding to the ith dimension of each vector $\boldsymbol{X}^k$, leaving the other dimensions unchanged. Once this generalized swap notion is defined, we extend the exchangeability property based on the invariance of the joint distribution to such transformations.

Definition 3.1. *We say that the concatenated vectors $[\boldsymbol{X}^0, \boldsymbol{X}^1, \dots \boldsymbol{X}^\kappa]$ satisfy the extended exchangeability property if the equality in distribution $[\boldsymbol{X}^0, \boldsymbol{X}^1, \dots \boldsymbol{X}^\kappa]_{swap(\sigma)} \overset{d}{=} [\boldsymbol{X}^0, \boldsymbol{X}^1, \dots \boldsymbol{X}^\kappa]$ holds for any σ as defined above.*

Definition 3.2. *We say that $(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$ is a multi-knockoff vector of $\boldsymbol{X}^0$ (or that they are κ multi-knockoffs of $\boldsymbol{X}^0$) if the joint vector $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$ satisfies extended exchangeability and the conditional independence requirement $(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \perp\!\!\!\perp Y | \boldsymbol{X}^0$.*

We will later on give examples of how to generate such multi-knockoffs. We state a lemma that is a direct generalization of Lemma 3.2 in Candès et al. (2018) and give a proof in Appendix B.1.

Lemma 3.1. *Consider a subset of nulls $S \subset \mathcal{H}_0$. Define a generalized swap σ as above, where σ_i is the identity permutation whenever $i \notin S$, and otherwise can be any permutation. Then we have the following equality in distribution for a multi-knockoff:*

$$([\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa], Y) \overset{d}{=} ([\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa]_{swap(\sigma)}, Y)$$

Once the multi-knockoff vector is sampled, consider the joint vector $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$ which takes values in $\mathbb{R}^{(\kappa+1)d}$. As for the single knockoff setting, we construct importance scores $\bar{\boldsymbol{T}} = (\boldsymbol{T}^0, \boldsymbol{T}^1, \dots, \boldsymbol{T}^\kappa)$ where each $\boldsymbol{T}^k$ is a d-dimensional vector with non-negative entries. The importance scores are associated to the features in the following sense: if we generate importance scores on a swapped joint vector (as in Definition 3.1), then we obtain the same result as if we had swapped the importance scores of the initial joint vector. That is, the function defining the importance scores must satisfy $[\boldsymbol{T}^0, \boldsymbol{T}^1, \dots, \boldsymbol{T}^\kappa]_{swap(\sigma)} = \bar{\boldsymbol{T}}([\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa]_{swap(\sigma)}, Y)$. Common examples of such constructions are the absolute values of the coefficients associated to each feature when regressing Y on $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$, with eventually a L^1 penalty for sparsity. Denoting an ordered sequence by indexing in parenthesis (i.e. for any real-valued sequence $(a_0, \dots, a_n)$, we have $a_{(0)} \ge a_{(1)} \ge \dots \ge a_{(n)}$), we can define feature-wise ordered importance scores $(T_i^{(k)})_{0\le k\le\kappa}$ for each feature $1 \le i \le d$. For all $1 \le i \le d$, define:

$$\kappa_i = \underset{0\le k\le\kappa}{\arg\max}\, T_i^k \qquad \tau_i = T_i^{(0)} - T_i^{(1)}$$

We no longer have the possibility of generating feature statistics W_i by taking an antisymmetric function of the importance scores. The extension to multi-knockoffs is done through these newly defined variables by noticing the analogy that if $\kappa = 1$ (single knockoff), then $\tau_i = |W_i|$ and κ_i corresponds to the sign of W_i ($\kappa_i = 0$ if and only if $W_i > 0$). In the single knockoff setting, the crucial distributional result is that, conditionally on $|W_i|$, the signs of the null W_i are i.i.d. flip coins. In the multi-knockoff case, the information encoded by the sign of W_i is contained in κ_i, which indicates whether among a given dimension i the original feature has a higher importance score than that of its knockoffs. In Appendix B.4 we provide a geometric explanation for such choices of κ_i, τ_i. The crucial result is that null κ_i behave uniformly and independently in distribution and can be used to estimate the number of false discoveries.

Lemma 3.2. *The random variables $(\kappa_i)_{i\in\mathcal{H}_0}$ are i.i.d. distributed uniformly on the set $\{0, 1, \dots, \kappa\}$, and independent of the remaining variables $(\kappa_i)_{i\notin\mathcal{H}_0}$, and of the feature-wise ordered importance scores $[(T_i^{(k)})_{0\le k\le\kappa}]_{1\le i\le d}$. In particular, conditionally on the variables $(\kappa_i)_{i\notin\mathcal{H}_0}$ and $(\tau_i)_{1\le i\le d}$, the random variables $(\kappa_i)_{i\in\mathcal{H}_0}$ are i.i.d. distributed uniformly on $\{0, 1, \dots, \kappa\}$.*

We prove this lemma in Appendix B.2. Following the steps that build the knockoff procedure as a particular case of the SeqStep+ procedure, we construct the following threshold $\hat{\tau}$ that defines the rejection set $\hat{\mathcal{S}}$ of our multi-knockoff procedure based on a FDP estimate

$$\widehat{FDP}_{\kappa KN+} = \frac{\frac{1}{\kappa} + \frac{1}{\kappa}\#\{i \in \{1, \dots, d\},\, \kappa_i \ge 1,\, \tau_i \ge t\}}{\#\{i \in \{1, \dots, d\},\, \kappa_i = 0,\, \tau_i \ge t\} \vee 1}$$

Essentially, the multi-knockoff procedure returns the features i where the original feature has higher importance score than any knockoffs (i.e. $\kappa_i = 0$), and the gap with the 2nd largest importance score is above some threshold.

Algorithm 1: Multi-knockoff Selection Procedure

Input : Concatenated vector $\bar{\boldsymbol{T}} = (\boldsymbol{T}^0, \boldsymbol{T}^1, \dots, \boldsymbol{T}^\kappa)$ of importance scores, target FDR level q

Output : Set of selected features $\hat{\mathcal{S}}$

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

  $\kappa_i = \arg\max_{0\le k\le\kappa} T_i^k\,, \quad \tau_i = T_i^{(0)} - T_i^{(1)}$

end

$\hat{\tau} = \min\left\{t > 0 : \frac{\frac{1}{\kappa} + \frac{1}{\kappa}\#\{i\in\{1,\dots,d\},\, \kappa_i \ge 1,\, \tau_i \ge t\}}{\#\{i\in\{1,\dots,d\},\, \kappa_i = 0,\, \tau_i \ge t\} \vee 1} \le q\right\}$

return $\hat{\mathcal{S}} = \{i \in \{1, \dots, d\},\, \kappa_i = 0,\, \tau_i \ge \hat{\tau}\}$

Proposition 3.3. *Fix a target FDR level $q \in (0,1)$. The procedure that selects the features in the set $\hat{\mathcal{S}}$ given by Algorithm 1 controls FDR at level q.*

We prove this result in Appendix B.3. One advantage of the multi-knockoff selection procedure lies on the new value of the offset parameter. By averaging over the κ multi-knockoffs, we are able to decrease the threshold of minimum number of rejections from $\lceil \frac{1}{q} \rceil$ to $\lceil \frac{1}{q\kappa} \rceil$, leading to an improvement in power and stability. We call $\lceil \frac{1}{q\kappa} \rceil$ the detection threshold of the multi-knockoff. We experimentally confirm such results in Section 4.

3.2 Gaussian Multi-knockoffs Based on Entropy Maximization

Most of the research and applications have focused around generating standard knockoffs when $\boldsymbol{X}$ comes from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, although more universal sampling algorithms exist, for which we provide in Appendix A a generalization to multi-knockoffs. Here we extend the existing procedures for Gaussian knockoffs to generate Gaussian multi-knockoffs for $\kappa \geq 2$. A sufficient condition for $(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \in \mathbb{R}^{d\kappa}$ to be a multi-knockoff vector is for $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \in \mathbb{R}^{d(\kappa+1)}$ to be jointly Gaussian such that: 1) all the $\boldsymbol{X}^k$ has the same mean μ; and 2) the covariance matrix has the form :

$$\Sigma_\kappa = \underbrace{\begin{pmatrix} \Sigma & \Sigma - D & \dots & \Sigma - D \\ \Sigma - D & \Sigma & \dots & \Sigma - D \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma - D & \Sigma - D & \dots & \Sigma \end{pmatrix}}_{\kappa+1 \text{ blocks}}$$

where D is a diagonal matrix chosen so that Σ_κ is positive semi-definite to ensure that it is a valid covariance.

Proposition 3.4. *If $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \in \mathbb{R}^{d(\kappa+1)}$ has the mean and covariance structure given above, then $(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$ is a valid κ multi-knockoff of $\boldsymbol{X}^0$.*

If a diagonal term D_{ii} is zero, then $\boldsymbol{X}_i^k = \boldsymbol{X}_i^0$ for $k \geq 1$. This generates a valid multi-knockoff but it has no power to discover the ith feature (regardless of whether it is null or non-null) since each multi-knockoff is indistinguishable from the original feature. The general intuition is that the more independent the knockoffs are from the original $\boldsymbol{X}^0$, the greater the power of discovering the non-null features (Candès et al., 2018). Therefore previous work for the standard single knockoff (corresponding to $\kappa = 1$) has focused on finding D as large as possible in some sense, while maintaining the positive semi-definiteness of the covariance matrix.

To construct D for Gaussian multi-knockoffs, we propose maximizing the entropy $H(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa)$ (which has a simple closed form for Gaussian distributions). This is equivalent in the single knockoff case to minimizing mutual information, as suggested in Candès et al. (2018). Indeed, $I(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = H(\boldsymbol{X}) + H(\tilde{\boldsymbol{X}}) - H(\boldsymbol{X}, \tilde{\boldsymbol{X}})$, and $H(\boldsymbol{X}) = H(\tilde{\boldsymbol{X}})$ do not depend on D, hence the equivalence.

Entropy Knockoffs The diagonal matrix $D(s) = diag(s_1, \dots, s_d)$ for constructing entropy multi-knockoffs is given by the following convex optimization problem:

$$\arg\min_s \; -\log\det\left(\frac{\kappa+1}{\kappa}\Sigma - D(s)\right) - \kappa \sum_{i=1}^{d} \log(s_i)$$
$$\text{subject to } \begin{cases} \frac{\kappa+1}{\kappa}\Sigma - D(s) \succ 0 \\ s_i \geq 0 \qquad \forall i \in \{1, \dots, d\} \end{cases}$$

This optimization problem is a convex optimization problem, by noticing that $s \mapsto -\log\det(\frac{\kappa+1}{\kappa}\Sigma - D(s))$ is convex. It can be solved efficiently and our implementation is based on the Python package CVXOPT (M. S. Andersen and Vandenberghe, 2012). This knockoff construction method avoids solutions where diagonal terms are extremely close to 0, and we provide the following lower bound on the diagonal terms of D:

$$\left(\lambda_{min}(s)\right)^{\frac{1}{\kappa-2}} \leq \min_{1 \leq j \leq d} s_j$$

where $\lambda_{min}(s)$ is the smallest (positive) eigenvalue of $\frac{\kappa+1}{\kappa}\Sigma - D(s)$. The fact that we maximize the value of the determinant of such matrix implies that we avoid having any extremely small eigenvalue, hence this bound proves useful. We provide additional analysis on the formulation of entropy maximization as a convex optimization problem and prove this lower bound in Appendix A. Once the diagonal matrix D is computed, we can generate the Gaussian multi-knockoffs by writing the conditional distribution given the original features $\boldsymbol{X}^0$.

$$(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) | \boldsymbol{X}^0 \sim \mathcal{N}\big((\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^\kappa), \tilde{\Sigma}\big), \text{ where}$$
$$\begin{cases} \boldsymbol{\mu}^i = D\Sigma^{-1}\boldsymbol{\mu} + (I_d - D\Sigma^{-1})\boldsymbol{X}^0 \quad \forall 1 \leq i \leq \kappa \\ \tilde{\Sigma} = \begin{pmatrix} C & \dots & C - D \\ C - D & \dots & C - D \\ \vdots & \vdots & \vdots \\ C - D & \dots & C \end{pmatrix} \text{ and } C = 2D - D\Sigma^{-1}D \end{cases}$$

In the single knockoff setting ($\kappa = 1$), the standard approach in literature is to solve a semidefinite program (SDP) to optimize D. An alternative approach in the literature, called equicorrelation, is to restrict D to be $D = sI_d$ and solve for s (where we consider Σ as a correlation matrix, the goal being having the same correlation between original features and knockoffs in every dimension). We provide natural generalizations of the SDP and equicorrelation to optimize the D matrix for multi-knockoffs (see Appendix A). The SDP knockoffs are based on an optimization problem that promotes sparsity: the fact that the objective function is a L^1-distance between the identity matrix and the diagonal matrix D implies that many diagonal terms

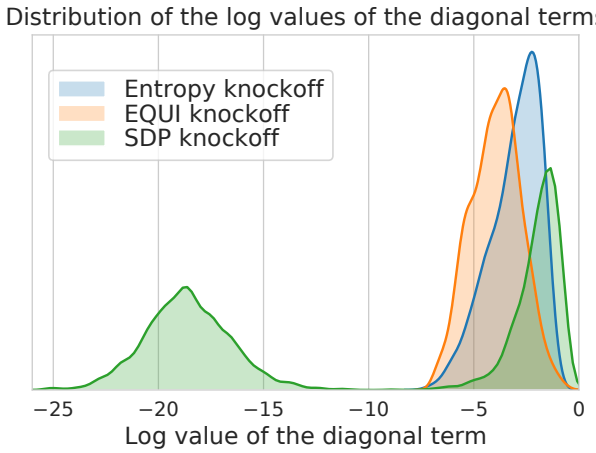

Figure 1: **Comparison between methods for generating Gaussian knockoffs** We plot the densities of the distributions of the diagonal terms generated by each method. The dimension of the covariance matrix is 60.

of the optimal solution will be set almost equal to 0. In addition, Candès et al. (2018) noticed that the equicorrelated knockoff method tends to have very low power, as in high dimensions the diagonal terms of D are proportional to the lowest eigenvalue of the covariance matrix Σ, which in high-dimensional settings tends to be extremely small. Currently, SDP knockoffs are chosen by default. We perform experiments to demonstrate the advantage of entropy over SDP and equicorrelation. We randomly generate correlation matrices with the function *make_spd_matrix* from the Python package scikit-learn (Pedregosa et al., 2011), and compute the diagonal matrix D with the SDP and equicorrelated methods (the diagonal terms of D are necessarily in the interval $(0, 1)$ so that we can compare them across several runs). The lower a diagonal term is, the higher the correlation is between the original feature and its corresponding knockoff, and the less powerful is the knockoff. In Figure 1, we plot the density of the distribution of the logarithm of the diagonal terms of D (that we approximate with the empirical distribution based on 50 runs).

We see that a significant proportion of diagonal terms based on the SDP construction have values extremely small, several orders of magnitude smaller than 10^{-10}, which effectively behave as 0 whenever we sample knockoffs and thus the corresponding features are essentially undiscoverable because their knockoffs are too similar. In Appendix C.1 we show more such comparisons of the distribution of the diagonal terms for varying dimensions of the correlation matrices and strength of the correlation. More concerning in the SDP construction case, the set of almost-zero diagonal terms is very unstable to perturbations in the correlation matrix. We report in Appendix C.2 the simulations proving the instability of such sets. The outcome is simple: the Jaccard similarity between two sets of SDP undiscoverable features generated from two empirical covariance matrices obtained from two batches of i.i.d. samples from the same distribution is on average very low. That is, two parallel runs of the knockoff procedure on different datasets coming from the exact same original distribution lead to different sets of undiscoverable features.

The equicorrelated construction does not suffer such issue, although the diagonal terms tend to be smaller compared to the SDP diagonal terms that are not almost 0. The SDP construction, due to its objective function, maximizes some diagonal terms at the expense of many others that are effectively set to 0, whereas the equicorrelated construction treats all coordinates more equally. Finally, the entropy construction achieves the best performance: the diagonal terms it constructs are generally a couple of orders of magnitude higher than the equicorrelated method, and when comparing to SDP, the entropy construction does not generate almost-zero terms, so that it does not create any catastrophic knockoff. We show in Appendix C.3 a concrete example of dataset $(\boldsymbol{X}, Y)$ where the SDP construction diagonal terms of the non-null features are zero so that SDP knockoffs have power 0, whereas the entropy method achieves almost full power. On top of that the whole procedure will be more stable with entropy knockoffs: there is no longer a highly variable set of undiscoverable features unrelated to the response that restricts the set of possible selections. Also, both methods have equivalent runtimes as they solve similar convex optimization problems (cf. Appendix A).

4 EXPERIMENTS

We first conduct systematic experiments on synthetic data, so that we know the ground truth. For each experiment, we evaluate both the power and the stability of multi-knockoffs and the standard knockoff. Then we evaluate the performance of knockoff on a real set from Genome Wide Association Studies (GWAS).

4.1 Analyzing Improvements with Synthetic Data

We run simulations with synthetic data to confirm the threshold phenomenon and the improvements brought by multi-knockoffs. We randomly generate a feature matrix $\boldsymbol{X}$ from a random covariance matrix, fix a number of non-nulls and create a binary response Y based on a logistic response of a weighted linear combination of the non-null features. Then, we sample multi-knockoffs

with $\kappa = 1$ (single knockoff), $\kappa = 2$ and $\kappa = 3$ from that same $\boldsymbol{X}$ and run the knockoff procedure based on a logistic regression to obtain a selection set, along with values for the power and an FDP. We then repeat this whole procedure 50 times to obtain estimates of the variance and get an empirical FDR. Knockoffs are generated based on the entropy construction to show that our multi-knockoff based improvement is made on top of the entropy improvement (which is only specific for Gaussian knockoffs). The dimension of the feature vectors is 100, and the signal strength is 5. Changing the signal strength affects power but does not change the comparative behavior that we observe between single and multi-knockoffs.

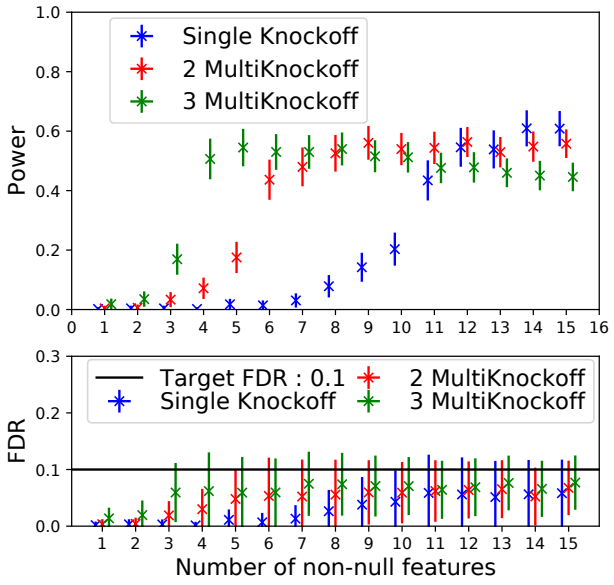

Figure 2: **Power and FDR comparison between single knockoffs and multi-knockoffs** 2 and 3 multi-knockoffs has greater power than the standard knockoff when the number of non-nulls is small. All three methods control FDR.

We set our target FDR level at $q = 0.1$, and compare the single knockoff setting with multi-knockoffs (with $\kappa = 2, 3$) over a range of number of non-null features. We report our results in Figure 2. We first point out that FDR is strongly controlled in all the experiments, as expected. We then estimate the threshold values for detection given by the estimates $\lceil \frac{1}{q\kappa} \rceil$: 10 for single knockoffs ($\kappa = 1$), 5 rejections for multi-knockoffs with $\kappa = 2$, and 3.3 whenever $\kappa = 3$. By plotting power as a function of the number of non-nulls we clearly confirm this threshold behavior. All three settings attain a high power regime whenever the number of non-nulls exceeds the expected detection threshold. This shows the advantage of using multi-knockoffs in settings where we expect a priori the number of non-nulls to be small, and want to make sure that our method has a chance of selecting such small set of non-nulls. We also see there is a small price to pay for using multi-knockoffs. Whenever the number of non-null features increases so that we are beyond the detection thresholds, power decreases with the number of multi-knockoffs. This is due to the fact that sampling multi-knockoffs imposes a more stringent constraint to construct the knockoff conditional distribution (cf. Appendix A), and therefore multi-knockoffs can have slightly "worse" power as κ increases.

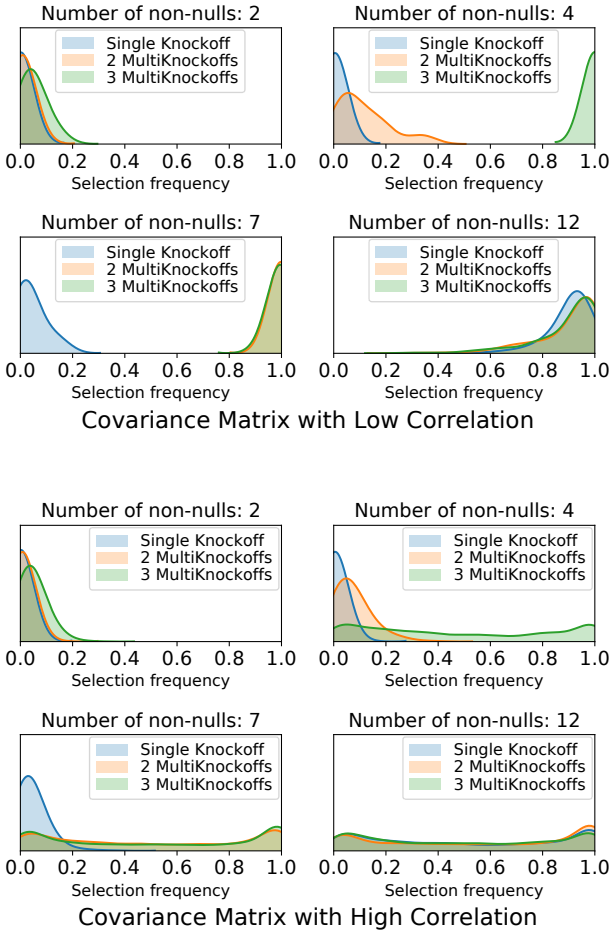

Figure 3: **Improvement in stability with multi-knockoffs: density of non-nulls by selection frequency** When X has low correlation setting (upper figure), and when X has high correlation (lower figure). The x-axis of each plot is the frequency that a non-null feature is selected, and the y axis indicates density.

Finally, multi-knockoffs not only substantially improve the power of the procedure in settings with a small number of non-nulls; they also help stabilizing the procedure. We plot in Figure 3, as a function of the selection frequency, the density of the distribution of

the non-nulls. In order to get the selection frequency, we run the same procedure as before, except this time we sample 200 (multi-)knockoffs out of one same $\boldsymbol{X}$ and run the procedure each time keeping that same $\boldsymbol{X}$ and the response Y we had generated. This allows us to compute how frequently each non-null is selected (by repeatedly sampling knockoffs from a same $\boldsymbol{X}$, FDR is no longer controlled. The point of these simulations is to stress the improvement in stability). For different settings where we vary the number of non-nulls, we see that the multi-knockoffs consistently reject a large fraction of the non-nulls whenever the threshold of detection is attained. In contrast, most non-null features are selected by the standard knockoff at low frequency, indicating instability.

The key aspect here is that the improvement in power whenever a threshold is crossed is not because of an overall increase in selection frequency of all the non-nulls: the densities in the above figures do not concentrate around intermediate selection frequency values. That is, multi-knockoffs do not increase the power by increasing instability.

4.2 Applications: GWAS Causal Variants

We apply our stabilizing procedures for fine mapping the causal variants in a genome wide association study (GWAS). These studies scan the whole genome in search of single nucleotide polymorphisms (SNPs) that are associated with a particular phenotype. In practice, they compute correlation scores for each SNP with respect to the phenotype, and select those beyond a certain significance threshold. Often times, the high correlation between SNPs (called linkage disequilibrium) implies that a large number of consecutive SNPs have a large association score and thus are selected. Fine mapping consists in finding the precise causal SNPs that really help explain the phenotype. Knockoffs can be useful in this setting, but the threshold phenomenon described earlier is an impediment to the application of knockoffs. We want to analyze several dozens, maybe hundreds of SNPs that have passed the selection threshold of the GWAS. However, the number of true causal SNPs may be very low, possibly less than 10. If we set a target FDR level of 0.1, the single knockoff procedure may be unable to make any detection.

We follow the lines of Hormozdiari et al. (2016, 2014) and run simulations analogous to those presented in Figure 2, where the features now correspond to individual genotypes. As it is not possible to actually know, for a given phenotype, which are the true causal SNPs without experimental confirmation, we generate synthetic responses (phenotypes) by randomly choosing a given number of SNPs as causal. Such semi-synthetic data (real X and simulated Y) is standard in literature

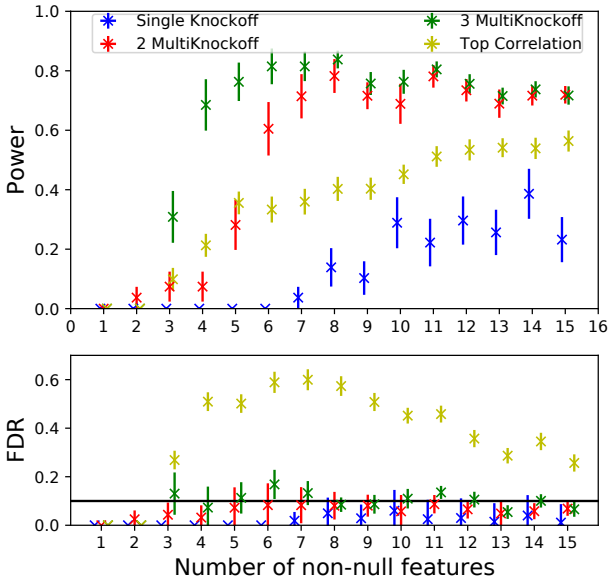

Figure 4: **Power and FDR comparison between single knockoffs, multi-knockoffs, and top correlation for a GWAS dataset.**

(Hormozdiari et al., 2014). In addition, we run a selection procedure without statistical guarantees that is commonly used: we pick the top correlated SNPs with the response. We give more details in Appendix C.4, and detail the impact of the approximation assumptions on the observed FDR. We recover the results obtained with synthetic data and report them in Figure 4: FDR is controlled with the multi-knockoff procedure, and the top correlation method fails to control FDR. We also observe the detection threshold effect: for a low number of causal SNPs, single knockoffs have almost no power, and multi-knockoffs have better power than picking the most correlated SNPs.

5 DISCUSSION

In this paper, we propose multi-knockoffs, an extension of the standard knockoff procedure. We show that multi-knockoff guarantees FDR control, and demonstrate how to generate Gaussian multi-knockoffs via a new entropy-maximization algorithm. Our extensive experiments show that multi-knockoffs are more stable and more powerful compared to the standard (single) knockoff. Finally we illustrate on the important problem of identifying causal GWAS mutations that multi-knockoff substantially outperforms the popular approach of selecting mutations with the highest correlation with the phenotype. The main contribution of this paper is in proposing the mathematical framework of multi-knockoffs; additional empirical analysis and applications is an important direction of future work.

Acknowledgments

J.R.G. was supported by a Stanford Graduate Fellowship. J.Z. is supported by a Chan–Zuckerberg Biohub Investigator grant and National Science Foundation (NSF) Grant CRII 1657155. The authors thank Emmanuel Candès and Nikolaos Ignatiadis for helpful discussions.

References

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

Colin L Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

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)*, 2018.

Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.

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

J. Dahl M. S. Andersen and L. Vandenberghe. CVXOPT: A Python package for convex optimization, version 1.1.5. 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, pages genetics–114, 2014.

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

Lihua Lei, Aaditya Ramdas, and William Fithian. Star: A general interactive framework for FDR control under structural constraints. *arXiv preprint arXiv:1710.02776*, 2017.

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

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526 (7571):68, 2015.