
Discovering Conditionally Salient Features with Statistical Guarantees

Jaime Roquero Gimenez¹ James Zou²

Abstract

The goal of feature selection is to identify important features that are relevant to explain an outcome variable. Most of the work in this domain has focused on identifying *globally* relevant features, which are features that are related to the outcome using evidence across the entire dataset. We study a more fine-grained statistical problem: *conditional feature selection*, where a feature may be relevant depending on the values of the other features. For example in genetic association studies, variant A could be associated with the phenotype in the entire dataset, but conditioned on variant B being present it might be independent of the phenotype. In this sense, variant A is globally relevant, but conditioned on B it is no longer locally relevant in that region of the feature space. We present a generalization of the knockoff procedure that performs *conditional feature selection* while controlling a generalization of the false discovery rate (FDR) to the conditional setting. By exploiting the feature/response model-free framework of the knockoffs, the quality of the statistical FDR guarantee is not degraded even when we perform conditional feature selections. We implement this method and present an algorithm that automatically partitions the feature space such that it enhances the differences between selected sets in different regions, and validate the statistical theoretical results with experiments.

1. Introduction

Given d features $\mathbf{X} = (X_1, \dots, X_d)$ and a response variable Y , in many settings only a small subset of X_i are relevant for Y . The goal of global statistical feature selection is to identify those relevant features, while ideally provid-

ing some statistical control on the rate of false discoveries. This problem has been extensively studied in the statistics and machine learning literature (Benjamini & Hochberg, 1995; Holm, 1979; Benjamini & Yekutieli, 2001; Efron, 2012). One scientific domain that requires extensive use of feature selection is genomics. There Y corresponds to a phenotype of interest (e.g. whether a patient has a disease or not), and \mathbf{X} contains information about individual mutations at genomic sites with high variability, or single nucleotide polymorphisms (SNPs). Researchers are interested to identify the small subset of features/SNPs that affect the disease risks based on information gathered from cohorts of participants with and without the phenotype of interest in genome-wide association studies (GWAS) (Hirschhorn & Daly, 2005; McCarthy et al., 2008).

We call this *global* feature selection because its goal is to identify the relevant features for Y regardless of their location in the feature space. GWAS looks for the subset of SNPs that best help to predict the phenotype, regardless of the actual values of the SNPs themselves. The purpose of this work is to extend the feature selection problem to a *conditional* feature selection problem, where we want to identify relevant features conditionally on being on a sub-region of the feature space. One motivation for this comes again from genomics. Given a SNP (or a set of SNPs) that we believe is relevant for understanding a disease based on prior biological knowledge, or from a previous GWAS study, we want to run a GWAS conditionally on such SNP having a prescribed value. As an example, a few SNPs in a few loci have been identified as major potential causes breast cancer, such as the mutations in the BRCA1/2 genes (O'donovan & Livingston, 2010). We would like to run a GWAS conditioned on not having these variants to identify the other sources of genetic variation that are related to breast cancer.

There are two main advantages when doing *conditional* feature selection. First, a feature that is selected through a global feature selection procedure may no longer be relevant conditioned on a local scale. In genomics, the presence or absence of a mutation can activate or deactivate different biological pathways. Therefore another mutation can be relevant or not for a given phenotype depending on whether its pathway was activated by the first mutation. Being able to distinguish between globally and locally relevant features

¹Department of Statistics, Stanford University, Stanford USA ²Department of Biomedical Data Science, Stanford University, Stanford USA. Correspondence to: James Zou <jamesz@stanford.edu>.

is therefore essential for understanding the complex interactions between features and outcome. Also, by focusing on a local region, one could increase power to identify locally relevant features, which could potentially remain undiscovered if running a global feature selection procedure.

Global feature selection procedures are unable to do this kind of conditional feature selection while preserving statistical guarantees (Sun et al., 2010; Tang et al., 2014). Statistical models that allow for inference while conditioning on some value of the features are very limited, relying upon strong assumptions such as simple linear models that are generally not appropriate for high-dimensional settings of interest. A second approach for conditional feature selection is to restrict the initial dataset to points lying in the region of interest, so that the outcomes of the selection procedure are only related to such region. However, this approach substantially reduces the amount of data available to fit the model, leading to more variance in the estimates. Furthermore, this undermines the validity of the assumptions upon which the construction of p-values relies, so that statistical false discovery guarantees are even less likely to hold.

Our Contributions We develop a new method which extends recently developed knockoff procedure to perform conditional feature selection while still guaranteeing that the local false discovery rate (defined in Sec. 2) is controlled. By using the whole dataset to fit our model of the feature distribution $P^{\mathbf{X}}$ during the first step of the procedure, we make sure that the FDR control for our local knockoff method is as valid as in the usual global method. We extend the knockoff machinery so that we are able to localize the second step of the procedure, where we construct local importance scores for each feature. The selection sets are obtained conditionally on being in a subregion of the space. The main contributions of this paper are in laying out the new framework for conditional feature selection along with proposing a new knockoff algorithm with mathematical guarantees. We also validate the algorithm on experiments.

2. Local Null Features and Local False Discovery Rate

We refer to conditional feature selection the task of identifying a relevant subset of features that explain the response conditionally on being in some subspace of the feature space. As such, we need to define a notion of null feature conditionally on being on a subspace. We will refer to such generalization of null feature as *local* null features; we will use *local* and *conditional* interchangeably to emphasize the conditioning on a local region of the feature space.

Preliminaries Given a positive integer d , let $[d] = \{1, \dots, d\}$ and $\mathcal{P}([d])$ denote the collection of all subsets

of $[d]$. For $x \in \mathbb{R}^d$, $r > 0$, let $B(x, r)$ be the ball centered at x of radius r for the sup norm. For any two mappings $\alpha, \beta : \mathbb{R}^d \rightarrow \mathcal{P}([d])$, we write $\alpha \subset \beta$ (resp. $\alpha \setminus \beta$) if $\alpha(\mathbf{x}) \subset \beta(\mathbf{x})$ (resp. $\alpha(\mathbf{x}) \setminus \beta(\mathbf{x})$) for all $\mathbf{x} \in \mathbb{R}^d$. We begin by introducing the usual setting of feature selection procedures. We consider the data $(\mathcal{X}, \mathcal{Y}) = (\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ as a sequence of n i.i.d. samples from some unknown joint distribution: $(\mathbf{X}_i, Y_i) = (X_{i1}, \dots, X_{id}, Y_i) \sim P^{\mathbf{X}Y}$, $i = 1, \dots, n$. We then define the set of null features $\mathcal{H}_0 \subset [d]$ such that $j \in \mathcal{H}_0$ if $X_j \perp\!\!\!\perp Y | \mathbf{X}_{-j}$ (where the $-j$ subscript indicates all variables except the j th, bold letters indicate vectors, and double-struck letters for the data matrix of stacked vectors). The set of non-null features $\mathcal{H}_1 = [d] \setminus \mathcal{H}_0$, also called alternatives, is important because these capture the truly influential effects and the goal of selection procedures is to identify them. We will denote by (\mathbf{X}, Y) the random variables whenever we make probabilistic statements.

Local Null Features We now consider a local notion of null feature, that is, we are interested in whether any given feature has an impact on the outcome conditionally on the feature set \mathbf{X} being on the neighborhood of some point $\mathbf{x} \in \mathbb{R}^d$. Indeed, the set of relevant features may depend on which region of the feature space we are in. We generalize the definition of a null feature as follows:

Definition 2.1. *Define mappings*

$$\mathcal{H}_0^0 : \begin{cases} \mathbb{R}^d \rightarrow \mathcal{P}([d]) \\ \mathbf{x} \mapsto \mathcal{H}_0^0(\mathbf{x}) \end{cases} \quad \text{and} \quad \mathcal{H}_1^0 = [d] \setminus \mathcal{H}_0^0 \quad (1)$$

such that $j \in \mathcal{H}_0^0(\mathbf{x})$ if $X_j \perp\!\!\!\perp Y | \mathbf{X}_{-j} = \mathbf{x}_{-j}$. We say that the j th feature is a local null at $\mathbf{x} \in \mathbb{R}^d$ if $j \in \mathcal{H}_0^0(\mathbf{x})$. For $r > 0$, define

$$\mathcal{H}_0^r : \begin{cases} \mathbb{R}^d \rightarrow \mathcal{P}([d]) \\ \mathbf{x} \mapsto \mathcal{H}_0^r(\mathbf{x}) = \bigcap_{\mathbf{z} \in B(\mathbf{x}, r)} \mathcal{H}_0^0(\mathbf{z}) \end{cases} \quad (2)$$

We say that the j th feature is a r -local null at $\mathbf{x} \in \mathbb{R}^d$ if $j \in \mathcal{H}_0^r(\mathbf{x})$. Finally, we define the set of r -local non-nulls at \mathbf{x} by $\mathcal{H}_1^r(\mathbf{x}) = [d] \setminus \mathcal{H}_0^r(\mathbf{x})$.

A straightforward generalization of this definition consists in defining local nulls based on a general subset of \mathbb{R}^d , although we will keep this simplified version in what follows. From now on we will assume the joint distribution of (\mathbf{X}, Y) , has a positive density $p(\mathbf{x}, y) > 0$ with respect to a base product measure, so that we can write the conditional distribution of $Y | \mathbf{X}$ as $p(y | \mathbf{x})$. Saying j is a (global) null (i.e. $X_j \perp\!\!\!\perp Y | \mathbf{X}_{-j}$) is equivalent to $p(y | \mathbf{x})$ not depending on x_j . In contrast, j is a local null at \mathbf{x} indicates that, if we fix the values of \mathbf{x}_{-j} , then the expression $p(y | \mathbf{x}) = p(y | x_j, \mathbf{x}_{-j})$ as a function of y and x_j does not depend on x_j . An immediate consequence is that the set of

local nulls does not vary if we move along the null features. Also, identifying local non-nulls at smaller r implies having a more detailed information about the distribution P^{XY} . Our ability to recover the set of local non-nulls for smaller r will be limited by the availability of data in the r -radius neighborhood of \mathbf{x} . We present these results more formally as follows:

Proposition 2.1. *For any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,*

$$\mathbf{x}_{[d] \setminus \mathcal{H}_0^0(\mathbf{x})} = \mathbf{z}_{[d] \setminus \mathcal{H}_0^0(\mathbf{z})} \Rightarrow \mathcal{H}_0^0(\mathbf{x}) = \mathcal{H}_0^0(\mathbf{z})$$

If $r' > r$, then $\mathcal{H}_0^\infty = \mathcal{H}_0 \subset \mathcal{H}_0^{r'} \subset \mathcal{H}_0^r$ the set of global nulls. Equivalently, $\mathcal{H}_1^r \subset \mathcal{H}_1^{r'}$.

Proof. If j is a local null at \mathbf{x} , then $p(y|\mathbf{x})$ does not depend on x_j when fixing \mathbf{x}_{-j} . Therefore j is still a local null whenever we change its value provided the other coordinates are fixed. The first assertion is a consequence of this, extended to the whole set of nulls. The second assertion is immediate following the definition $\mathcal{H}_0^r(\mathbf{x}) = \bigcap_{\mathbf{z} \in B(\mathbf{x}, r)} \mathcal{H}_0^0(\mathbf{z})$. \square

Local False Discovery Rate The output of a global feature selection procedure is a data-dependent set of selected features $\hat{S} \subset \{1, \dots, d\}$. One popular statistical criterion to evaluate the performance of the procedure is the False Discovery Rate (FDR), which stands for the expected value of the proportion of false discoveries: $\text{FDR} = \mathbb{E} \left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|} \right]$.

The ratio $\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|}$ is also called False Discovery Proportion (FDP). Given a prescribed target $q \in (0, 1)$, we want to guarantee that our procedure's FDR is upper bounded by q . If we now switch to a local/conditional perspective, we are now looking for procedures that construct mappings $\mathbf{x} \mapsto \hat{S}(\mathbf{x})$ where we want $\hat{S}(\mathbf{x})$ to match the set of local nulls at \mathbf{x} . The criterion to evaluate these new selection mapping procedures should also be point-dependent. We can now accordingly extend the definition of FDP and FDR to the local setting.

Definition 2.2. *For $r > 0$, and a given feature selection procedure that for $\mathbf{x} \in \mathbb{R}^d$ outputs a data-dependent set of selected features $\hat{S}(\mathbf{x})$, we define the r -local False Discovery Proportion (FDP r) and r -local False Discovery Rate (FDR r) as the following functions of $\mathbf{x} \in \mathbb{R}^d$:*

$$\text{FDP}^r(\mathbf{x}) = \frac{|\hat{S}(\mathbf{x}) \cap \mathcal{H}_0^r(\mathbf{x})|}{1 \vee |\hat{S}(\mathbf{x})|} \quad \text{FDR}^r(\mathbf{x}) = \mathbb{E}[\text{FDP}^r(\mathbf{x})]$$

The randomness in the FDP r expression comes from the fact that the construction of $\hat{S}(\mathbf{x})$ depends on the random dataset of observations (\mathbb{X}, \mathbb{Y}) . The motivation for such criterion is the following: in a global feature selection procedure with FDR control, we want to construct \hat{S} as close as possible from \mathcal{H}_1^∞ , while penalizing whenever $j \in \hat{S} \cap \mathcal{H}_0^\infty$. If we

want to have a more granular approximation of the local non-nulls at \mathbf{x} by constructing $\hat{S}(\mathbf{x})$, we need to penalize whenever $j \in \hat{S}(\mathbf{x}) \cap \mathcal{H}_0^r(\mathbf{x})$. Indeed, given that $\mathcal{H}_0^\infty \subset \mathcal{H}_0^r$ by Proposition 2.1, if we were to construct local selection sets using the global FDR we wouldn't penalize the fact that in a neighborhood of \mathbf{x} a feature j is locally null even though it is non-null at a far away part of the feature space. A new issue that arises from this statistical objective is that we no longer need to control for a single real value, but for a function $\mathbb{R}^d \rightarrow [0, 1]$. From now on, we will focus on controlling point-wise FDR r at a finite number of distinct points far apart, in an independent manner. A future research direction will consist in defining a more comprehensive FDR objective across selection sets at several points in the space.

The strength of the local procedure that we now present is that the validity of the statistical guarantees on the local FDR is as strong as for the global procedure. The ability of making true local discoveries at a given point will be limited by the availability of data in the neighborhood of such point. However the rate of false positives will still be controlled.

3. Controlling Local FDR Through Knockoffs with Local Importance Scores

We now present the global knockoff procedure, and then define the generalization to a local setting through an extension of the importance scores. Our procedure requires the analyst to provide information on the choice of the partitioning of the feature space. We end this section by providing two algorithms based on heuristics to automate this choice.

3.1. The Knockoff Procedure

Traditionally feature selection is based on a parametric model we assume captures the relationship between \mathbf{X} and Y . By fitting such model to the available data, with the appropriate statistical assumptions we can assess the probability of each parameter to be different from 0, and use it as a proxy for the likelihood that the corresponding feature is relevant for predicting Y (Lippert et al., 2011; Miller, 2002; Tibshirani, 1996). These assumptions on the distribution of $Y|\mathbf{X}$ are often inadequate, as the complexity of the distribution of the response given the features is seldom limited to a linear regression (or even to other more complex models such as generalized linear models). This model mismatch can break down the statistical guarantees of the procedure, or drastically reduce its ability to find relevant features (Friedman et al., 2001). Furthermore, in high-dimensional settings the classic distributional results upon which p-values are constructed (the summary value for assessing the statistical relevance of a feature) are no longer valid, so the statistical guarantees for any selection based on those methods may completely break down (Sur & Candès, 2018).

The knockoff procedure is a two-step procedure that has recently emerged which avoids the pitfalls of the model-based global feature selection procedures by shifting the bulk of the statistical hypothesis from modeling $Y|\mathbf{X}$ to modeling the distribution $P^{\mathbf{X}}$ of \mathbf{X} (Barber et al., 2015; Candès et al., 2018). False Discovery Rate control (i.e. control of the number of false positives in expectation) is guaranteed during the first step of the procedure, by assuming only knowledge of $P^{\mathbf{X}}$: during this step the analyst does not make any assumption on the behavior of $Y|\mathbf{X}$, and still FDR is controlled no matter which choices are made downstream. This first step constructs what are called knockoff variables $\tilde{\mathbf{X}}$ that are used as controls when running the second step of the procedure. That second step aims at getting a high number of true selections (referred to as the power of the procedure), which requires the analyst to fit some appropriate procedure to the joint data (\mathbf{X}, Y) among a large pool of options. To summarize, the knockoff procedure decouples the assumptions underlying the statistical guarantee (knowledge of $P^{\mathbf{X}}$ to generate $\tilde{\mathbf{X}}$) from the assumptions needed to have a high power (fitting a model for $Y|\mathbf{X}$).

We now introduce some notation for a more formal presentation of the knockoff procedure. For $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^d$, we define the concatenated vector by $[\mathbf{x}, \tilde{\mathbf{x}}] \in \mathbb{R}^{2d}$, and the following swap operation: given a subset $S \subset [d]$ of indices, the vector $[\mathbf{x}, \tilde{\mathbf{x}}]_{\text{swap}(S)} \in \mathbb{R}^d \times \mathbb{R}^d$ corresponds to $[\mathbf{x}, \tilde{\mathbf{x}}]$ where the coordinates indexed by S are swapped between \mathbf{x} and $\tilde{\mathbf{x}}$ (we use the same notation whenever the vectors are stacked in a matrix). Also, for any subset of indices $S \subset [d]$ we denote by $\mathbf{x}_S \in \mathbb{R}^{|S|}$ the restriction of \mathbf{x} to coordinates in S . For a set A , denote $\#A$ its size.

Assuming we know the ground truth for the distribution $P^{\mathbf{X}}$, the first step of the knockoff procedure is to obtain a *knockoff sample* $\tilde{\mathbf{X}}$ that satisfies the following conditions (Candès et al., 2018):

Definition 3.1 (Knockoff sample). *A knockoff sample $\tilde{\mathbf{X}}$ of a d -dimensional random variable \mathbf{X} is a d -dimensional random variable such that two properties are satisfied:*

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

$$[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)} \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}] \quad \forall S \subset \{1, \dots, d\}$$

where $\stackrel{d}{=}$ stands for equality in distribution.

The first condition is immediately satisfied as long as knockoffs are sampled conditioned on the sample \mathbf{X} without considering any information about Y . Satisfying the exchangeability condition is more technical and has recently received widespread attention (Sesia et al., 2017; Jordon et al., 2019; Gimenez et al., 2018; Liu & Zheng, 2018; Romano et al., 2018). Indeed satisfying the exchangeability condition is the cornerstone of the knockoff procedure upon

which the FDR control relies, thus the importance of having accurate mechanisms to generate valid knockoffs.

The next step requires the analyst to apply some procedure on the dataset $[\mathbb{X}, \tilde{\mathbb{X}}], \mathbb{Y}$ such that it constructs *importance scores* $[\mathbf{T}, \tilde{\mathbf{T}}] \in \mathbb{R}_+^d \times \mathbb{R}_+^d$: a large value for T_j (or \tilde{T}_j) indicates that the procedure considers that the feature X_j (or \tilde{X}_j) has a strong influence on Y . Notice that here we consider the full dataset of i.i.d. samples \mathbb{X} , and the corresponding dataset of knockoff variables $\tilde{\mathbb{X}}$, that are generated sample-wise from \mathbb{X} . The importance scores are random variables constructed based on the whole set of samples. Some examples consist in taking the absolute values of the coefficients in a linear or logistic regression, the time of first entry of a feature in a LASSO path in a L^1 -regularized regression, the drops in accuracy in a trained classifier by perturbing one feature at a time (Barber et al., 2015; Candès et al., 2018; Gimenez et al., 2018; Lu et al., 2018). The only crucial requirement on such importance scores to be valid (for the knockoff procedure) is that they are *associated* to their corresponding feature, and agnostic to the ordering of the remaining features. That is, for any $S \subset [d]$, if the analyst feeds the procedure with the dataset $[\mathbb{X}, \tilde{\mathbb{X}}]_{\text{swap}(S)}, \mathbb{Y}$, then the resulting importance scores are $[\mathbf{T}, \tilde{\mathbf{T}}]_{\text{swap}(S)}$.

The strength and flexibility of the knockoff procedure is that there is no need to assume any model on the joint distribution of (\mathbf{X}, Y) , and that we are free to choose any method that generates valid importance scores: the FDR control will hold because the knockoff importance scores \tilde{T}_j will serve as controls for the T_j . For a target FDR level $q \in (0, 1)$, the output of the procedure is the set $\hat{\mathcal{S}} = \{j : (T_j - \tilde{T}_j) \geq \hat{\tau}\}$ based on a data-driven threshold

$$\hat{\tau} = \min \left\{ t > 0 : \frac{1 + \#\{j : (T_j - \tilde{T}_j) \leq -t\}}{\#\{j : (T_j - \tilde{T}_j) \geq t\}} \leq q \right\} \quad (3)$$

The j th feature is selected when the difference between the importance scores of X_j and \tilde{X}_j is larger than $\hat{\tau}$.

Why subsetting the data first does not work The first naive way to extend the knockoff procedure to a local setting is to run the whole procedure after restricting the initial dataset. For $\mathbf{x} \in \mathbb{R}^d$, if we are interested in identifying the set of local non-nulls $\mathcal{H}_1^r(\mathbf{x})$, we could run the knockoff procedure starting with the set $\{(\mathbf{X}_i, Y_i), s.t. \|\mathbf{X}_i - \mathbf{x}\|_\infty \leq r\}$. The reason why this approach is difficult to implement is that generating $\tilde{\mathbf{X}}$ for the conditional distribution of \mathbf{X} being in a neighborhood of \mathbf{x} is generally not tractable. Restricting the amount of data harms how well we can fit our model of $P^{\mathbf{X}}$ to the data. Moreover conditioning on the values of \mathbf{X} requires a different process for sampling the knockoffs $\tilde{\mathbf{X}}|\mathbf{X}$ than what is currently available (Barber et al., 2015; Candès et al., 2018; Gimenez et al., 2018; Lu et al., 2018). In order to construct

Algorithm 1 Knockoffs with Local Importance Scores Feature Selection Procedure

Input: Dataset \mathcal{X}, \mathcal{Y} , set of L points $\mathcal{L} = \{z_l\}_{1 \leq l \leq L}$, radius $r > 0$.

Generate dataset of knockoff variables $\tilde{\mathcal{X}}$.

for $l = 1$ **to** L **do**

 Subsample data $(\mathcal{X}^{(l)}, \tilde{\mathcal{X}}^{(l)}, \mathcal{Y}^{(l)}) = \{(\mathbf{X}_\cdot, \tilde{\mathbf{X}}_\cdot, Y)\}$ such that $\|\mathbf{X}_\cdot - z_l\|_\infty \leq r$ & $\|\tilde{\mathbf{X}}_\cdot - z_l\|_\infty \leq r$

 Generate (r -local) importance scores $\mathbf{T}^{(l)}, \tilde{\mathbf{T}}^{(l)}$ from $(\mathcal{X}^{(l)}, \tilde{\mathcal{X}}^{(l)}, \mathcal{Y}^{(l)})$

 Construct $\hat{\tau}^{(l)}$ as in equation 3 from $\mathbf{T}^{(l)}, \tilde{\mathbf{T}}^{(l)}$

 Construct $\hat{\mathcal{S}}_l = \{j \in [d] : \mathbf{T}_j^{(l)} - \tilde{\mathbf{T}}_j^{(l)} \geq \hat{\tau}^{(l)}\}$

end for

Return: Selected sets $(\hat{\mathcal{S}}_l)_{1 \leq l \leq L}$

local selection sets at a given point \mathbf{x} , we need to keep the original knockoff generation process intact, but construct local importance scores. Keeping the statistical guarantee in place requires reformulating the whole knockoff procedure in a local approach.

3.2. Knockoff with Local Importance Scores

Our extension of the knockoff procedure assumes that we have valid knockoffs available for the whole dataset: that is, the first step of constructing knockoffs is exactly the same as in the global knockoff procedure. By using the whole available dataset, by making the same assumptions as in the global knockoff setting, we can reuse all the available tools already developed for building valid knockoffs, and will not leave out any useful information which could hurt the validity of the statistical guarantees.

Fix $r > 0$, consider a set of L points z_1, \dots, z_L whose pairwise distances are lower bounded by $2r$. We now present Algorithm 1, a procedure that constructs L sets $\hat{\mathcal{S}}_1, \dots, \hat{\mathcal{S}}_L \subset [d]$ of selected features at each point z_l , such that we control FDR^r at any desired FDR target level $q \in (0, 1)$, uniformly for all sets. The crucial aspect of this procedure is that we are able to make a local statement for local FDR uniformly for several points. This algorithm starts by constructing knockoffs $\tilde{\mathcal{X}}$ from \mathcal{X} as in the global knockoff procedure. It then subsets the whole dataset into L datasets $(\mathcal{X}^{(l)}, \tilde{\mathcal{X}}^{(l)}, \mathcal{Y}^{(l)})$ where for each $1 \leq l \leq L$ the samples in the l th subset are such that the original vector $\mathbf{X}^{(l)}$ and the knockoff sample $\tilde{\mathbf{X}}^{(l)}$ are in a r -ball neighborhood of z_l (we denote by \mathbf{X}_\cdot a generic sample of the vector \mathbf{X} among the samples in $\mathcal{X}^{(l)}$). We now run construct r -local importance scores independently for each subset, that we use to construct a selected set of features as in the global knockoff procedure. As for the global knockoff procedure, a wise choice of importance scores may lead to a higher power of the procedure, i.e. a correct modelling of the re-

lationship between the original features and the response for the dataset $(\mathcal{X}^{(l)}, \tilde{\mathcal{X}}^{(l)}, \mathcal{Y}^{(l)})$. As long as the importance scores satisfy the association requirement stated above, one can use different methods for each l : we can use the absolute coefficients when fitting a logistic regression (Candès et al., 2018) for one of the subsets if we know it is a good model for that subset, and a more generic accuracy drop (Gimenez et al., 2018) for another region where we may lack such knowledge.

Theorem 3.1. *The output of Algorithm 1 is such that FDR^r is controlled uniformly for z_l . That is, denoting for each $1 \leq l \leq L$,*

$$FDP^r(z_l) = \frac{|\hat{\mathcal{S}}_l \cap \mathcal{H}_0^r(z_l)|}{1 \vee |\hat{\mathcal{S}}_l|}$$

we have that:

$$\sup_l \mathbb{E}[FDP^r(z_l)] \leq q$$

In order to prove this result we need to show that the usual knockoffs can be reused in a local way: this requires to redefine and extend the distributional properties underlying the proof of FDR control from the global knockoff procedure as in (Candès et al., 2018). We present the technical tools needed to prove this theorem in Section 5.

3.3. Algorithms for Partitioning the Feature Space into Meaningful Sub-regions

We now present the last element needed to obtain a reasonable local selection procedure. The implementation of Algorithm 1 requires the analyst to indicate the points at which the selection sets need to be computed. Such knowledge may not be available: it even requires to identify some of the non-nulls that are needed to partition the space into sub-regions with different sets of non-nulls. We now present an approach to hierarchically identify relevant "switch" features for the purpose of partitioning the feature space and the values that these switch features take at the boundaries of the partitions, and a second one in Appendix D. These algorithms based on heuristics only provide input values to Algorithm 1. As such, the statistical guarantees described in Theorem 3.1 always hold.

Importance Score-based Partitioning Finding the best "switch" feature and its boundary value can be done in a greedy manner. The output will be relevant depending on the choice of importance scores that we use as a subroutine of this algorithm. Fix a subset $I \subset [d]$ of features, fix a set $\Lambda = (\lambda_1, \dots, \lambda_L) \in \mathbb{R}^L$ of L boundary values. For each pair (c, λ) of feature and split value, we split the full matrix dataset $(\mathcal{X}, \mathcal{Y})$ into two sets $(\mathcal{X}^{>,c,\lambda}, \mathcal{Y}^{>,c,\lambda})$ and $(\mathcal{X}^{<,c,\lambda}, \mathcal{Y}^{<,c,\lambda})$ based on the value of the d th covariate of

Algorithm 2 One-Step Greedy Feature Space Partition

Input: Subset of indices I , Split values Λ , Dataset $\{\mathbb{X}, \mathbb{Y}\}$, radius r .
for d **in** I **do**
 for λ **in** Λ **do**
 Partition the data (\mathbb{X}, \mathbb{Y}) according to $X_{\cdot d} > \lambda$ or $X_{\cdot d} < \lambda$ into $(\mathbb{X}^{>,c\lambda}, \mathbb{Y}^{>,c\lambda})$ and $(\mathbb{X}^{<,c\lambda}, \mathbb{Y}^{<,c\lambda})$.
 Generate importance scores $\mathbf{T}^{>,c\lambda}$ and $\mathbf{T}^{<,c\lambda}$.
 Compute $gap_{d,\lambda} = \|\mathbf{T}^{>,c\lambda} - \mathbf{T}^{<,c\lambda}\|_\infty$
 end for
end for
 Select $d^*, \lambda^* = \arg \max gap_{d,\lambda}$.
if Final recursive step **then**
 Return $z_{<}, z_{>}$ s.t. $z_{<,d^*} \leq \lambda^* - r$ and $z_{>,d^*} \geq \lambda^* + r$.
else
 Return $(\mathbb{X}^{>,c\lambda}, \mathbb{Y}^{>,c\lambda})$ and $(\mathbb{X}^{<,c\lambda}, \mathbb{Y}^{<,c\lambda})$.
end if

each sample \mathbf{X}_i and compute importance scores $\mathbf{T}^{>,c\lambda}$ and $\mathbf{T}^{<,c\lambda}$ independently on each (notice that importance scores do not require the presence of knockoffs: moreover, importance scores are constructed based on a set of features agnostic to whether those features were original or knock-off). As a method to generate importance scores, we will train a classifier on the dataset and compute, for each feature, the accuracy drops when reevaluating the trained classifier on the same initial data except for one feature that we shuffle across samples. We then select the optimal split as the one with highest gap in importance scores in L^∞ norm between the two splits. We can then recursively, depending on the amount of data, run the algorithm on each of the splits, or, if it corresponds to the final step, return a set of vectors lying in distinct subregions, spaced by at least $2r$.

The main disadvantage of this method is the computational cost. Indeed, for a given subset of relevant features, identifying the optimal choice for the switch feature and the boundary value is quadratic in the size of the subset. As such we may want to choose an appropriate subset I and set Λ as inputs for Algorithm 2 (any global feature selection procedure can be used to identify an appropriate set I). Given that we need to repeatedly fit the model, this method may become prohibitive whenever the number of relevant features is too high (i.e. non sparse settings or high-dimensional settings). We present an alternative approach to overcome computational bottlenecks in Appendix D.

4. Experiments

We now run experiments and the first goal is to show that our main theorem holds. Among the many available distributions for $P^{\mathbf{X}}$ that allow for efficient sampling of knockoffs, we choose the Hidden Markov Model (HMM) (Sesia et al.,

2017) to illustrate a GWAS study. We therefore sample the datasets $\mathbb{X}, \tilde{\mathbb{X}}$ so that \mathbb{X} simulates the SNPs of a cohort of patients, i.e. a matrix of 0, 1, and 2. We generate a complex response Y in the following way: we randomly pick a subset of q features $\mathcal{S}_0 = \{s_1, \dots, s_q\}$ that correspond to the global non-nulls. However, we design the response in such way that not all global non-nulls are local non-nulls at every point. Our design of the response begins by choosing switch features: based on the value of X_{s_1} (greater or lower than 1.5) we select the first or second half of the remaining global non-null indices $\{s_2, \dots, s_q\}$, and we repeat one more time this process (i.e. for example whenever $X_{s_1} > 1.5$ look at whether $X_{s_2} > 1.5$, in which case we end up picking the first half of $\{s_3, \dots, s_{q/2}\}$). We end up, for each individual sample $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,d})$ in the dataset \mathbb{X} , with a subset of indices $\mathcal{S}_i \subset \mathcal{S}_0 \subset [d]$ (in the case where $X_{i,s_1}, X_{i,s_2} > 1.5$, we have then $\mathcal{S}_i = \{s_3, \dots, s_{q/4}\}$). We then generate Y_i as a Bernoulli whose probability is given by a linear combination of the values $\mathbf{X}_{i,\mathcal{S}_i}$ (i.e. can be modeled through a logistic regression on that subset of the space). This way, the set of local non-nulls varies depending on the values of the switch features.

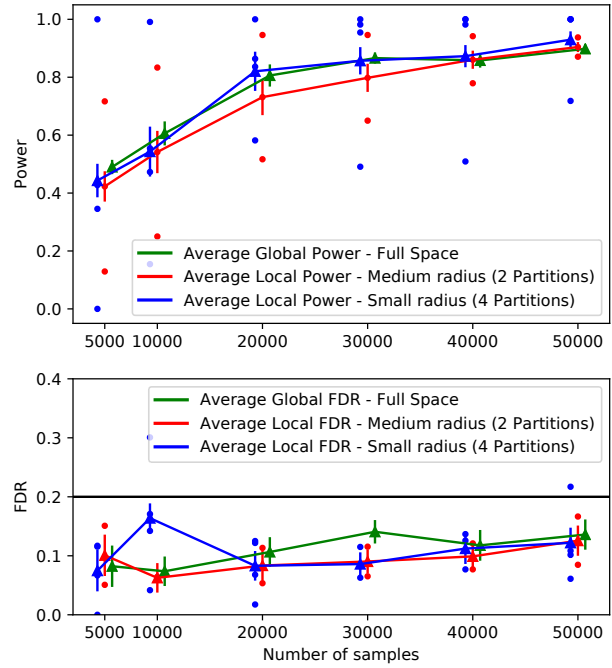


Figure 1. Local Power and Local FDR as a Function of Sample size. For several partitions, the dots indicate the (averaged over 20 runs) local power (top) / local FDR (bottom) at each of the partitions (1, 2 or 4). The lines correspond to the averages across partitions and 20 runs, with estimates of the standard errors. We report the target FDR at 0.2 with a horizontal line.

We now run Algorithm 1 at three "resolution levels", i.e. for $L = 1, 2$ and 4 points $(z_l)_{1 \leq l \leq L}$ (so that we respectively

construct 1,2,4 selection sets $(\hat{S}_l)_{1 \leq l \leq L}$, using each time an optimal choice of points that we assume are given and correspond to the optimal partitions in our response generating process. Our target FDR is $q = 0.2$ and locally we consider as importance scores the absolute values of the coefficients in a logistic regression. We report the results in Figure 1, where the dots refer (for each resolution level) to the local FDR/power at each individual set of the partition. It is crucial to notice that here the FDR/Power scale is related to the procedure: as we increase the partition we consider local FDR/Power on a smaller scale: at the highest resolution, i.e. $L = 4$, for the point z_l corresponding to the region $X_{s_1}, X_{s_2} > 1.5$, we will evaluate the FDR and power of the output \hat{S}_l of the procedure locally with respect to $\mathcal{S}_{0l} = \{s_3, \dots, s_{q/4}\}$. Our procedure is able, in every setting, *to retrieve the correct subset of non-nulls at the corresponding scale while controlling for FDR at the corresponding scale*. That is, the run with just 1 point corresponds to the usual global feature selection problem and selects features that are global nulls. Whenever we choose 2 or 4 points, we output several selection sets that correspond to the local nulls on the corresponding regions.

However to appreciate the real interest of the local feature selection problem one needs to consider what happens to the features that are non-null in a subspace but null in another (for example X_{s_3} is non-null only whenever $X_{s_1}, X_{s_2} > 1.5$). The fact that our procedure controls local FDR indicates that, at a given subspace, it is properly identifying the true local non-nulls, and *not selecting features that are null in that subspace, albeit non-null somewhere else*. Figure 2 illustrates this phenomenon: we run again the procedure (for different values of L) but report the averaged local FDR across the four partitions whenever $L = 4$. This indicates that, if we were only interested in a phenomenon in a subregion of the feature space, the output of the global feature selection procedure would report a large proportion of features that are actually irrelevant in the corresponding subregion.

Notice also that our simulations were done in dimension 200, and we vary the size of the dataset to show that our method still controls local FDR even though the number of points in a given subregion is very limited: we indeed have lower power in these settings, but our statistical guarantees always hold.

5. Theoretical Analysis of Local Feature FDR

Our main result presented in Theorem 3.1 holds regardless of the method used to partition the space. It is based upon a generalization of several notions defined in the first introductions of the knockoff procedure (Barber et al., 2015; Candès et al., 2018) which we now present.

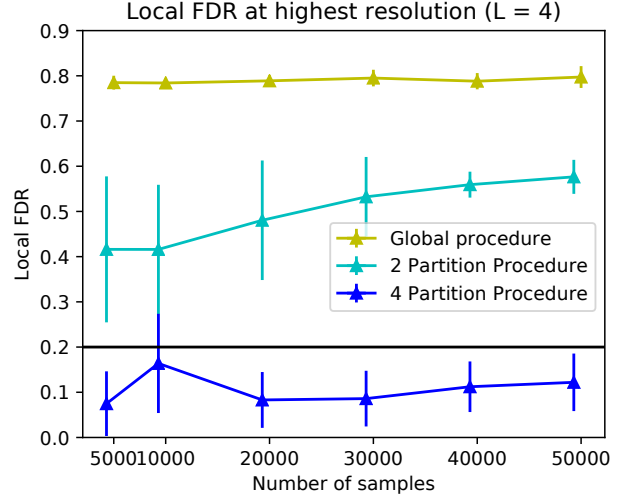


Figure 2. **Local FDR at highest resolution for procedures at different resolutions.** Whenever we evaluate the results of a low-resolution procedure (global feature selection) with a high resolution FDR (small radius with 4 partitions), the coarser methods can not control the finer FDR.

Local swaps and Local Exchangeability We extend the definition of the swap to mappings from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d \times \mathbb{R}^d$, where the set of indices S used to swap depends on the inputs of the mapping.

Definition 5.1. We call a mapping $\sigma : \mathbb{R}^d \rightarrow \mathcal{P}([d])$ a generic swap or a swap. In addition, we say that a swap is a local swap if for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$\mathbf{x}_{[d] \setminus \sigma(\mathbf{x})} = \mathbf{z}_{[d] \setminus \sigma(\mathbf{x})} \Rightarrow \sigma(\mathbf{x}) = \sigma(\mathbf{z})$$

Given a mapping

$$(\mathbf{F}, \tilde{\mathbf{F}}) : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d \\ (\mathbf{x}, \tilde{\mathbf{x}}) \mapsto \begin{cases} (F_1(\mathbf{x}, \tilde{\mathbf{x}}), \dots, F_d(\mathbf{x}, \tilde{\mathbf{x}}), \\ \tilde{F}_1(\mathbf{x}, \tilde{\mathbf{x}}), \dots, \tilde{F}_d(\mathbf{x}, \tilde{\mathbf{x}})) \end{cases} \end{cases}$$

and swap σ , define the operation $[\mathbf{F}, \tilde{\mathbf{F}}]_{\text{swap}(\sigma)}$ as the mapping $[\mathbf{F}, \tilde{\mathbf{F}}]_{\text{swap}(\sigma)} : (\mathbf{x}, \tilde{\mathbf{x}}) \mapsto [(\mathbf{F}, \tilde{\mathbf{F}})(\mathbf{x}, \tilde{\mathbf{x}})]_{\text{swap}(\sigma(\mathbf{x}))}$.

We develop further this notion and additional notation in Appendix A. Our goal now is to show that the exchangeability condition that defines a knockoff variable implies a stronger distributional result.

Proposition 5.1. Let σ be a local swap. If $\tilde{\mathbf{X}}$ is a knockoff random variable for \mathbf{X} (i.e. satisfies exchangeability), then

$$[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\sigma)} \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}] \quad (4)$$

which we refer to as local exchangeability. If $\sigma \subset \mathcal{H}_0^0$, then

$$[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\sigma)}, Y \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}], Y \quad (5)$$

We prove this result in Section B.1. This result extends the exchangeability property and the Lemma 3.2 in (Candès et al., 2018). Instead of swapping a fixed set of features, we now allow the swapping indices to depend on the features. Notice that the knockoffs $\tilde{\mathbf{X}}$ are constructed as in the general case, the local exchangeability does not require a different definition for knockoff variables. We extend the definition of a swap to probability distributions: for $\mu \in Pr(\mathbb{R}^d \times \mathbb{R}^d)$, we denote $\mu_{\text{swap}(\sigma)} := \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\sigma)})$ whenever $\mu = \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}])$. Abusing notation, whenever $\mu = \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}], Y)$ we will still denote $\mu_{\text{swap}(\sigma)} := \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\sigma)}, Y)$.

Local Feature Statistics The next step is to extend the construction of feature statistics to the local setting.

Definition 5.2. Define local importance scores as a mapping:

$$\Phi : \begin{cases} Pr(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}) \longrightarrow (\mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d) \\ \mu \mapsto (\mathbf{T}_\mu, \tilde{\mathbf{T}}_\mu) \end{cases} \quad (6)$$

where

$$(\mathbf{T}_\mu, \tilde{\mathbf{T}}_\mu) : \begin{cases} \mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d \\ \mathbf{z} \mapsto (\mathbf{T}_\mu(\mathbf{z}), \tilde{\mathbf{T}}_\mu(\mathbf{z})) \end{cases} \quad (7)$$

such that, for any $S \subset [d]$, we have

$$\Phi(\mu_{\text{swap}(S)}) = [\Phi(\mu)]_{\text{swap}(S)}$$

For $r > 0$, we say that such importance scores Φ are r -local if, for any $\mu \in Pr(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$, we have that $\Phi(\mu)(\mathbf{z}) = (\mathbf{T}_\mu(\mathbf{z}), \tilde{\mathbf{T}}_\mu(\mathbf{z}))$ only depends on μ through the restriction of μ to $B(\mathbf{z}, r) \times B(\mathbf{z}, r) \times \mathbb{R}$. That is, if μ, μ' are two probability measures on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ such that they coincide on $B(\mathbf{z}, r) \times B(\mathbf{z}, r) \times \mathbb{R}$, then $(\mathbf{T}_\mu(\mathbf{z}), \tilde{\mathbf{T}}_\mu(\mathbf{z})) = (\mathbf{T}_{\mu'}(\mathbf{z}), \tilde{\mathbf{T}}_{\mu'}(\mathbf{z}))$.

The next goal consists in translating the swap operation in $\mu_{\text{swap}(\sigma)} = \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\sigma)}, Y)$ into a swap of $[\mathbf{T}_\mu, \tilde{\mathbf{T}}_\mu]_{\text{swap}(\sigma)}$. This step does not require $\tilde{\mathbf{X}}$ to be a knockoff of \mathbf{X} : in what follows we do not make any assumption on μ . Notice that the swap operation has been defined (Definition 5.1) as a transformation of a mapping $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d$, but it can be immediately extended to mappings $\mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d$. We are able to relate $[\mathbf{T}_{\mu_{\text{swap}(\sigma)}}, \tilde{\mathbf{T}}_{\mu_{\text{swap}(\sigma)}}]$ and $[\mathbf{T}_\mu, \tilde{\mathbf{T}}_\mu]_{\text{swap}(\sigma)}$ if we assume locality constraints on the importance scores.

Definition 5.3. For $\sigma : \mathbb{R}^d \rightarrow \mathcal{P}([d])$ local swap, and $r > 0$, define

$$A^r = \{\mathbf{z} \in \mathbb{R}^d, \sigma(\mathbf{z}) = \sigma(\mathbf{y}) \forall \mathbf{y} \in B(\mathbf{z}, r)\}$$

$$\sigma^r : \mathbf{z} \mapsto \begin{cases} \sigma(\mathbf{z}) & \text{if } \mathbf{z} \in A^r \\ \emptyset & \text{else} \end{cases}$$

Proposition 5.2. Assume there exists $r > 0$ such that the importance scores are r -local. Then, assuming A^r is non-empty, for $\mathbf{z} \in A^r$, we have:

$$[(\mathbf{T}_\mu(\mathbf{z}), \tilde{\mathbf{T}}_\mu(\mathbf{z}))]_{\text{swap}(\sigma(\mathbf{z}))} = [\mathbf{T}_{\mu_{\text{swap}(\sigma)}(\mathbf{z})}, \tilde{\mathbf{T}}_{\mu_{\text{swap}(\sigma)}(\mathbf{z})}]$$

We prove this result in Section B.2. Whenever $\mu = \mathcal{L}([\mathbf{X}, \tilde{\mathbf{X}}], Y)$ where $\tilde{\mathbf{X}}$ is a knockoff of \mathbf{X} , consider a local swap σ such that $\sigma \subset \mathcal{H}_0^0$. By proposition 5.1, we get that $\mu = \mu_{\text{swap}(\sigma)}$, and therefore $(\mathbf{T}_\mu, \tilde{\mathbf{T}}_\mu) = (\mathbf{T}_{\mu_{\text{swap}(\sigma)}}, \tilde{\mathbf{T}}_{\mu_{\text{swap}(\sigma)}})$. In practice, we can imagine that instead of using μ as an input when constructing importance scores, we use $\hat{\mu}_n$, an empirical measure defined by the dataset of n i.i.d. samples from μ that we feed as input to the algorithm. Therefore a consequence of proposition 5.1 will be that, taking also into account the eventual randomness when generating importance scores, we have for any $\mathbf{z} \in \mathbb{R}^d$,

$$(\mathbf{T}_{\hat{\mu}_n}(\mathbf{z}), \tilde{\mathbf{T}}_{\hat{\mu}_n}(\mathbf{z})) \stackrel{d}{=} (\mathbf{T}_{\hat{\mu}_{n, \text{swap}(\sigma)}}(\mathbf{z}), \tilde{\mathbf{T}}_{\hat{\mu}_{n, \text{swap}(\sigma)}}(\mathbf{z}))$$

We now combine this equality with that of Proposition 5.2. Fix $r > 0$ such that we have r -local importance scores. Consider a set of L points $\mathbf{z}_1, \dots, \mathbf{z}_N \in \mathbb{R}^d$ that are pairwise $2r$ far apart, that is, for any $1 \leq l, l' \leq N$, $\|\mathbf{z}_l - \mathbf{z}_{l'}\|_\infty \geq 2r$. We can now conclude thanks to Proposition B.3 in the Appendix, which implies the flip-sign property introduced first in (Candès et al., 2018). This is all that is needed to construct an adaptive threshold τ^l based on $\mathbf{T}_{\hat{\mu}_n}(\mathbf{z}_l), \tilde{\mathbf{T}}_{\hat{\mu}_n}(\mathbf{z}_l)$, and the corresponding selected features \hat{S}_l such that FDR is controlled.

That is, according to Theorem 3.4 in (Candès et al., 2018), the set \hat{S}_l is such that

$$\mathbb{E}\left[\frac{|\hat{S}_l \cap \mathcal{H}_0^r(\mathbf{x})|}{1 \vee |\hat{S}_l|}\right] \leq q$$

hence the result. We provide additional details in Appendix A.

6. Discussion

We develop a new framework for selecting features conditionally on being in a specific region of the feature space, for which we define the notions of local null features and local FDR. The fact that FDR is controlled even though the analyst does not need to model the relationship between the outcome and the feature variables makes the knockoff procedure stand out among feature selection procedures. We extend the usage of the knockoff procedure to deal with our local feature selection problem, in such way that there is no need to strengthen the assumptions required for the statistical guarantees to hold.

ACKNOWLEDGMENTS

J.R.G. is 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.

References

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

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

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Candès, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: ‘model-x’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

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

Efron, B. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA., 2001.

Gimenez, J. R., Ghorbani, A., and Zou, J. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214*, 2018.

Hirschhorn, J. N. and Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95, 2005.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Jordon, J., Yoon, J., and van der Schaar, M. Knockoff-GAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=ByeZ5jC5YQ>.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10): 833, 2011.

Lipton, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Liu, Y. and Zheng, C. Auto-encoding knockoff generator for fdr controlled variable selection. *arXiv preprint arXiv:1809.10765*, 2018.

Lu, Y., Fan, Y., Lv, J., and Noble, W. S. Deepink: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 8690–8700, 2018.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356, 2008.

Miller, A. *Subset selection in regression*. Chapman and Hall/CRC, 2002.

O’donovan, P. J. and Livingston, D. M. Brca1 and brca2: breast/ovarian cancer susceptibility gene products and participants in dna double-strand break repair. *Carcinogenesis*, 31(6):961–967, 2010.

Romano, Y., Sesia, M., and Candès, E. J. Deep knockoffs. *arXiv preprint arXiv:1811.06687*, 2018.

Sesia, M., Sabatti, C., and Candès, E. J. Gene hunting with knockoffs for hidden markov models. *arXiv preprint arXiv:1706.04677*, 2017.

Sun, Y., Todorovic, S., and Goodison, S. Local-learning-based feature selection for high-dimensional data analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1610–1626, 2010.

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

Tang, J., Alelyani, S., and Liu, H. Feature selection for classification: A review. *Data classification: Algorithms and applications*, pp. 37, 2014.

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