Submitted to the Annals of Statistics

ROBUST INFERENCE WITH KNOCKOFFS

By Rina Foygel Barber[*], Emmanuel J. Candès[†], and Richard J. Samworth[‡]

The University of Chicago[*], *Stanford University*[†] *and University of Cambridge*[‡]

We consider the variable selection problem, which seeks to identify important variables influencing a response Y out of many candidate features $X_1, \dots, X_p$. We wish to do so while offering finite-sample guarantees about the fraction of false positives—selected variables X_j that in fact have no effect on Y after the other features are known. When the number of features p is large (perhaps even larger than the sample size n), and we have no prior knowledge regarding the type of dependence between Y and X, the model-X knockoffs framework nonetheless allows us to select a model with a guaranteed bound on the false discovery rate, as long as the distribution of the feature vector $X = (X_1, \dots, X_p)$ is exactly known. This model selection procedure operates by constructing "knockoff copies" of each of the p features, which are then used as a control group to ensure that the model selection algorithm is not choosing too many irrelevant features. In this work, we study the practical setting where the distribution of X can only be estimated, rather than known exactly, and the knockoff copies of the X_j's are therefore constructed somewhat incorrectly. Our results, which are free of any modeling assumption whatsoever, show that the resulting model selection procedure incurs an inflation of the false discovery rate that is proportional to our errors in estimating the distribution of each feature X_j conditional on the remaining features $\{X_k : k \neq j\}$. The model-X knockoffs framework is therefore robust to errors in the underlying assumptions on the distribution of X, making it an effective method for many practical applications, such as genome-wide association studies, where the underlying distribution on the features $X_1, \dots, X_p$ is estimated accurately but not known exactly.

1. Introduction.

Our methods of data acquisition are such that we often obtain information on an exhaustive collection of possible explanatory variables. We know a priori that a large proportion of these are irrelevant for our purposes, but in an effort to cover all bases, we gather data on all what we can measure and rely on subsequent analysis to identify the relevant vari-

Keywords and phrases: knockoffs, variable selection, false discovery rate (FDR), high-dimensional regression, robustness

ables. For instance, to achieve a better understanding of biological processes behind a disease, we may evaluate variation across the entire DNA sequence and collect single nucleotide polymorphism (SNP) information, or quantify the expression level of all genes, or consider a large panel of exposures, and so on. We then expect the statistician or the scientist to sort through all these and select those important variables that truly influence a response of interest. For example, we would like the statistician to tell us which of the many genetic variations affect the risk of a specific disease, or which of the many gene expression profiles help determine the severity of a tumor.

This paper is about this variable selection problem. We consider situations where we have observations on a response Y and a large collection of variables $X_1, \ldots, X_p$. With the goal of identifying the important variables, we want to recover the smallest set $\mathcal{S} \subseteq \{1, \ldots, p\}$ such that, conditionally on $\{X_j\}_{j\in\mathcal{S}}$, the response Y is independent of all the remaining variables $\{X_j\}_{j\notin\mathcal{S}}$. In the literature on graphical models, the set $\mathcal{S}$ would be called the *Markov blanket* of Y. Effectively, this means that the explanatory variables $X_1, \ldots, X_p$ provide information about the outcome Y only through the subset $\{X_j\}_{j\in\mathcal{S}}$. To ensure reproducibility, we are interested in methods that result in the estimation of a set $\widehat{\mathcal{S}}$ with false discovery rate (FDR) control [Benjamini and Hochberg, 1995], in the sense that

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\{j : j \in \widehat{\mathcal{S}} \setminus \mathcal{S}\}}{\#\{j : j \in \widehat{\mathcal{S}}\}}\right] \le q,$$

i.e. a bound on the expected proportion of our discoveries $\widehat{\mathcal{S}}$ which are *not* in the smallest explanatory set $\mathcal{S}$.[1] (Here q is some predetermined target error rate, e.g. $q = 0.1$.)

In truth, there are not many variable selection methods that would control the FDR with finite-sample guarantees, especially when the number p of variables far exceeds the sample size n. That said, one solution is provided by the recent model-X knockoffs approach of Candès et al. [2018], which is a new read on the earlier knockoff filter of Barber and Candès [2015]; see also Barber and Candès [2019]. One singular aspect of the method of model-X knockoffs is that it makes assumptions that are substantially different from those commonly encountered in the statistical literature. Most of the model selection literature relies on a specification of the model that links together the response and the covariates, making assumptions on $P_{Y|X}$, the distribution of Y conditional on X—for instance, assuming that the

[1] As is standard in the FDR literature, in this expected value we treat $0/0$ as 0, to incur no penalty in the event that no variables are selected, i.e. when $\widehat{\mathcal{S}} = \emptyset$.

form of this distribution follows a generalized linear model or some other parametrized model. In contrast, model-X knockoffs makes no assumption whatsoever on the relation between the response Y and the variables $X = (X_1, \ldots, X_p)$; in other words, the distribution $P_{Y|X}$ of Y conditional on X is “model free”. The price of this generality is that we need to be able to specify the distribution of the feature variables $X = (X_1, \ldots, X_p)$, which we denote by P_X. This distribution is then used to construct knockoff feature variables $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$, where each $\widetilde{X}_j$ mimics the real feature X_j and acts a “negative control” in any variable selection algorithm—if our variable selection algorithm selects any of the knockoff features, this alerts us to a high false positive rate in the algorithm. Knowledge of the distribution of X is needed in order to construct the $\widetilde{X}_j$'s appropriately—for instance, if X_1 is a real signal while X_2 is null, then we need $\widetilde{X}_2$ to mimic X_2's dependence with X_1 in order to act as an appropriate negative control.

As argued in Candès et al. [2018] and Janson [2017], this “shift” of the burden of knowledge is interesting because we must recall that the object of inference is on how Y relates to X, that is, on $P_{Y|X}$. It is, therefore, a strong premise to posit the form of this relationship $P_{Y|X}$ a priori—and indeed, there are many applications in which we objectively do not have any understanding of how Y depends on X. Further, the shift is also appropriate whenever we know much more about the distribution of X than on the conditional distribution of $Y \mid X$. For instance, it is easy to imagine applications in which we have many unlabeled samples—samples of X—whereas it may be much harder to acquire labeled data or samples with a given value of the response Y. A typical example is offered by genetic studies, where we now have available hundreds of thousands or even millions of genotypes across many different populations. At the same time, it may be difficult to recruit patients with a given phenotype (the response variable Y), and therefore, we have substantially more data with which to estimate P_X than $P_{Y|X}$.

The ease with which we can gather information about X does not imply that we know the distribution P_X exactly, but we often do have substantial information about this distribution. Returning to our genetic example, it has been shown that the joint distribution of SNPs may be accurately modeled by hidden Markov models (see Stephens et al. [2001], Zhang et al. [2002], Qin et al. [2002], Li and Stephens [2003] for some early formulations), and there certainly is an abundance of genotype data to estimate the various model parameters; compare for instance the success of a variety of methods for genotype imputation [Marchini and Howie, 2010, Howie et al., 2012] based on such models. More generally, if a large amount of unlabeled data is available, the “deep knockoffs” methodology of Romano et al. [2018]

proposes using a deep generative model to generate knockoffs, subject to constraints that ensure that the knockoffs have approximately replicated the dependencies among the X_j's. Empirically, they find that this method is extremely effective at producing knockoff distributions that successfully control the FDR.

The purpose of this paper, then, is precisely to investigate common situations of this kind, namely, what happens when we run model-X knockoffs and only assume *approximate* knowledge of the distribution of X rather than exact knowledge, or equivalently, a construction of the knockoff features $\widetilde{X}$ that only *approximately* replicates the distribution of X. Our contribution is a considerable extension of the original work on model-X knockoffs [Candès et al., 2018], which assumed a perfect knowledge of the distribution of X to achieve FDR control. If we only have access to an approximation of the distribution of X, then it is certainly possible for model-X knockoffs to fail to control FDR—for instance, see [Romano et al., 2018, Sec. 6.5,6.6] for examples where estimating the distribution of X using only its first two moments is not sufficient for FDR control if the true distribution is heavy-tailed.

Here, we develop a new theory, which quantifies very precisely the inflation in FDR when running the knockoff filter with estimates of the distribution of X in place of the true distribution P_X. We develop non-asymptotic bounds which show that the possible FDR inflation is well-behaved whenever the estimated distribution is reasonably close to the truth. These bounds are general and apply to all possible statistics that the researcher may want to use to tease out the signal from the noise. We also develop converse results for some settings, showing that our bounds are fairly sharp in that it is impossible to obtain tighter FDR control bounds in full generality. Thus, our theory offers finite-sample guarantees that hold for any algorithm that the analyst decides to employ, assuming no knowledge of the form of the relationship between Y and X and only an estimate of the distribution of X itself. On the other hand, since our bounds are worst-case, they may be pessimistic in the sense that the realized FDR in any practical situation may be much lower than that achieved in the worst possible case.

Underlying our novel model-X knockoffs theory is a completely new mathematical analysis and understanding of the knockoffs inferential machine. The technical innovation here is essentially twofold. First, with only partial knowledge of the distribution of X, we can no longer achieve a perfect exchangeability between the test statistics for the null variables and for their knockoffs. Hence, we need tools that can deal with only a form of approximate exchangeability. Second, our methods to prove FDR control no longer rely on martingale arguments, and rather, involve leave-one-out type of arguments. These new arguments are likely to have applications far outside the scope of the present paper.

2. Robust inference with knockoffs.

To begin with, imagine we have data consisting of n i.i.d. draws from a joint distribution on (X, Y), where $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ is the feature vector while $Y \in \mathbb{R}$ is the response variable. We will gather the n observed data points into a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and vector $\mathbf{Y} \in \mathbb{R}^n$—that is, the pairs $(\mathbf{X}_{i,*}, \mathbf{Y}_i)$ are i.i.d. copies of the pair (X, Y). The joint distribution of (X, Y) is unknown—specifically, we do not assume any information about the conditional distribution of Y given X as discussed above. We work under the assumption that P_X, the marginal distribution of X, is known only approximately.

Since the Markov blanket of Y may be ill-defined (e.g. if two features are identical then the choice of the minimal set $\mathcal{S}$ may not be unique), we follow Candès et al. [2018] and define X_j to be a null variable if $X_j \perp\!\!\!\perp Y \mid X_{-j}$, that is if X_j and the response Y are independent conditionally on all the other variables. (We use the terms "features" and "variables" interchangeably.) Under very mild identifiability conditions, the set of non-nulls is nothing other than the Markov blanket of Y. Writing $\mathcal{H}_0$ to denote the set of indices corresponding to null variables, we can then reformulate the error we would like to control as $\mathbb{E}\left[|\widehat{\mathcal{S}} \cap \mathcal{H}_0| / |\widehat{\mathcal{S}}|\right] \le q$.

2.1. *Exact model-X knockoffs.*

Consider first an ideal setting where the distribution P_X is known. The model-X knockoffs method [Candès et al., 2018] is defined by constructing knockoff features satisfying the following conditions: $\widetilde{X}$ is drawn conditional on the feature vector X without looking at the response Y (i.e. $\widetilde{X} \perp\!\!\!\perp Y \mid X$), such that the joint distribution of $(X, \widetilde{X})$ satisfies a pairwise exchangeability condition, namely

$$(X,\, \widetilde{X})_{\mathrm{swap}(\mathcal{A})} \overset{\mathrm{d}}{=} (X,\, \widetilde{X}) \tag{1}$$

for any subset $\mathcal{A} \subseteq \{1, \dots, p\}$, where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. (In fact, to achieve FDR control, this condition only needs to hold for subsets $\mathcal{A} \subseteq \mathcal{H}_0$ containing only null variables.) Above, the family $(X,\, \widetilde{X})_{\mathrm{swap}(\mathcal{A})}$ is obtained from $(X,\, \widetilde{X})$ by swapping the entries X_j and $\widetilde{X}_j$ for each $j \in \mathcal{A}$; for example, with $p = 3$ and $\mathcal{A} = \{2, 3\}$,

$$(X_1, X_2, X_3, \widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3)_{\mathrm{swap}\,(\{2,3\})} = (X_1, \widetilde{X}_2, \widetilde{X}_3, \widetilde{X}_1, X_2, X_3).$$

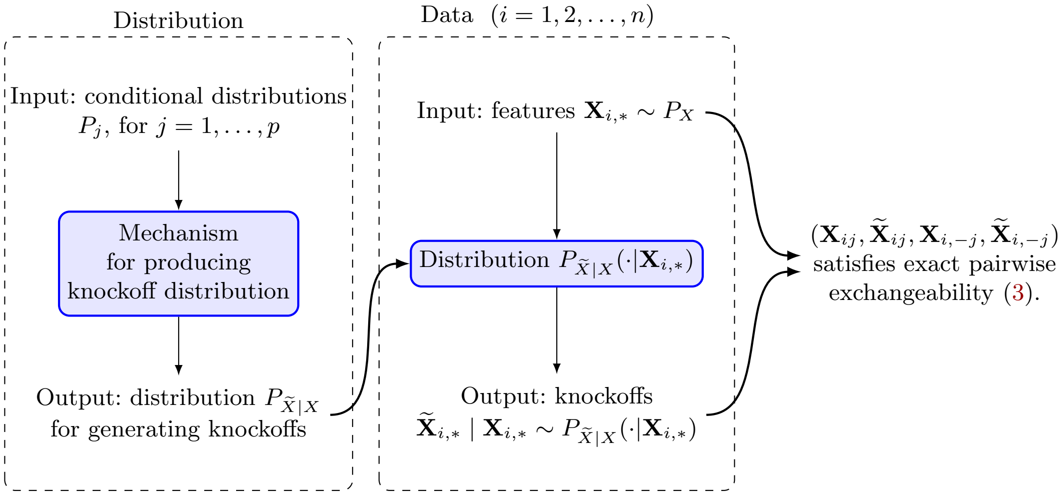
Distribution
Data $(i = 1, 2, \dots, n)$
Input: conditional distributions P_j, for $j = 1, \dots, p$
Mechanism for producing knockoff distribution
Output: distribution $P_{\widetilde{X}|X}$ for generating knockoffs
Input: features $\mathbf{X}_{i,*} \sim P_X$
Distribution $P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*})$
Output: knockoffs $\widetilde{\mathbf{X}}_{i,*} \mid \mathbf{X}_{i,*} \sim P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*})$
$(\mathbf{X}_{ij}, \widetilde{\mathbf{X}}_{ij}, \mathbf{X}_{i,-j}, \widetilde{\mathbf{X}}_{i,-j})$ satisfies exact pairwise exchangeability (3).

Fig 1: Schematic representation of the exact model-X knockoffs construction.

As a consequence of the pairwise exchangeability property (1), we see that the null knockoff variables $\{\widetilde{X}_j\}_{j \in \mathcal{H}_0}$ are distributed in exactly the same way as the original nulls $\{X_j\}_{j \in \mathcal{H}_0}$ but some dependence is preserved: for instance, for any pair $j \neq k$ where k is a null, we have that $(X_j, \widetilde{X}_k) \overset{\mathrm{d}}{=} (X_j, X_k)$.

Given knowledge of the true distribution P_X of the features X, our first step to implement the method of model-X knockoffs is to construct a distribution for drawing $\widetilde{X}$ conditional on X such that the pairwise exchangeability property (1) holds for all subsets of features $\mathcal{A}$. We can think of this mechanism as constructing some probability distribution $P_{\widetilde{X}|X}(\cdot|x)$, which is a conditional distribution of $\widetilde{X}$ given $X = x$, chosen so that the resulting joint distribution of $(X, \widetilde{X})$, which is equal to

$$P_X(x)\, P_{\widetilde{X}|X}(\widetilde{x}|x),$$

is symmetric in the pairs $(x_j, \widetilde{x}_j)$, and thus will satisfy the exchangeability property (1). Now, when working with data $(\mathbf{X}, \mathbf{Y})$, we will treat each data point $(\mathbf{X}_{i,*}, \mathbf{Y}_i)$ independently. Specifically, after observing the data $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$, the rows $\widetilde{\mathbf{X}}_{i,*}$ of the knockoff matrix are drawn from $P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*})$, independently for each i and also independently of $\mathbf{Y}$. Figure 1 shows a schematic representation of the exact model-X knockoffs construction.

It is important to point out that mechanisms for producing the pairwise exchangeability property (1) do exist and can be very concrete. As a specific example, suppose we wish to sample knockoff copies of Gaussian features, which follow a known Gaussian distribution $P_X = \mathcal{N}_p(\mathbf{0}_p, \Sigma)$. Then Candès et al. [2018] show that the knockoffs $\widetilde{\mathbf{X}}_{i,*}$ can be drawn from the conditional distribution

$$P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*}) = \mathcal{N}_p\big((\mathbf{I}_p - D\Sigma^{-1})\mathbf{X}_{i,*}, 2D - D\Sigma^{-1}D\big) \tag{2}$$

for any fixed diagonal matrix D satisfying $0 \preceq D \preceq 2\Sigma$. (This mechanism provides valid knockoffs because it ensures that the joint distribution of $(\mathbf{X}_{i,*}, \widetilde{\mathbf{X}}_{i,*})$ is given by

$$\mathcal{N}_{2p}\left(\mathbf{0}_{2p}, \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}\right),$$

which satisfies pairwise exchangeability (1).) There are also fast algorithms for the case where X follows either a Markov or a hidden Markov model [Sesia et al., 2018]. More broadly, Candès et al. [2018] develop a general abstract mechanism termed the Sequential Conditional Independent Pairs (SCIP) algorithm, which always produces exchangeable knockoff copies and can be applied to any distribution P_X. Looking ahead, all of these algorithms can be used in the case where P_X is known only approximately, where the exchangeability property (1) will be required to hold only with reference to the *estimated* distribution of X, discussed in Section 2.2 below.

For assessing a model selection algorithm, the knockoff feature vectors $\widetilde{\mathbf{X}}_j$ can be used as a "negative control"—a control group for testing the algorithm's ability to screen out false positives, since $\widetilde{\mathbf{X}}_j$ is known to have no real effect on $\mathbf{Y}$. Although details are given in Section 2.3, it is helpful to build some intuition already at this stage. Imagine for simplicity that we wish to assess the importance of a variable by measuring the strength of the marginal correlation with the response, i.e. we compute $Z_j = \left|\mathbf{X}_j^\top \mathbf{Y}\right|$. Then we can compare Z_j with $\widetilde{Z}_j = \left|\widetilde{\mathbf{X}}_j^\top \mathbf{Y}\right|$, the marginal correlation for the corresponding knockoff variable. The crucial point is that the pairwise exchangeability property (1) implies that if j is null (recall that this means that X_j and Y are conditionally independent given X_{-j}), then

$$(Z_j, \widetilde{Z}_j) \overset{\mathrm{d}}{=} (\widetilde{Z}_j, Z_j).$$

This holds without any assumptions on the form of the relationship $P_{Y|X}$ between Y and X [Candès et al., 2018]. In particular, this means that the test statistic $W_j = Z_j - \widetilde{Z}_j$ is equally likely to be positive or negative. Thus to reject the null, we would need to observe a large positive value of W_j. As we will see in Section 2.3, this way of reasoning extends to any choice of statistic Z_j; whatever statistic we choose, knockoff variables obeying (1) offer corresponding values of the statistic which can be used as "negative controls" for calibration purposes.

Throughout this paper, we will pay close attention to the distribution we obtain when swapping only one variable and its knockoff (and do not swap any of the other variables). In this context, we can reformulate the broad exchangeability condition (1) in terms of single variable swaps.

PROPOSITION 1 (Candès et al. [2018, Prop. 3.5]). *The pairwise exchangeability property* (1) *holds for a subset $\mathcal{A} \subseteq \{1, \dots, p\}$ if and only if*

$$\big(X_j, \widetilde{X}_j, X_{-j}, \widetilde{X}_{-j}\big) \overset{\mathrm{d}}{=} \big(\widetilde{X}_j, X_j, X_{-j}, \widetilde{X}_{-j}\big) \tag{3}$$

holds for all $j \in \mathcal{A}$.

In other words, we can restrict our attention to the question of whether a single given feature X_j and its knockoff $\widetilde{X}_j$ are exchangeable with each other (in the joint distribution that also includes X_{-j} and $\widetilde{X}_{-j}$).

2.2. *Approximate model-X knockoffs and pairwise exchangeability.* Now we will work towards constructing a version of this method when the true distribution of the feature vector X is not known exactly. Here, we need to relax the pairwise exchangeability assumption, since choosing a useful mechanism $P_{\widetilde{X}|X}$ that satisfies this condition would generally require a very detailed knowledge of the distribution of X, which is typically not available. This section builds towards a definition of pairwise exchangeability with respect to an approximate estimate of the distribution of X, in two steps.

From this point on, we will write P_X to denote the *assumed* joint distribution of X, and P_j for its conditionals; $P^\star_X$, and its conditionals $P^\star_j$, will denote the unknown *true* distribution of X. Throughout we will assume that P_X, our assumed or estimated distribution of X, is fixed or is independent from the data set $(\mathbf{X}, \mathbf{Y})$—for example, it may have been estimated from a separate unlabeled data set.

2.2.1. *Exchangeability with respect to an input distribution P_X.* We are provided with data $\mathbf{X}$ and conditional distributions $P_j(\cdot|x_{-j})$ for each j. As a warm-up, assume first that these conditionals are mutually compatible in the sense that there is a joint distribution P_X over $\mathbb{R}^p$ that matches these

p estimated conditionals—we will relax this assumption very soon. Then as shown in Figure 2, we repeat the construction from Figure 1, only with the P_j's as inputs. In words, the algorithm constructs knockoffs, which are samples from $P_{\widetilde{X}|X}$, a conditional distribution whose construction is based on the conditionals P_j or, equivalently, the joint distribution P_X. In place of requiring that pairwise exchangeability of the features X_j and their knockoffs $\widetilde{X}_j$ holds relative to the true distribution of X, as in (1) and (3), we instead require that the knockoff construction mechanism satisfy pairwise exchangeability conditions relative to the estimated joint distribution P_X that it receives as input:

$$\begin{aligned}&\text{If } (X, \widetilde{X}) \text{ is drawn as } X \sim P_X \text{ and } \widetilde{X} \mid X \sim P_{\widetilde{X}|X}(\cdot|X), \text{ then}\\ &\qquad (X, \widetilde{X})_{\mathrm{swap}(\mathcal{A})} \overset{\mathrm{d}}{=} (X, \widetilde{X}), \quad \text{for any subset } \mathcal{A} \subseteq \{1, \ldots, p\}.\end{aligned} \tag{4}$$

When only estimated compatible conditionals are available, original and knockoff features are required to be exchangeable with respect to the distribution P_X, which is provided as input (but not with respect to the true distribution of X, which is unknown). To rephrase, if the distribution of X were in fact equal to P_X, then we would have exchangeability.

2.2.2. *Exchangeability with respect to potentially incompatible conditionals P_j.* We wish to provide an extension of (4) to cover the case where the conditionals may not be compatible; that is, when a joint distribution with the P_j's as conditionals may not exist. To understand why this is of interest, imagine we have unlabeled data that we can use to estimate the distribution of X. Then we may construct P_j by regressing the jth feature X_j onto the $p-1$ remaining features X_{-j}. For instance, we may use a regression technique promoting sparsity or some other assumed structure. In such a case, it is easy to imagine that such a strategy may produce incompatible conditionals. It is, therefore, important to develop a framework adapted to this setting. To address this, we shall work throughout the paper with the following definition:

Definition 1. *$P_{\widetilde{X}|X}$ is* pairwise exchangeable with respect to P_j *if it satisfies the following property:*

$$\begin{aligned}&\textit{For any distribution } D^{(j)} \textit{ on } \mathbb{R}^p \textit{ with } j\textit{th conditional } P_j,\\ &\textit{if } (X, \widetilde{X}) \textit{ is drawn as } X \sim D^{(j)} \textit{ and } \widetilde{X} \mid X \sim P_{\widetilde{X}|X}(\cdot|X),\\ &\ \textit{then } \big(X_j, \widetilde{X}_j, X_{-j}, \widetilde{X}_{-j}\big) \overset{\mathrm{d}}{=} \big(\widetilde{X}_j, X_j, X_{-j}, \widetilde{X}_{-j}\big).\end{aligned} \tag{5}$$

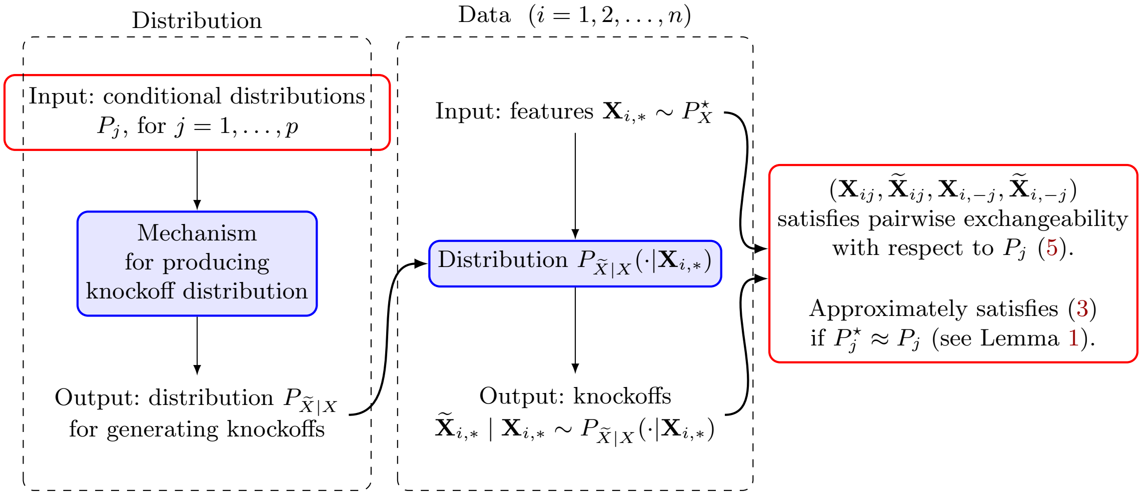
Distribution
Data $(i = 1, 2, \dots, n)$
Input: conditional distributions P_j, for $j = 1, \dots, p$
Mechanism for producing knockoff distribution
Output: distribution $P_{\widetilde{X}|X}$ for generating knockoffs
Input: features $\mathbf{X}_{i,*} \sim P^\star_X$
Distribution $P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*})$
Output: knockoffs $\widetilde{\mathbf{X}}_{i,*} \mid \mathbf{X}_{i,*} \sim P_{\widetilde{X}|X}(\cdot|\mathbf{X}_{i,*})$
$(\mathbf{X}_{ij}, \widetilde{\mathbf{X}}_{ij}, \mathbf{X}_{i,-j}, \widetilde{\mathbf{X}}_{i,-j})$ satisfies pairwise exchangeability with respect to P_j (5).
Approximately satisfies (3) if $P^\star_j \approx P_j$ (see Lemma 1).

Fig 2: Schematic representation of the approximate model-X knockoffs construction. The two differences relative to Figure 1 are circled in red.

Above, $D^{(j)}$ is the product of an arbitrary marginal distribution for X_{-j} and of the conditional P_j.

In words, with estimated conditionals P_j, we choose $P_{\widetilde{X}|X}$ to satisfy pairwise exchangeability with respect to these P_j's, for every j. (As before, we remark that this only needs to hold for $j \in \mathcal{H}_0$ to ensure FDR control, but since in practice we do not know which features are null, we require (5) to hold for every j.)

To see why this is an extension of (4), note that if the P_j's are mutually compatible, i.e. there is some distribution P_X with conditionals P_j for each j, then any algorithm operating such that (5) holds for each j, obeys (4) as well—this is because, for each j, we can apply (5) with the distribution $D^{(j)} = P_X$.

Now let's consider the question of how we might generate knockoff copies obeying (5). In the setting where our estimated conditionals P_j are all compatible with some joint distribution P_X on X, constructing knockoff copies in this approximate scenario is no different from the exact model-X knockoffs framework—if we have some mechanism which, when we input the joint distribution $P^\star_X$ of X, will produce exchangeable knockoffs obeying (1), then we can instead provide our estimated joint distribution P_X as input to produce knockoff copies that satisfy (4) and, by extension, satisfy (5). Hence, if the P_j's are mutually compatible, then all the mechanisms producing valid knockoffs under exact knowledge of $P^\star_X$—we mentioned a few in the previous section—can be readily used for our purposes. Later in Section 4, we will also give an example of a mechanism producing valid knockoffs satisfying (5) under incompatible P_j's.

2.2.3. *Probability of a swap.* We next develop a key lemma that will allow us to characterize the quality of our constructed knockoffs. In an exact model-X knockoffs framework, the key idea is that the knockoffs $\widetilde{X}_j$ act as controls for null variables X_j, because even after all observing all of the data—all the covariates, and the response Y—we are unable to tell which of the two, i.e. X_j and $\widetilde{X}_j$, is the real variable versus the knockoff. More precisely, each of the two is *equally likely* to be the real variable or the knockoff. Our next step in the approximate setting, therefore, is to determine whether this is approximately true when the estimated conditionals P_j are not too far from the true conditionals $P^\star_j$.

From this point on, we will assume without comment that for each j, either X_j and $\widetilde{X}_j$ are both discrete variables or are both continuous variables, and abusing notation, in these two settings we will use $P^\star_j(\cdot|x_{-j})$ and $P_j(\cdot|x_{-j})$ to denote the conditional probability mass function or conditional density, respectively, for the true and estimated conditional distribution of X_j given $X_{-j} = x_{-j}$. Furthermore, we assume that $P^\star_j(\cdot|x_{-j})$ and $P_j(\cdot|x_{-j})$ are supported on the same (discrete or continuous) set for any x_{-j}. Our theory can be generalized to the setting of mixed distributions and/or varying supports, but for clarity of the results we do not present these generalizations here.

The construction of the knockoff features as in Figure 2 yields the following approximate pairwise exchangeability result (proved in Appendix A).

LEMMA 1. *Fix any feature index j such that pairwise exchangeability (5) with respect to P_j is satisfied. If $X_j, \widetilde{X}_j$ are discrete, then for any*[2] *a, b,*

$$\frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b \;\middle|\; X_{-j}, \widetilde{X}_{-j}\right\}}{\mathbb{P}\left\{X_j = b, \widetilde{X}_j = a \;\middle|\; X_{-j}, \widetilde{X}_{-j}\right\}} = \frac{P^\star_j(a|X_{-j})P_j(b|X_{-j})}{P_j(a|X_{-j})P^\star_j(b|X_{-j})}.$$

Furthermore, if index j corresponds to a null feature (i.e. $X_j \perp\!\!\!\perp Y \mid X_{-j}$) and we additionally assume that $\widetilde{X} \mid X$ is drawn from $P_{\widetilde{X}|X}$ independently

[2] Formally, this result holds only for a, b lying in the support of $P^\star_j(\cdot|X_{-j})$, which is assumed to be equal to the support of $P_j(\cdot|X_{-j})$, as otherwise the ratio is 0/0; we ignore this possibility here and throughout the paper since these results will be applied only in settings where a, b do lie in this support.

of Y, then the same result holds when we also condition on Y:

$$\frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b \;\middle|\; X_{-j}, \widetilde{X}_{-j}, Y\right\}}{\mathbb{P}\left\{X_j = b, \widetilde{X}_j = a \;\middle|\; X_{-j}, \widetilde{X}_{-j}, Y\right\}} = \frac{P_j^\star(a|X_{-j})P_j(b|X_{-j})}{P_j(a|X_{-j})P_j^\star(b|X_{-j})}. \tag{6}$$

The conclusion in the continuous case is identical except with ratios of probabilities replaced with ratios of densities.

To better understand the roles of the various distributions at play, consider the two following scenarios for the joint distribution of the feature vector X and its knockoff copy $\widetilde{X}$:

$$\text{True distribution}\left\{\begin{array}{r} X_{-j} \sim \text{(any distribution)} \\ X_j \mid X_{-j} \sim P_j^\star(\cdot|X_{-j}) \\ \widetilde{X} \mid X \sim P_{\widetilde{X}|X}(\cdot|X) \end{array}\right. \qquad \text{Assumed distrib.}\left\{\begin{array}{r} X_{-j} \sim \text{(any distribution)} \\ X_j \mid X_{-j} \sim P_j(\cdot|X_{-j}) \\ \widetilde{X} \mid X \sim P_{\widetilde{X}|X}(\cdot|X) \end{array}\right.$$

The knockoff generating mechanism $P_{\widetilde{X}|X}$ is designed with the estimated conditional P_j in mind, and therefore by construction, X_j and $\widetilde{X}_j$ are exchangeable under the “Assumed distribution” scenario on the right, defined with the incorrect estimate P_j of the jth conditional. The real distribution of $(X, \widetilde{X})$ instead follows the scenario labeled as the “True distribution”, on the left. When $P_j^\star \neq P_j$, this means that X_j and $\widetilde{X}_j$ are only *approximately* exchangeable under the true distribution of the data. Lemma 1 quantifies the extent to which the pair $(X_j, \widetilde{X}_j)$ deviate from exchangeability, giving a useful formula for computing the ratio between the likelihoods of the two configurations $(X_j, \widetilde{X}_j) = (a, b)$ and $(X_j, \widetilde{X}_j) = (b, a)$ (after conditioning on the remaining data).

It is important to observe that if we are working in the exact model-X framework, where the true distribution and assumed distribution are the same (i.e. $P_j^\star = P_j$), then in this case the lemma yields

$$\frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b \;\middle|\; X_{-j}, \widetilde{X}_{-j}, Y\right\}}{\mathbb{P}\left\{X_j = b, \widetilde{X}_j = a \;\middle|\; X_{-j}, \widetilde{X}_{-j}, Y\right\}} = 1 \tag{7}$$

for each null j. That is, the two configurations are equally likely. This result for the exact model-X setting is proved in Candès et al. [2018, Lemma 3.2] and is critical for establishing FDR control properties. When we use estimates P_j rather than the true conditionals $P_j^\star$, however, the property (7) is no longer true, since Lemma 1 shows that the ratio is no longer equal to 1 in general. We can no longer use the knockoff statistics as exact negative controls; only as approximate controls. This is where the major difficulty comes in: if a knockoff statistic is only approximately distributed like its corresponding null, what is the potential inflation of the type-I error that this could cause? In other words, if $P_j \approx P_j^\star$ so that the ratio in (6) is slightly different from 1, how much might this inflate the resulting FDR?

Before proceeding with this question, we first give some additional background on the knockoff filter, to see how the knockoff variables $\widetilde{\mathbf{X}}_j$ will be used to test our hypotheses. We will then return in Section 3 to the question of how errors in constructing the knockoffs can affect the resulting FDR.

2.3. *The knockoff filter.* After constructing the variables $\widetilde{\mathbf{X}}_j$, we apply the knockoff filter to select important variables. We here quickly rehearse the main ingredients of this filter and refer the reader to Barber and Candès [2015] and Candès et al. [2018] for additional details; our exposition borrows from Barber and Candès [2019]. Suppose that for each variable $\mathbf{X}_j$ (resp. each knockoff variable $\widetilde{\mathbf{X}}_j$), we compute a score statistic Z_j (resp. $\widetilde{Z}_j$), such that

$$(Z_1, \ldots, Z_p, \widetilde{Z}_1, \ldots, \widetilde{Z}_p) = z\big([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{Y}\big),$$

with the idea that Z_j (resp. $\widetilde{Z}_j$) measures the importance of X_j (resp. $\widetilde{X}_j$) in explaining Y. Assume that the scores are "knockoff agnostic" in the sense that switching a variable with its knockoff simply switches the components of Z in the same way. This means that

$$z\big([\mathbf{X}, \widetilde{\mathbf{X}}]_{\mathrm{swap}(\mathcal{A})}, \mathbf{Y}\big) = z\big([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{Y}\big)_{\mathrm{swap}(\mathcal{A})} \tag{8}$$

i.e. swapping X_1 and $\widetilde{X}_1$ before calculating Z has the effect of swapping Z_1 and $\widetilde{Z}_1$, and similarly swapping X_2 and $\widetilde{X}_2$ swaps Z_2 and $\widetilde{Z}_2$, and so on. Here, we emphasize that Z_j may be an arbitrarily complicated statistic. For instance, it can be defined as the absolute value of a lasso coefficient, or some random forest feature importance statistic; or, we may fit both a lasso model and a random forest, and choose whichever one has the lowest cross-validated error.

These scores are then combined in a single importance statistic for the variable X_j as

$$W_j = f_j(Z_j, \widetilde{Z}_j) =: w_j\big([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{Y}\big),$$

where f_j is any anti-symmetric function, meaning that $f_j(v,u) = -f_j(u,v)$. As an example, we may have $W_j = Z_j - \widetilde{Z}_j$, where the Z_j's and $\widetilde{Z}_j$'s are the magnitudes of regression coefficients estimated by the lasso at a value of the regularization parameter given by cross-validation, say. Again, any choice of anti-symmetric function f_j and score statistic Z_j, no matter how complicated, is allowed. By definition, the statistics W_j obey the *flip-sign property*, which says that swapping the jth variable with its knockoff has the effect of changing the sign of W_j (since, by (8) above, if we swap feature vectors $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$ then Z_j and $\widetilde{Z}_j$ get swapped):

$$w_j\big([\mathbf{X},\,\widetilde{\mathbf{X}}]_{\mathrm{swap}(\mathcal{A})},\,\mathbf{Y}\big) = \begin{cases} w_j\big([\mathbf{X},\,\widetilde{\mathbf{X}}],\mathbf{Y}\big), & j \notin \mathcal{A}, \\ -w_j\big([\mathbf{X},\,\widetilde{\mathbf{X}}],\mathbf{Y}\big), & j \in \mathcal{A}. \end{cases} \tag{9}$$

The W_j's are the statistics that the knockoff filter will use. The idea is that large positive values of W_j provide evidence against the hypothesis that the distribution of Y is conditionally independent of X_j, while in contrast, if $j \in \mathcal{H}_0$, then W_j has a symmetric distribution and, therefore, is equally likely to take on positive or negative values.

In fact, it is equally valid for us to define $W_j = w_j\big([\mathbf{X},\,\widetilde{\mathbf{X}}],\mathbf{Y}\big)$ for any function w_j satisfying the flip-sign property (8), without passing through the intermediate stage of defining Z_j's and $\widetilde{Z}_j$'s, and from this point on we do not refer to the feature importance scores $Z_j, \widetilde{Z}_j$ in our theoretical results. However, for better understanding of the intuition behind the method, we should continue to think of W_j as comparing the apparent importance of the feature $\mathbf{X}_j$ versus its knockoff $\widetilde{\mathbf{X}}_j$ for modeling the response $\mathbf{Y}$.

Now that we have test statistics for each variable, we need a selection rule. For the knockoff filter, we choose a threshold $T_0 > 0$ by setting[3]

$$T_0 = \min\left\{ t > 0 : \frac{\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\}} \le q \right\} \ , \tag{10}$$

where q is the target FDR level. The output of the procedure is the selected model $\widehat{\mathcal{S}} = \{j : W_j \ge T_0\}$. In Barber and Candès [2015], it is argued that the ratio appearing in the right-hand side of (10) is an estimate of the false discovery proportion (FDP) if we were to use the threshold t—this is true because $\mathbb{P}\left\{W_j \ge t\right\} = \mathbb{P}\left\{W_j \le -t\right\}$ for any null feature $j \in \mathcal{H}_0$, and so we

[3] We want T_0 to be positive and the formal definition is that the minimum in (10) is taken over all $t > 0$ taking on values in the set $\{|W_1|, \dots, |W_p|\}$.

would roughly expect

$$\begin{aligned}(\#\text{ false positives at threshold } t) &= \#\{j \in \mathcal{H}_0 : W_j \geq t\} \\ &\approx \#\{j \in \mathcal{H}_0 : W_j \leq -t\} \\ &\leq \#\{j : W_j \leq -t\},\end{aligned} \tag{11}$$

that is, the numerator in (10) is an (over)estimate of the number of false positives selected at the threshold t. Hence, the selection rule can be interpreted as a step-up rule, stopping the first time our estimate falls below our target level. A slightly more conservative procedure, the knockoff+ filter, is given by incrementing the number of negatives by one, replacing the threshold in (10) with the choice

$$T_+ = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}, \tag{12}$$

and setting $\widehat{\mathcal{S}} = \{j : W_j \geq T_+\}$. Formalizing the intuition of our rough calculation (11), the false discovery rate control properties of these two procedures are studied in Barber and Candès [2015] under an exact pairwise exchangeability setting.

3. FDR control results.

3.1. *Measuring errors in the distribution.* If the knockoff features are generated using a mechanism designed to mimic the estimated conditionals P_j rather than the true conditional distributions $P^\star_j$, when can we hope for error control? Intuitively, if the conditional distributions $P^\star_j$ and P_j are similar, then we might hope that the knockoff feature $\widetilde{\mathbf{X}}_j$ is a reasonably good control group for the original feature $\mathbf{X}_j$.

In order to quantify this, we begin by measuring the discrepancy between the true conditional $P^\star_j$ and its estimate P_j. Define the random variable

$$\widehat{\mathrm{KL}}_j := \sum_i \log\left(\frac{P^\star_j(\mathbf{X}_{ij}|\mathbf{X}_{i,-j}) \cdot P_j(\widetilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j})}{P_j(\mathbf{X}_{ij}|\mathbf{X}_{i,-j}) \cdot P^\star_j(\widetilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j})}\right), \tag{13}$$

where the notation $\widehat{\mathrm{KL}}_j$ suggests the KL divergence. In fact, $\widehat{\mathrm{KL}}_j$ is the *observed* KL divergence between $(\mathbf{X}_j, \widetilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j})$ and $(\widetilde{\mathbf{X}}_j, \mathbf{X}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j})$. To prove this, working in the discrete case for simplicity, Lemma 1 tells us that

$$\sum_i \log\left(\frac{P^\star_j(\mathbf{x}_{ij}|\mathbf{x}_{i,-j})\cdot P_j(\widetilde{\mathbf{x}}_{ij}|\mathbf{x}_{i,-j})}{P_j(\mathbf{x}_{ij}|\mathbf{x}_{i,-j})\cdot P^\star_j(\widetilde{\mathbf{x}}_{ij}|\mathbf{x}_{i,-j})}\right) \\ = \log\left(\frac{\mathbb{P}\left\{(\mathbf{X}_j,\widetilde{\mathbf{X}}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j}) = (\mathbf{x}_j,\widetilde{\mathbf{x}}_j,\mathbf{x}_{-j},\widetilde{\mathbf{x}}_{-j})\right\}}{\mathbb{P}\left\{(\widetilde{\mathbf{X}}_j,\mathbf{X}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j}) = (\mathbf{x}_j,\widetilde{\mathbf{x}}_j,\mathbf{x}_{-j},\widetilde{\mathbf{x}}_{-j})\right\}}\right)$$

for any $\mathbf{x}_j,\widetilde{\mathbf{x}}_j,\mathbf{x}_{-j},\widetilde{\mathbf{x}}_{-j}$. Therefore, we see that

$$\mathbb{E}[\widehat{\mathrm{KL}}_j] = \mathrm{d}_{\mathrm{KL}}\Big((\mathbf{X}_j,\widetilde{\mathbf{X}}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j}) \,\big\|\, (\widetilde{\mathbf{X}}_j,\mathbf{X}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j})\Big),$$

where d_{KL} is the usual KL divergence between distributions. (Recall that the approximate conditionals P_j and the knockoff mechanism $P_{\widetilde{X}|X}$ are assumed to be chosen independent of the data $(\mathbf{X},\mathbf{Y})$, and so this KL divergence measures the difference between two *fixed* distributions.)

In the exact model-X setting, where the knockoff construction mechanism $P_{\widetilde{X}|X}$ satisfies the pairwise exchangeability property (1), Proposition 1 immediately implies that $(\mathbf{X}_j,\widetilde{\mathbf{X}}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j}) \overset{\mathrm{d}}{=} (\widetilde{\mathbf{X}}_j,\mathbf{X}_j,\mathbf{X}_{-j},\widetilde{\mathbf{X}}_{-j})$ and, thus, $\mathbb{E}[\widehat{\mathrm{KL}}_j] = 0$—and in fact, since we are using the true conditionals $P^\star_j$, or in other words $P_j = P^\star_j$, we would have $\widehat{\mathrm{KL}}_j = 0$ always.

In the approximate model-X framework, where $P_j \neq P^\star_j$, we will instead have $\mathbb{E}[\widehat{\mathrm{KL}}_j] > 0$ (although of course, for a given draw of the data, it may occur that $\widehat{\mathrm{KL}}_j$ is zero or even negative.) We can interpret $\widehat{\mathrm{KL}}_j$ as measuring the extent to which the pairwise exchangeability property (3) is violated for a specific feature j. We will see in our results below that controlling the $\widehat{\mathrm{KL}}_j$'s is sufficient to ensure control of the false discovery rate for the approximate model-X knockoffs method. More precisely, we will be able to bound the false positives coming from those null features which have small $\widehat{\mathrm{KL}}_j$.

3.2. *FDR control guarantee.*

We now present our guarantee for robust error control with the model-X knockoffs filter. The proof of this theorem appears in Appendix A.

THEOREM 1. *Under the definitions above, for any $\epsilon \geq 0$, consider the null variables for which $\widehat{\mathrm{KL}}_j \leq \epsilon$. If we use the knockoff+ filter, then the fraction of the rejections that correspond to such nulls obeys*

$$\mathbb{E}\left[\frac{\big|\{j : j \in \widehat{\mathcal{S}} \cap \mathcal{H}_0 \text{ and } \widehat{\mathrm{KL}}_j \leq \epsilon\}\big|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq q\cdot e^{\epsilon}. \tag{14}$$

In particular, this implies that the false discovery rate is bounded as

$$\mathrm{FDR} \leq \min_{\epsilon \geq 0} \left\{ q \cdot e^{\epsilon} + \mathbb{P}\left(\max_{j \in \mathcal{H}_0} \widehat{\mathrm{KL}}_j > \epsilon \right) \right\}. \tag{15}$$

Similarly, for the knockoff filter, for any $\epsilon \geq 0$, a slightly modified fraction of the rejections that correspond to nulls with $\widehat{\mathrm{KL}}_j \leq \epsilon$ obeys

$$\mathbb{E}\left[\frac{\left|\{j : j \in \widehat{\mathcal{S}} \cap \mathcal{H}_0 \text{ and } \widehat{\mathrm{KL}}_j \leq \epsilon\}\right|}{|\widehat{\mathcal{S}}| + q^{-1}} \right] \leq q \cdot e^{\epsilon},$$

and therefore, we obtain a bound on a modified false discovery rate:

$$\mathbb{E}\left[\frac{\left|\widehat{\mathcal{S}} \cap \mathcal{H}_0\right|}{|\widehat{\mathcal{S}}| + q^{-1}} \right] \leq \min_{\epsilon \geq 0} \left\{ q \cdot e^{\epsilon} + \mathbb{P}\left(\max_{j \in \mathcal{H}_0} \widehat{\mathrm{KL}}_j > \epsilon \right) \right\}.$$

In Section 4, we will see concrete examples where $\max_{j=1,\ldots,p} \widehat{\mathrm{KL}}_j$ is small with high probability, yielding a meaningful result on FDR control.

It worth pausing to unpack our main result a little. Clearly, we cannot hope to have error control over *all* nulls if we have done a poor job in constructing some of their knockoff copies, because our knockoff "negative controls" may be completely off. Having said this, (14) tells us that that if we restrict our definition of false positives to only those nulls for which we have a reasonable "negative control" via the knockoff construction, then the FDR is controlled. Since we do not make any assumptions, this type of result is all one can really hope for. In other words, exact model-X knockoffs make the assumption that the knockoff features provide exact controls for each null, thus ensuring control of the false positives; our new result removes this assumption, and provides a bound on the false positives when counting only those nulls for which the corresponding knockoff feature provides an approximate control.

In a similar fashion, imagine running a multiple comparison procedure, e.g. the Benjamini–Hochberg procedure, with p-values that are not uniformly distributed under the null. Then in such a situation, we cannot hope to achieve error control over *all* nulls if some of the null p-values follow grossly incorrect distributions. However, we may still hope to achieve reasonable control over those nulls for which the p-value is close to uniform.

A noteworthy aspect of this result is that it makes no modeling assumption whatsoever. Indeed, our FDR control guarantees hold in any setting—no matter the relationship $P_{Y|X}$ between Y and X, no matter the true distribution $P^{\star}_X$ of the feature vector X, and no matter the test statistics W the data analyst has decided to employ (as long as W obeys the flip-sign condition). What the theorem says is that when we use estimated conditionals P_j, if the P_j's are close to the true conditionals $P_j^\star$ in the sense that the quantities $\widehat{\mathrm{KL}}_j$ are small, then the FDR is well under control. (In the ideal case where we use the true conditionals, then $\widehat{\mathrm{KL}}_j = 0$ for all $j \in \mathcal{H}_0$, and we automatically recover the FDR-control result from Candès et al. [2018]; that is, we get FDR control at the nominal level q since we can take $\epsilon = 0$.)

Finally, we close this section by emphasizing that the proof of Theorem 1 employs arguments that are completely different from those one finds in the existing literature on knockoffs. We discuss the novelties in our techniques in Appendix A.

3.2.1. *Is KL the right measure?.* As mentioned above, our theorem applies to any construction of the statistics W, including adversarial constructions that might be chosen deliberately to try to detect the differences between the X_j's and the $\widetilde{X}_j$'s. It is therefore expected that in any practical scenario, the achieved FDR would be lower than that suggested by our upper bounds. In practice, W would be chosen to try to identify strong correlations with Y, and we would not expect that this type of statistic is worst-case in terms of finding discrepancies between the distributions of X_j and $\widetilde{X}_j$. In fact, empirical studies [Candès et al., 2018, Sesia et al., 2018] have already reported on the robustness of model-X knockoffs vis-à-vis possibly large model misspecifications when W is chosen to identify a strong dependence between X and Y.

Examining our result more closely, we can see that our theorem applies to any statistic W because the $\widehat{\mathrm{KL}}_j$'s measure our ability to distinguish between each $\mathbf{X}_j$ and its knockoff copy $\widetilde{\mathbf{X}}_j$, and therefore if the two are virtually indistinguishable (i.e. $\widehat{\mathrm{KL}}_j$ is small), then *any* importance statistic W is almost equally likely to have $W_j > 0$ or $W_j < 0$ (as long as W obeys the "flip-sign" property (9)). In other words, if $\widehat{\mathrm{KL}}_j$ is low, then $\widetilde{\mathbf{X}}_j$ provides a high quality "control group" for the null $\mathbf{X}_j$, under any choice of W. However, when we run the knockoff filter in practice, our statistics $W = (W_1, \ldots, W_p)$ provide only a coarse summary of the data $\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y}$. Even if the p-dimensional vectors $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$ contain sufficient information for us to distinguish between the original null variable and its knockoff (due to a poor approximation P_j of $P_j^\star$), it is likely that much of this information is lost when we observe only W instead of the full data. Therefore, a small $\widehat{\mathrm{KL}}_j$ is sufficient, but by no means necessary, for FDR control—$\widehat{\mathrm{KL}}_j$ being small means that we are unable to distinguish between a null and its knockoff when viewing the full data, while for FDR control we only need to establish that the two are indistinguishable when viewing the statistics $W_1, \dots, W_p$. To formalize this idea, suppose that we fix some choice of statistic W (i.e. a map from the data $(\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y})$ to the statistic $W = (W_1, \dots, W_p)$). Suppose that random variables $E_1, \dots, E_p$ satisfy the following property:

$$\begin{aligned}(16)\quad &\mathbb{P}\left\{W_j > 0, E_j \leq \epsilon \mid |W_j|, W_{-j}\right\} \\ &\qquad\qquad \leq e^{\epsilon} \cdot \mathbb{P}\left\{W_j < 0 \mid |W_j|, W_{-j}\right\} \ \forall\, \epsilon \geq 0,\ j \in \mathcal{H}_0.\end{aligned}$$

(We would generally choose the E_j's to be functions of $(\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y})$, and would then interpret the probability as being taken with respect to the joint distribution of the data $(\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y})$.) For each null j, if E_j is low then this means that, if we are only given access to the statistic W (rather than viewing the full data), then we do not have much hope of distinguishing between the jth feature and its knockoff copy. The following lemma verifies that the $\widehat{\mathrm{KL}}_j$'s satisfy this property universally, i.e. for any choice of the feature importance statistic W.

LEMMA 2. *For any choice of statistic W that obeys the "flip-sign" property* (9)*, the random variables $\widehat{\mathrm{KL}}_j$ defined in* (13) *satisfy the property* (16)*.*

We will now generalize our FDR control result, Theorem 1, to replace $\widehat{\mathrm{KL}}_j$ with any knockoff quality measure E_j satisfying the property (16). The proof of this theorem, and the lemma above, appear in Appendix A.

THEOREM 2. *Under the definitions above, let W be a statistic satisfying the "flip-sign" property* (9)*. Suppose that, for this choice of W, the random variables $E_1, \dots, E_p$ satisfy the property* (16)*, meaning that they measure the quality of the knockoffs with respect to W. Then the conclusions of Theorem 1 hold with E_j in place of $\widehat{\mathrm{KL}}_j$ for each j.*

In particular, if the statistic W reveals much less information than the full data set $\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y}$, then it may be possible to construct E_j's that are in general much lower than the $\widehat{\mathrm{KL}}_j$'s, thus yielding a tighter bound on FDR. It remains to be seen whether, in specific settings for the distribution of the data, there are natural examples of the statistic W that are amenable to constructing tightly controlled E_j's to yield tighter bounds on the resulting FDR. We aim to explore this question in future work, but here we give one potential example. Suppose that the statistic W depends on the data

$\mathbf{X}, \widetilde{\mathbf{X}}, \mathbf{Y}$ only through some coarse summary statistics, for example, only through $\mathbf{X}^\top\mathbf{Y}$ and $\widetilde{\mathbf{X}}^\top\mathbf{Y}$. In this setting, for any values $a, b \in \mathbb{R}$, define

$$E_j(a,b) = \log\left(\frac{\mathbb{P}\left\{(\mathbf{X}_j^\top\mathbf{Y}, \widetilde{\mathbf{X}}_j^\top\mathbf{Y}) = (a,b) \;\middle|\; \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}}{\mathbb{P}\left\{(\mathbf{X}_j^\top\mathbf{Y}, \widetilde{\mathbf{X}}_j^\top\mathbf{Y}) = (b,a) \;\middle|\; \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}}\right)$$

(where the numerator and denominator are interpreted as conditional probabilities or conditional densities, as appropriate). We can then take

$$E_j = E_j(\mathbf{X}_j^\top\mathbf{Y}, \widetilde{\mathbf{X}}_j^\top\mathbf{Y})$$

and, by our assumption on W, we can verify that these E_j's satisfy the desired property (16). Now, will E_j yield a better bound on FDR? We can see that E_j measures the extent to which the one-dimensional random variables $\mathbf{X}_j^\top\mathbf{Y}$ and $\widetilde{\mathbf{X}}_j^\top\mathbf{Y}$ are distinguishable from each other, after observing the remaining data, i.e. $\mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}$. In contrast, $\widehat{\mathrm{KL}}_j$ measures the same question for the full n-dimensional random vectors $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$, and therefore will in general be much larger than E_j.

3.3. *A lower bound on FDR.* Next, we ask whether it is possible to prove a converse to Theorem 1, which guarantees FDR control as long as the $\widehat{\mathrm{KL}}_j$'s are small. We are interested in knowing whether bounding the $\widehat{\mathrm{KL}}_j$'s is in fact necessary for FDR control—or is it possible to achieve an FDR control guarantee even when the $\widehat{\mathrm{KL}}_j$'s are large? Of course, as discussed in Section 3.2.1, for a predefined choice of the statistic W, the $\widehat{\mathrm{KL}}_j$'s may yield very conservative results. Here, however, we are interested in determining whether the $\widehat{\mathrm{KL}}_j$'s are indeed the right measure of FDR inflation when we are aiming for a result that is universal over *all* FDR control methods.

Theorem 3 below proves that, if there is a feature j for which $\widehat{\mathrm{KL}}_j$ does not concentrate near zero, then we can construct an honest model selection method that, when assuming that the conditional distribution of $X_j \mid X_{-j}$ is given by P_j, *fails* to control FDR at the desired level if the true conditional distribution is in fact $P_j^\star$. By "honest", we mean that the model selection method would successfully control FDR at level q if P_j were the true conditional distribution. Our construction does not run a knockoff filter on the data; it is instead a hypothesis testing based procedure, meaning that the $\widehat{\mathrm{KL}}_j$'s govern whether it is possible to control FDR in a general sense. Hence, our converse is information-theoretic in nature and not specific to the knockoff filter. The proof of Theorem 3 is given in the Supplementary Material [Barber et al., 2018].

THEOREM 3. *Fix any distribution $P^\star_X$, any feature index j, and any estimated conditional distribution P_j. Suppose that there exists a knockoff sampling mechanism $P_{\widetilde{X}|X}$ that is pairwise exchangeable with respect to P_j (5), such that*

$$\mathbb{P}\left\{\widehat{\mathrm{KL}}_j \geq \epsilon\right\} \geq c$$

for some $\epsilon, c > 0$ when $(\mathbf{X}, \widetilde{\mathbf{X}})$ is drawn from $P^\star_X \times P_{\widetilde{X}|X}$. Then there exists a conditional distribution $P_{Y|X}$, and a testing procedure $\widehat{\mathcal{S}}$ that maps data $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$ to a selected set of features $\widehat{\mathcal{S}}(\mathbf{X}, \mathbf{Y}) \subseteq \{1, \dots, p\}$, such that:

- *If the data points $(\mathbf{X}_{i,*}, \mathbf{Y}_i)$ are i.i.d. draws from the distribution $P_X \times P_{Y|X}$, where P_X is any distribution whose jth conditional is P_j (that is, our estimated conditional distribution P_j for feature X_j is correct), then*
$$\mathrm{FDR}\big(\widehat{\mathcal{S}}\big) = q.$$
- *On the other hand, if the data points $(\mathbf{X}_{i,*}, \mathbf{Y}_i)$ are i.i.d. draws from the distribution $P^\star_X \times P_{Y|X}$ (i.e. our estimated conditional distribution P_j is not correct, as the true conditional distribution is $P^\star_j$), then*
$$\mathrm{FDR}\big(\widehat{\mathcal{S}}\big) \geq q\big(1 + c(1 - e^{-\epsilon})\big).$$

For the last case (where $P^\star_X$ is the true distribution), if $c \approx 1$ (i.e. $\widehat{\mathrm{KL}}_j \geq \epsilon$ with high probability) then $\mathrm{FDR}\big(\widehat{\mathcal{S}}\big) \gtrapprox q(2 - e^{-\epsilon})$; when $\epsilon \approx 0$ is small, we have $2 - e^{-\epsilon} \approx 1 + \epsilon \approx e^\epsilon$, which is the same inflation factor on the FDR on the upper bound in Theorem 1. In other words, Theorems 1 and 3 provide (nearly) matching upper and lower bounds. With these theorems, we do not aim to claim that the knockoffs methodology is universally robust, but rather, to determine and quantify the robustness properties of this already existing method. It is indeed true that substantial mistakes in the model of X can lead to a loss of FDR control, and the theorems above show that the $\widehat{\mathrm{KL}}_j$'s quantify exactly when, and to what extent, this issue has the potential to occur. Of course, as discussed above in Section 3.2.1, if we restrict our attention to prespecified statistics W, then the actual loss of FDR control maybe much less severe than that predicted by the bounds in Theorem 1.

4. Examples. To make our FDR control results more concrete, we will give two examples of settings where accurate estimates P_j of the conditionals $P^\star_j$ ensure that the $\widehat{\mathrm{KL}}_j$'s are bounded near zero. Examining the definition (13) of $\widehat{\mathrm{KL}}_j$, we see that $\widehat{\mathrm{KL}}_j$ is a sum of n i.i.d. terms, and we can therefore expect that large deviation bounds such as Hoeffding's inequality can be used to provide an upper bound uniformly across all p features. (Of course, as noted in Section 3.2.1, measuring knockoff quality via the $\widehat{\mathrm{KL}}_j$'s is a "worst-case" analysis that will bound FDR universally over all statistics W, and may therefore give a very conservative result; for a specific predefined choice of W, it may be possible to compute a tighter bound.)

All theoretical results in this section are proved in the Supplementary Material.

4.1. *Bounded errors in the likelihood ratio.* First, suppose that our estimates P_j of the conditional distribution $P^\star_j$ satisfy a likelihood ratio bound uniformly over any values for the variables:

$$\log\left(\frac{P^\star_j(x_j \mid x_{-j}) \cdot P_j(x'_j \mid x_{-j})}{P_j(x_j \mid x_{-j}) \cdot P^\star_j(x'_j \mid x_{-j})}\right) \leq \delta \tag{17}$$

for all j, all x_j, x'_j, and all x_{-j}. In this setting, the following lemma, proved via Hoeffding's inequality, gives a bound on the $\widehat{\mathrm{KL}}_j$'s:

LEMMA 3. *If the condition* (17) *holds uniformly for all j and all x_j, x'_j, x_{-j}, then with probability at least $1 - \frac{1}{p}$,*

$$\max_{j=1,\dots,p} \widehat{\mathrm{KL}}_j \leq \frac{n\delta^2}{2} + 2\delta\sqrt{n\log(p)}.$$

In other words, if P_j satisfies (17) for some $\delta = o\left(\frac{1}{\sqrt{n\log(p)}}\right)$, then with high probability every $\widehat{\mathrm{KL}}_j$ will be small. By Theorem 1, then, the FDR for model-X knockoffs in this setting is controlled near the target level q.

4.2. *Gaussian knockoffs.* For a second example, suppose that the distribution of the feature vector X is mean zero and has covariance Θ^{-1}, where Θ is some unknown precision matrix. (We assume zero mean for simplicity, but these results can of course be generalized to an arbitrary mean.) Suppose that we have estimated Θ with some approximation $\widetilde{\Theta}$, and let Θ_j and $\widetilde{\Theta}_j$ denote the jth columns of these matrices. Our results below will assume that the error in estimating each column of Θ is small, i.e. $\widetilde{\Theta}_j - \Theta_j$ is small for all j.

As described earlier in (2), Candès et al. [2018, eqn. (3.2)]'s Gaussian knockoff construction consists of drawing the knockoffs according to the conditional distribution $P_{\widetilde{X}|X}(\cdot|X)$ given by

$$\widetilde{X} \mid X \sim \mathcal{N}_p\big((\mathbf{I}_p - D\widetilde{\Theta})X, 2D - D\widetilde{\Theta}D\big), \tag{18}$$

where $D = \mathrm{diag}\{d_j\}$ is a nonnegative diagonal matrix chosen to satisfy $2D - D\widetilde{\Theta}D \succeq 0$, or equivalently, $D \preceq 2\widetilde{\Theta}^{-1}$. If the true precision matrix of X were given by $\widetilde{\Theta}$ (assumed to be positive definite), then we can calculate that the joint distribution of the pair $(X, \widetilde{X})$ has first and second moments given by

$$\mathbb{E}\left[\begin{pmatrix} X \\ \widetilde{X} \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathrm{Var}\left(\begin{pmatrix} X \\ \widetilde{X} \end{pmatrix}\right) = \begin{pmatrix} \widetilde{\Theta}^{-1} & \widetilde{\Theta}^{-1} - D \\ \widetilde{\Theta}^{-1} - D & \widetilde{\Theta}^{-1} \end{pmatrix}.$$

In other words, for every j, X_j and $\widetilde{X}_j$ are exchangeable if we only look at the first and second moments of the joint distribution.

If the true distribution of X is in fact Gaussian, again with mean zero and covariance $\widetilde{\Theta}^{-1}$, then a stronger claim follows—the joint distribution of $(X, \widetilde{X})$ is then multivariate Gaussian and therefore $(X, \widetilde{X})_{\mathrm{swap}(\mathcal{A})} \stackrel{\mathrm{d}}{=} (X, \widetilde{X})$ for every subset $\mathcal{A} \subseteq [p]$. In other words, the knockoff construction determined by $P_{\widetilde{X}|X}$ satisfies pairwise exchangeability, as defined in (4), with respect to the distribution $P_X = \mathcal{N}_p(0, \widetilde{\Theta}^{-1})$. To frame this property in terms of conditionals, P_j, rather than an estimated joint distribution, P_X, we can calculate the estimated conditional distributions $P_j(\cdot|X_{-j})$ as

$$X_j \mid X_{-j} \sim \mathcal{N}\left(X_{-j}^\top \left(\frac{-\widetilde{\Theta}_{-j,j}}{\widetilde{\Theta}_{jj}}\right), \frac{1}{\widetilde{\Theta}_{jj}}\right), \tag{19}$$

where $\widetilde{\Theta}_{-j,j} \in \mathbb{R}^{p-1}$ is the column $\widetilde{\Theta}_j$ with entry $\widetilde{\Theta}_{jj}$ removed.

As noted in Section 2.2, we may want to work with estimated precision matrices, which are not positive semidefinite (PSD). The rationale is that if $\widetilde{\Theta}$ is fitted by regressing each X_j on the remaining features X_{-j} to produce the jth column, $\widetilde{\Theta}_j$, then the result will not be PSD in general. If $\widetilde{\Theta}$ is not PSD, although there is no corresponding joint distribution, the conditionals P_j (19) are still well-defined as long as $\widetilde{\Theta}_{jj} > 0$ for all j; they are just not compatible. (Note that symmetry is a far easier constraint to enforce, e.g. by simply replacing our initial estimate $\widetilde{\Theta}$ with $(\widetilde{\Theta} + \widetilde{\Theta}^\top)/2$, which preserves desirable features such as sparsity that might be present in the initial $\widetilde{\Theta}$;

in contrast, projecting to the PSD cone while enforcing sparsity constraints may be computationally challenging in high dimensions.)

Our first result verifies that this construction of $P_{\widetilde{X}|X}$ satisfies pairwise exchangeability with respect to the conditional distributions P_j given in (19):

LEMMA 4. *Let $\widetilde{\Theta} \in \mathbb{R}^{p\times p}$ be a symmetric matrix with a positive diagonal, and let $P_{\widetilde{X}|X}$ be defined as in (18). Then, for each $j = 1,\dots,p$, $P_{\widetilde{X}|X}$ is pairwise exchangeable with respect to the conditional distribution P_j given in (19)—that is, the exchangeability condition (5) is satisfied.*

In practice, we would construct Gaussian knockoffs in situations where the distribution of X might be well approximated by a multivariate normal. The lemma below gives a high probability bound on the $\widehat{\mathrm{KL}}_j$'s in the case where the features are indeed Gaussian but with an unknown covariance matrix Θ^{-1}. Here, Gaussian concentration results can be used to control the $\widehat{\mathrm{KL}}_j$'s, which then yields FDR control. (We note that recent work by Fan et al. [2017] also studies the Gaussian model-X knockoffs procedure with an estimated precision matrix $\widetilde{\Theta}$, under a different framework.)

LEMMA 5. *Let $\Theta, \widetilde{\Theta} \in \mathbb{R}^{p\times p}$ be any matrices, where Θ is positive definite and $\widetilde{\Theta}$ is symmetric with a positive diagonal. Suppose that $\mathbf{X}_{i,*} \overset{\text{iid}}{\sim} \mathcal{N}_p(0,\Theta^{-1})$, while $\widetilde{\mathbf{X}} \mid \mathbf{X}$ is drawn according to the distribution $P_{\widetilde{X}|X}$ given in (18). Define*

$$\delta_\Theta = \max_{j=1,\dots,p} (\Theta_{jj})^{-1/2} \cdot \|\Theta^{-1/2}(\widetilde{\Theta}_j - \Theta_j)\|_2. \tag{20}$$

Then with probability at least $1 - \frac{1}{p}$,

$$\max_{j=1,\dots,p} \widehat{\mathrm{KL}}_j \leq 4\delta_\Theta \sqrt{n\log(p)} \cdot (1 + o(1)),$$

where the $o(1)$ term refers to terms that are vanishing when we assume that $\frac{\log(p)}{n} = o(1)$ and that this upper bound is itself bounded by a constant.

(A formal bound making the $o(1)$ term explicit is provided in the proof.) In particular, comparing to our FDR control result, Theorem 1, we see that as long as the columnwise error in estimating the precision matrix Θ satisfies $\delta_\Theta = o\left(\frac{1}{\sqrt{n\log(p)}}\right)$, the FDR will be controlled near the target level q.

When might we be able to attain such a bound on the error in estimating Θ? As mentioned earlier, in many applied settings, we may have access to substantially more unlabeled data (i.e. the feature vector X without an associated response Y) than labeled data (pairs (X, Y)). Suppose that, for the purpose of estimating Θ, we have access to $N \gg n$ draws of the feature vector $X \sim P^\star_X$. When the distribution of X is multivariate Gaussian with a sparse inverse covariance matrix Θ, the graphical Lasso [Yuan, 2007, Friedman et al., 2008] estimates Θ as

$$\widehat{\Theta}_\lambda = \arg\min_{A \succeq 0} \Big\{ -\log\det(A) + \langle A, \widehat{S}_N \rangle + \lambda \sum_{j \neq k} |A_{jk}| \Big\},$$

where $\widehat{S}_N$ is the sample covariance matrix of the unlabeled training data while $\lambda > 0$ is a penalty parameter inducing sparsity in the resulting solution. Ravikumar et al. [2011] proved that, if Θ is sufficiently sparse, then under certain additional assumptions and with an appropriate choice of penalty parameter λ, the graphical Lasso solution $\widehat{\Theta}_\lambda$ satisfies an entrywise error bound $\|\widehat{\Theta}_\lambda - \Theta\|_\infty \lesssim \sqrt{\frac{\log(p)}{N}}$, and furthermore, is asymptotically guaranteed to avoid any false positives (i.e. if $\Theta_{jk} = 0$ then $(\widehat{\Theta}_\lambda)_{jk} = 0$). Therefore, if each column of Θ has at sparsity at most s_Θ (i.e. at most s_Θ nonzeros) and Θ has bounded condition number, this then proves that the bound (20) on the error in estimating Θ holds with $\delta_\Theta \asymp \sqrt{\frac{s_\Theta \log(p)}{N}}$. We conclude that the results of Lemma 5 give a meaningful bound on FDR control as long as

$$4\delta_\Theta \cdot \sqrt{n \log(p)} \asymp \sqrt{\frac{s_\Theta \log(p)}{N}} \cdot \sqrt{n \log(p)} = o(1).$$

Equivalently, it is sufficient to have an unlabeled sample size N satisfying

$$N \gg n \cdot s_\Theta \log^2(p).$$

5. Discussion. In this paper, we established that the method of model-X knockoffs is robust to errors in the underlying assumptions on the distribution of the feature vector X, making it an effective method for many practical applications, such as genome-wide association studies, where the underlying distribution on the features $X_1, \ldots, X_p$ can be estimated accurately. One notable aspect is that our theory is free of any modeling assumptions, since our theoretical guarantees hold no matter the data distribution or the statistics that the data analyst wishes to use, even if they are designed to exploit some weakness in the construction of knockoffs. Looking forward, it would be interesting to develop a theory for fixed statistics, as outlined in Section 3.2.1. For instance, if the researcher commits to using a pre-specified random forest feature importance statistic, or some statistic based on the magnitudes of lasso coefficients (perhaps calculated at a data-dependent value of the regularization parameter), then what can be said about FDR control? In other words, what can we say when the statistics W only probe the data in certain directions? We leave such interesting questions for further research.

Acknowledgements. R. F. B. was partially supported by the National Science Foundation via grant DMS 1654076, and by an Alfred P. Sloan fellowship. E. C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, by the National Science Foundation via DMS 1712800, by the Math + X Award from the Simons Foundation and by a generous gift from TwoSigma. E. C. would like to thank Chiara Sabatti for useful conversations related to this project. R. J. S. was partially supported by Engineering and Physical Sciences Research Council Fellowships EP/J017213/1 and EP/P031447/1, and by grant RG81761 from the Leverhulme Trust.

References.

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

R. F. Barber and E. J. Candès. A knockoff filter for high-dimensional selective inference. *Ann. Statist. (to appear)*, 2019.

R. F. Barber, E. J. Candès, and R. J. Samworth. Supplementary material to 'Robust inference with knockoffs'. *Submitted*, 2018.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1): 289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.

E. J. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 80 (3):551–577, 2018.

Y. Fan, E. Demirkaya, G. Li, and J. Lv. RANK: large-scale inference with graphical nonlinear knockoffs. *arXiv preprint arXiv:1709.00092*, 2017.

J. A. Ferreira and A. H. Zwinderman. On the Benjamini–Hochberg method. *Ann. Statist.*, 34(4):1827–1849, 2006.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, 44(8):955–959, 2012.

L. Janson. *A model-free approach to high-dimensional inference*. PhD thesis, Stanford University, 2017.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, 11(7):499–511, 2010.

Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, 71 (5):1242–1247, 2002.

P. Ravikumar, M. J Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–980, 2011.

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

M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, page asy033, 2018. .

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68(4):978–989, 2001.

J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004.

Y. Yuan, M.and Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. U.S.A.*, 99(11):7335–7339, 2002.

APPENDIX A: PROOFS OF MAIN RESULTS

Whereas all proofs of FDR control for the knockoff methods thus far have relied on martingale arguments (see Barber and Candès [2015], Barber and Candès [2019], Candès et al. [2018]), here we will prove our main theorem using a novel leave-one-out argument. Before we begin, we would like to draw a loose analogy. To prove FDR controlling properties of the Benjamini–Hochberg procedure under independence of the p-values, Storey et al. [2004] developed a very elegant martingale argument. Other proof techniques, however, operate by removing or leaving out one hypothesis (or one p-value); see Benjamini and Yekutieli [2001], Ferreira and Zwinderman [2006] for examples. At a very high level, our own methods are partially inspired by the latter approach.

A.1. Proofs of FDR control results, Theorems 1 and 2.

Theorem 1 follows directly from Theorem 2 combined with Lemma 2, and thus requires no separate proof. To prove Theorem 2, for any $\epsilon \geq 0$ and for any threshold $t > 0$, define

$$R_\epsilon(t) := \frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\left\{W_j \geq t, \widehat{\mathrm{KL}}_j \leq \epsilon\right\}}{1 + \sum_{j\in\mathcal{H}_0} \mathbb{1}\left\{W_j \leq -t\right\}}.$$

Then, for the knockoff+ filter with threshold T_+, we can write

$$\begin{aligned}\frac{\left|\{j : j \in \widehat{\mathcal{S}} \cap \mathcal{H}_0 \text{ and } \widehat{\mathrm{KL}}_j \leq \epsilon\}\right|}{|\widehat{\mathcal{S}}| \vee 1} &= \frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\left\{W_j \geq T_+, \widehat{\mathrm{KL}}_j \leq \epsilon\right\}}{1 \vee \sum_j \mathbb{1}\left\{W_j \geq T_+\right\}}\\ &= \frac{1+\sum_j \mathbb{1}\left\{W_j \leq -T_+\right\}}{1 \vee \sum_j \mathbb{1}\left\{W_j \geq T_+\right\}} \cdot \frac{\sum_{j\in\mathcal{H}_0} \mathbb{1}\left\{W_j \geq T_+, \widehat{\mathrm{KL}}_j \leq \epsilon\right\}}{1+\sum_j \mathbb{1}\left\{W_j \leq -T_+\right\}}\\ &\leq \frac{1+\sum_j \mathbb{1}\left\{W_j \leq -T_+\right\}}{1 \vee \sum_j \mathbb{1}\left\{W_j \geq T_+\right\}} \cdot R_\epsilon(T_+) \leq q \cdot R_\epsilon(T_+),\end{aligned}$$

where the next-to-last step holds by definition of R_ϵ, and the last step holds by the construction of the knockoff+ filter. If we instead use the knockoff filter (rather than knockoff+), then we use the threshold T_0 and similarly obtain

$$\begin{aligned}\frac{\left|\{j : j \in \widehat{\mathcal{S}} \cap \mathcal{H}_0 \text{ and } \widehat{\mathrm{KL}}_j \leq \epsilon\}\right|}{q^{-1} + |\widehat{\mathcal{S}}|} &\leq \frac{1+\sum_j \mathbb{1}\left\{W_j \leq -T_0\right\}}{q^{-1} + \sum_j \mathbb{1}\left\{W_j \geq T_0\right\}} \cdot R_\epsilon(T_0)\\ &\leq q \cdot R_\epsilon(T_0),\end{aligned}$$

where the two steps hold by definition of R_ϵ and the construction of the knockoff filter, respectively. Either way, then, it is sufficient to prove that $\mathbb{E}\left[R_\epsilon(T)\right] \leq e^\epsilon$, where T is either T_+ or T_0.

Next, given a threshold rule $T = T(W)$ mapping statistics $W \in \mathbb{R}^p$ to a threshold $T > 0$ (i.e. the knockoff or knockoff+ filter threshold, T_0 or T_+), for each index $j = 1, \dots, p$ we define

$$T_j = T\Big((W_1, \dots, W_{j-1}, |W_j|, W_{j+1}, \dots, W_p)\Big) > 0,$$

i.e. the threshold that we would obtain if W_j were replaced with $|W_j|$. The following lemma (proved in the Supplementary Material) establishes a property of the T_j's in the context of the knockoff filter:

Lemma 6. *Let $T = T(W)$ be the threshold for either the knockoff or the knockoff+.*[4] *For any j, k,*

$$\text{(21)} \qquad \textit{If } W_j \leq -\min\{T_j, T_k\} \textit{ and } W_k \leq -\min\{T_j, T_k\}, \textit{ then } T_j = T_k.$$

[4] More generally, this result holds for any function $T = T(W)$ that satisfies a "stopping time condition" with respect to the signs of the W_j's, defined as follows: for any $t > 0$, the event $\mathbb{1}\{T \leq t\}$ depends on W only through (1) the magnitudes $|W|$, (2) $\mathrm{sign}(W_j)$ for each j with $|W_j| < t$, and (3) $\sum_{j:|W_j|\geq t} \mathrm{sign}(W_j)$.

Now with T being either the knockoff or knockoff+ thresholding rule, we have

$$\begin{aligned}\mathbb{E}\left[R_\epsilon(T)\right] &= \mathbb{E}\left[\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}\left\{W_j \geq T, E_j \leq \epsilon\right\}}{1+\sum_{j\in\mathcal{H}_0}\mathbb{1}\left\{W_j \leq -T\right\}}\right]\\ &= \sum_{j\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{1}\left\{W_j \geq T_j, E_j \leq \epsilon\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right],\end{aligned}$$

where the last step holds since $T > 0$ by definition, so if $W_j \geq T$ then $W_j \not\leq -T$, and, by definition of T_j, we also have $T = T_j$ in this case. Continuing from this last step, we can rewrite the expectation as

$$\begin{aligned}\mathbb{E}\left[R_\epsilon(T)\right] &= \sum_{j\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{1}\left\{W_j > 0, E_j \leq \epsilon\right\}\cdot\mathbb{1}\left\{|W_j| \geq T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right]\\ &\overset{(*)}{=} \sum_{j\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{P}\left\{W_j > 0, E_j \leq \epsilon \mid |W_j|, W_{-j}\right\}\cdot\mathbb{1}\left\{|W_j| \geq T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right]\\ &\leq e^\epsilon\cdot\sum_{j\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{P}\left\{W_j < 0 \mid |W_j|, W_{-j}\right\}\cdot\mathbb{1}\left\{|W_j| \geq T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right]\\ &\overset{(*)}{=} e^\epsilon\cdot\sum_{j\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{1}\left\{W_j < 0\right\}\cdot\mathbb{1}\left\{|W_j| \geq T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right]\\ &= e^\epsilon\cdot\mathbb{E}\left[\sum_{j\in\mathcal{H}_0}\frac{\mathbb{1}\left\{W_j \leq -T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}}\right],\end{aligned}$$

where the two steps marked with (*) hold because T_j is a function of $|W_j|, W_{-j}$ by its definition, and so we can treat it as known when we condition on $|W_j|, W_{-j}$.

Finally, the summation inside the last expected value above can be simplified as follows: if for all null j, $W_j > -T_j$, then the sum is equal to zero, while otherwise, we can write

$$\begin{aligned}\sum_{j\in\mathcal{H}_0}\frac{\mathbb{1}\left\{W_j \leq -T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_j\right\}} &= \sum_{j\in\mathcal{H}_0}\frac{\mathbb{1}\left\{W_j \leq -T_j\right\}}{1+\sum_{k\in\mathcal{H}_0,k\neq j}\mathbb{1}\left\{W_k \leq -T_k\right\}}\\ &= \sum_{j\in\mathcal{H}_0}\frac{\mathbb{1}\left\{W_j \leq -T_j\right\}}{\sum_{k\in\mathcal{H}_0}\mathbb{1}\left\{W_k \leq -T_k\right\}} = 1,\end{aligned}$$

where the first step applies Lemma 6. Combining everything, we have shown that $\mathbb{E}\left[R_\epsilon(T)\right] \leq e^\epsilon$, which proves the theorem.

A.2. Proof of Lemma 2. We need to prove that

$$\mathbb{P}\left\{W_j > 0, \widehat{\mathrm{KL}}_j \leq \epsilon \;\middle|\; |W_j|, W_{-j}\right\} \leq e^{\epsilon} \cdot \mathbb{P}\left\{W_j < 0 \mid |W_j|, W_{-j}\right\}$$

for any null j and any $\epsilon \geq 0$. To proceed, we will be conditioning on observing $\mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}$, and on observing the *unordered* pair $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$—that is, we observe both the original and knockoff features but do not know which is which. It follows from the flip-sign property that having observed all this, we know all the knockoff statistics W except for the sign of the jth component W_j. Put differently, W_{-j} and $|W_j|$ are both functions of the variables we are conditioning on, but $\mathrm{sign}(W_j)$ is not. Without loss of generality, label the unordered pair of feature vectors $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$, as $\mathbf{X}_j^{(0)}$ and $\mathbf{X}_j^{(1)}$, such that:

$$\begin{cases} \text{If } \mathbf{X}_j = \mathbf{X}_j^{(0)} \text{ and } \widetilde{\mathbf{X}}_j = \mathbf{X}_j^{(1)}, \text{ then } W_j \geq 0; \\ \text{If } \mathbf{X}_j = \mathbf{X}_j^{(1)} \text{ and } \widetilde{\mathbf{X}}_j = \mathbf{X}_j^{(0)}, \text{ then } W_j \leq 0. \end{cases} \tag{22}$$

We can therefore write

$$\begin{aligned}&\mathbb{P}\left\{W_j > 0, \widehat{\mathrm{KL}}_j \leq \epsilon \;\middle|\; |W_j|, W_{-j}\right\} \\ &\qquad = \mathbb{E}\left[\mathbb{P}\left\{W_j > 0, \widehat{\mathrm{KL}}_j \leq \epsilon \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\} \;\middle|\; |W_j|, W_{-j}\right]\end{aligned}$$

and similarly

$$\begin{aligned}&\mathbb{P}\left\{W_j < 0 \mid |W_j|, W_{-j}\right\} \\ &\qquad = \mathbb{E}\left[\mathbb{P}\left\{W_j < 0 \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\} \;\middle|\; |W_j|, W_{-j}\right].\end{aligned}$$

Therefore, it will be sufficient to prove that

$$\begin{aligned}&\mathbb{P}\left\{W_j > 0, \widehat{\mathrm{KL}}_j \leq \epsilon \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\} \\ &\qquad \leq e^{\epsilon} \cdot \mathbb{P}\left\{W_j < 0 \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}.\end{aligned} \tag{23}$$

Now, if $\mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}$ are such that $|W_j| = 0$, clearly this bound holds trivially, so from this point on we ignore this trivial case and assume that $|W_j| > 0$. By our definition (22) of $\mathbf{X}_j^{(0)}$ and $\mathbf{X}_j^{(1)}$, we have

$$\begin{aligned}&\frac{\mathbb{P}\left\{W_j > 0 \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}}{\mathbb{P}\left\{W_j < 0 \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}} \\ &\qquad = \frac{\mathbb{P}\left\{(\mathbf{X}_j, \widetilde{\mathbf{X}}_j) = (\mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}) \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}}{\mathbb{P}\left\{(\mathbf{X}_j, \widetilde{\mathbf{X}}_j) = (\mathbf{X}_j^{(1)}, \mathbf{X}_j^{(0)}) \;\middle|\; \mathbf{X}_j^{(0)}, \mathbf{X}_j^{(1)}, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}},\end{aligned} \tag{24}$$

where this last ratio should be interpreted as a ratio of conditional probabilities or conditional densities, as appropriate. Since the observations $i = 1, \dots, n$ are independent, this can be rewritten as

$$\prod_{i=1}^{n} \frac{\mathbb{P}\left\{(\mathbf{X}_{ij}, \widetilde{\mathbf{X}}_{ij}) = (\mathbf{X}^{(0)}_{ij}, \mathbf{X}^{(1)}_{ij}) \,\middle|\, \mathbf{X}^{(0)}_{ij}, \mathbf{X}^{(1)}_{ij}, \mathbf{X}_{i,-j}, \widetilde{\mathbf{X}}_{i,-j}, \mathbf{Y}_i\right\}}{\mathbb{P}\left\{(\mathbf{X}_{ij}, \widetilde{\mathbf{X}}_{ij}) = (\mathbf{X}^{(1)}_{ij}, \mathbf{X}^{(0)}_{ij}) \,\middle|\, \mathbf{X}^{(0)}_{ij}, \mathbf{X}^{(1)}_{ij}, \mathbf{X}_{i,-j}, \widetilde{\mathbf{X}}_{i,-j}, \mathbf{Y}_i\right\}} \tag{25}$$
$$= \prod_{i=1}^{n} \frac{P^\star_j(\mathbf{X}^{(0)}_{ij} \mid \mathbf{X}_{i,-j}) \cdot P_j(\mathbf{X}^{(1)}_{ij} \mid \mathbf{X}_{i,-j})}{P_j(\mathbf{X}^{(0)}_{ij} \mid \mathbf{X}_{i,-j}) \cdot P^\star_j(\mathbf{X}^{(1)}_{ij} \mid \mathbf{X}_{i,-j})} =: e^{\rho_j},$$

where the first equality holds by Lemma 1 (recalling that j is assumed to be a null feature). Next, from the definition (13) of $\widehat{\mathrm{KL}}_j$ and the definition (22) of $\mathbf{X}^{(0)}_j$ and $\mathbf{X}^{(1)}_j$, we can see that $\widehat{\mathrm{KL}}_j = \rho_j$ if $W_j > 0$, or otherwise $\widehat{\mathrm{KL}}_j = -\rho_j$ if $W_j < 0$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{W_j > 0, \widehat{\mathrm{KL}}_j \le \epsilon \,\middle|\, \mathbf{X}^{(0)}_j, \mathbf{X}^{(1)}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}\\
&= \mathbb{P}\left\{W_j > 0, \rho_j \le \epsilon \,\middle|\, \mathbf{X}^{(0)}_j, \mathbf{X}^{(1)}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}\\
&= \mathbb{1}\left\{\rho_j \le \epsilon\right\} \cdot \mathbb{P}\left\{W_j > 0 \,\middle|\, \mathbf{X}^{(0)}_j, \mathbf{X}^{(1)}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\}\\
&= \mathbb{1}\left\{\rho_j \le \epsilon\right\} \cdot e^{\rho_j} \cdot \mathbb{P}\left\{W_j < 0 \,\middle|\, \mathbf{X}^{(0)}_j, \mathbf{X}^{(1)}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}\right\},
\end{aligned}$$

where the next-to-last step holds since ρ_j is a function of $\mathbf{X}^{(0)}_j, \mathbf{X}^{(1)}_j, \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}, \mathbf{Y}$, while the last step uses our work in (24) and (25). Since $\mathbb{1}\left\{\rho_j \le \epsilon\right\} \cdot e^{\rho_j} \le e^{\epsilon}$ trivially, we have proved the desired bound (23), which concludes the proof of the lemma.

A.3. Proof of Lemma 1. We prove the lemma in the case where all features are discrete; the case where some of the features may be continuous is proved analogously. First, consider any null feature index j. By definition of the nulls, we know that $X_j \perp\!\!\!\perp Y \mid X_{-j}$. Furthermore, $\widetilde{X} \perp\!\!\!\perp Y \mid X$ by construction. Therefore, the distribution of $Y \mid (X, \widetilde{X})$ depends only on X_{-j}, and in particular, $Y \perp\!\!\!\perp (X_j, \widetilde{X}_j) \mid (X_{-j}, \widetilde{X}_{-j})$. This proves that

$$\frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b \,\middle|\, X_{-j}, \widetilde{X}_{-j}, Y\right\}}{\mathbb{P}\left\{X_j = b, \widetilde{X}_j = a \,\middle|\, X_{-j}, \widetilde{X}_{-j}, Y\right\}} = \frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b \,\middle|\, X_{-j}, \widetilde{X}_{-j}\right\}}{\mathbb{P}\left\{X_j = b, \widetilde{X}_j = a \,\middle|\, X_{-j}, \widetilde{X}_{-j}\right\}}, \tag{26}$$

because the numerator and denominator are each unchanged whether we do or do not condition on Y. Thus, for null features j, it is now sufficient to prove only the first claim of the lemma, namely that the right-hand side above is equal to $\frac{P^\star_j(a|X_{-j})P_j(b|X_{-j})}{P_j(a|X_{-j})P^\star_j(b|X_{-j})}$.

From this point on, let j be any feature (null or non-null). We will now prove the first claim in the lemma. Recalling the assumption that $P_{\widetilde{X}|X}$ is pairwise exchangeable with respect to P_j (5), we introduce a pair of random variables drawn as follows: first, draw $X'_{-j} \sim P^\star_{X_{-j}}$, where $P^\star_{X_{-j}}$ is the distribution of X_{-j}; then draw $X'_j \mid X'_{-j} \sim P_j(\cdot|X'_{-j})$; and finally, draw $\widetilde{X}' \mid X' \sim P_{\widetilde{X}|X}(\cdot|X')$. Then by (5),

$$\big(X'_j, \widetilde{X}'_j, X'_{-j}, \widetilde{X}'_{-j}\big) \stackrel{\mathrm{d}}{=} \big(\widetilde{X}'_j, X'_j, X'_{-j}, \widetilde{X}'_{-j}\big). \tag{27}$$

By construction, the joint distribution of $(X', \widetilde{X}')$ is given by

$$\mathbb{P}\left\{X' = x, \widetilde{X}' = \widetilde{x}\right\} = P^\star_{X_{-j}}(x_{-j})P_j(x_j\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}\,|\,x).$$

Now, fixing any $x_{-j}, \widetilde{x}_{-j} \in \mathbb{R}^{p-1}$, write x^a as the vector in $\mathbb{R}^p$ with entry j given by a and all other entries given by x_{-j}, and define $x^b, \widetilde{x}^a, \widetilde{x}^b$ analogously. Then (27) is equivalent to

$$P^\star_{X_{-j}}(x_{-j})P_j(a\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}^b\,|\,x^a) = P^\star_{X_{-j}}(x_{-j})P_j(b\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}^a\,|\,x^b). \tag{28}$$

Now we turn to the true distribution of the data, generated as $X \sim P^\star_X$ and $\widetilde{X} \mid X \sim P_{\widetilde{X}|X}$. This means that the joint distribution of $(X, \widetilde{X})$ is given by

$$\mathbb{P}\left\{X = x, \widetilde{X} = \widetilde{x}\right\} = P^\star_{X_{-j}}(x_{-j})P^\star_j(x_j\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}\,|\,x).$$

We can therefore calculate

$$\frac{\mathbb{P}\left\{X_j = a, \widetilde{X}_j = b, X_{-j} = x_{-j}, \widetilde{X}_{-j} = \widetilde{x}_{-j}\right\}}{\mathbb{P}\left\{X'_j = a, \widetilde{X}'_j = b, X'_{-j} = x_{-j}, \widetilde{X}'_{-j} = \widetilde{x}_{-j}\right\}}$$
$$= \frac{P^\star_{X_{-j}}(x_{-j})P^\star_j(a\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}^b\,|\,x^a)}{P^\star_{X_{-j}}(x_{-j})P^\star_j(b\,|x_{-j})P_{\widetilde{X}|X}(\widetilde{x}^a\,|\,x^b)} = \frac{P^\star_j(a \mid x_{-j})}{P_j(a \mid x_{-j})} \cdot \frac{P_j(b \mid x_{-j})}{P^\star_j(b \mid x_{-j})},$$

where the last step holds by (28). This proves the lemma.

Department of Statistics
The University of Chicago
Chicago, IL, U.S.A.
E-mail: rina@uchicago.edu

Departments of Statistics and Mathematics
Stanford University
Stanford, CA, U.S.A.
E-mail: candes@stanford.edu

Statistical Laboratory
University of Cambridge
Cambridge, U.K.
E-mail: r.samworth@statslab.cam.ac.uk