Knockoffs for the Mass: New Feature Importance Statistics with False Discovery Guarantees

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

Amirata Ghorbani
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
amiratag@stanford.edu

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

Abstract

An important problem in machine learning and statistics is to identify features that causally affect the outcome. This is often impossible to do from purely observational data, and a natural relaxation is to identify features that are correlated with the outcome even conditioned on all other observed features. For example, we want to identify that smoking really is correlated with cancer conditioned on demographics. The knockoff procedure is a recent breakthrough in statistics that, in theory, can identify truly correlated features while guaranteeing that false discovery rate is controlled. The idea is to create synthetic data—knockoffs—that capture correlations among the features. However, there are substantial computational and practical challenges to generating and using knockoffs. This paper makes several key advances that enable knockoff application to be more efficient and powerful. We develop an efficient algorithm to generate valid knockoffs from Bayesian Networks. Then we systematically evaluate knockoff test statistics and develop new statistics with improved power. The paper combines new mathematical guarantees with systematic experiments on real and synthetic data.

1 INTRODUCTION

Identifying important features is an ubiquitous problem in machine learning and statistics. In relatively simple settings, the importance of a given feature is measured via the fitted parameters of some model. Consider Generalized Linear Models (GLM) for example, users often add a LASSO penalty (Tibshirani, 1996) to promote sparsity in the coefficients, and subsequently select those features with non-zero coefficients. Step-wise procedures where we sequentially modify the model are another way of doing feature selection (Mallows, 1973; Akaike, 1974; Raftery, 1986).

These standard methods are all plagued by correlations between the features: a feature that is not really relevant for the outcome, in a precise sense we will define in Section 2, can be selected by LASSO or Step-wise procedure, because it is correlated with relevant features. The feature selection problem becomes even more difficult in more complex settings where we may not have a clean parametric model relating input features to the labels. Moreover, in these settings we usually lack statistical guarantees on the validity of the selected features. Finally, even procedures with statistical guarantees usually depend on having *valid p-values*, which are based on a correct modeling of $Y|X$ and (sometimes) assume some asymptotic regime. However, there are many common settings where these assumptions fail and we cannot perform inference based on those p-values (Sur and Candès, 2018).

A powerful new approach called *Model-X knockoff procedure* (Candès et al., 2018) has recently emerged to deal with these issues. This method introduces a new paradigm: we no longer assume any model for the distribution of $Y|X$ in order to do feature selection (and therefore do not compute p-values), but we assume that we have full knowledge of the feature distribution P^X – or at least we can accurately model it. This knowledge

of the ground truth P^X allows us to sample new *knock-off* variables $\tilde{X}$ satisfying some precise distributional conditions. Although we make no assumption on $Y|X$, we can run any procedure on the data set extended with the knockoffs and perform feature selection while controlling the False Discovery Rate (FDR). Statistical guarantees no longer rely on *valid p-values* whenever our model for $Y|X$ is correct, but exclusively on having *valid knockoffs*, obtained whenever we know P^X (or our model for P^X is correct).

There are two main obstacles to using knockoffs in practice (i.e. "for the mass"): 1) tractably generating valid knockoffs and 2) once we have generated knockoffs, computing powerful test statistics. Current tractable methods for generating knockoffs are restricted to the settings where X is modeled as a multivariate Gaussian (Candès et al., 2018) or as the set of observed nodes in a Hidden Markov Model (HMM) (Sesia et al., 2017). Even though knockoff methods enjoy some robustness properties when approximating P^X (Barber et al., 2018), the guarantee on FDR control breaks down even in very simple settings if these approximations turn out to be too crude, as we demonstrate in this paper. Sequential conditional independent pairs algorithm (SCIP) (Candès et al., 2018) is a universal knockoff sampling scheme but is computationally intractable as soon as we model P^X by more complex distributions. Constructing tractable knockoff sampling procedures for more flexible classes of distributions used to model P^X is essential to improve the validity of knockoffs, which is always subject to the quality of the approximation. Regarding test statistics, current methods (Barber et al., 2015; Candès et al., 2018; Sesia et al., 2017; Barber et al., 2018) focus on LASSO-based statistics to obtain proxies for the importance of any given feature: such methods need to fit a GLM to the data, which is too restrictive in many supervised learning tasks.

Our Contributions. First, we formulate a novel tractable procedure for sampling knockoffs in settings where features can be modeled as the observed variables in a Bayesian Network. This allows for great flexibility when modeling the feature distribution P^X. We show that this procedure is different from SCIP, which is the current state-of-the-art method to sample knockoffs. We construct valid knockoffs in settings where previous knockoff sampling procedures assumed a very restrictive model for P^X and therefore failed to control FDR, and provide a unified framework for several different sampling procedures. In addition, we systematically evaluate and compare several nonlinear knockoff test statistics which can be applied in general supervised learning problems. We develop a new statistic, *Swap Integral*, which has significantly better power than other

methods. These two advances enable us to perform feature selection using black-box classifiers to represent $Y|X$ while still retaining statistical guarantees under correct modeling of P^X.

2 KNOCKOFF BACKGROUND

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

Assuming we know the ground truth for the distribution P^X (we'll come back to the realistic setting where we only have samples of X), the first step of the knockoff procedure is to obtain a *knockoff* sample $\tilde{X}$ that satisfies the following conditions:

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

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

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

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

The first condition is immediately satisfied as long as knockoffs are sampled conditioned on the sample X without considering any information about Y, which will always be the case in our sampling methods. More generally we say that any $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $f([x, \tilde{x}]_{swap(S)}) = f([x, \tilde{x}])$ satisfies the exchangeability property.

Once we have the knockoff sample $\tilde{X}$, the next step of the procedure constructs *feature statistics* $\boldsymbol{W} = (W_1, \ldots, W_d)$, such that a high value for $W_j \in \mathbb{R}$ is evidence that the jth feature is non-null. Feature statistics described by Candès et al. (2018) depend

only on $[X, \tilde{X}] \in \mathbb{R}^{N \times 2d}, Y \in \mathbb{R}^N$ such that for each $j \in \{1, \dots, d\}$ we can write $W_j = w_j([X, \tilde{X}], Y)$ for some function w_j. The only restriction these statistics must satisfy is the *flip-sign property*: swapping the jth feature and its corresponding knockoff feature should flip the sign of the statistic W_j while leaving other feature statistics unchanged. More formally, for a subset $S \subset \{1, \dots, d\}$ of features, denoting $[X, \tilde{X}]_{swap(S)}$ the data matrix where the original jth variable and its corresponding knockoff have been transposed whenever $j \in S$, we require that w_j satisfies:

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

As suggested by Candès et al. (2018), the choice of feature statistics can be done in two steps: first, find a statistic $Z = Z([X, \tilde{X}], Y) = (Z_1, \dots, Z_d, \tilde{Z}_1, \dots, \tilde{Z}_d) \in \mathbb{R}^{2d}$ where each coordinate corresponds to the "importance" — hence we will call them *importance scores* — of the corresponding feature (either original or knockoff). For example, Z_j would be the absolute value of the regression coefficient of the jth feature.

After obtaining the importance score for the original and knockoff feature, we compute the feature statistic $W_j = Z_j - \tilde{Z}_j$. The intuition is that importance scores of knockoffs serve as a control, such that larger importance score of the jth feature compared to that of its knockoff implies larger W_j (and therefore is evidence against the null). Given that feature statistics of the null features behave "symmetrically" around 0 by the flip-sign property, the final selection procedure compares, at a given threshold $q > 0$ the number of features such that $W_j > q$ vs. the number of those such that $W_j < -q$ (Candès et al., 2018). By keeping this ratio "under control" for a given target FDR that we fix in advance, we obtain an adaptive procedure that selects features whose value W_j is above some adaptive threshold. FDR control is guaranteed with this procedure, as long as the knockoffs are sampled correctly (i.e., satisfying the two conditions), regardless of the choices made to construct the feature statistics W (Candès et al., 2018).

The power of the knockoff procedure is that it makes *no assumptions* on the relationship between X and Y; however, it requires full knowledge of the feature distribution P^X which is often not known and needs to modeled. It is also important to notice that for any model we use for P^X, we are required to know how to sample knockoffs from that model distribution. Current classes of distributions having a tractable knockoff sampling method have limited expressivity, and we can find simple settings where using models limited to these classes fails to provide valid knockoffs, and consequently the FDR is not controlled. We provide

an example of such setting in Section 5. However, the knockoff procedure is robust to small errors in the estimation of P^X (Barber et al., 2018). Our method for sampling knockoffs from a Bayesian network allows us to fit a better distribution to the data resulting in a better control of FDR up to some error term with respect to the FDR we target (under the assumption that our model distribution of P^X is perfect, this procedure controls FDR at the target FDR). We discuss in Appendix B the reasons why we expect our method to be robust to model misspecification.

Note that setting $\tilde{X} = X$ satisfies the knockoff properties, but no discovery would be made since all of the W_j would be zero. Therefore, in addition to generating valid knockoffs in order to control FDR, it is also important to consider the statistical power of the procedure, which is the probability that a true alternative is selected. This depends not only on the knockoff sampling procedure, but also on the choice of the importance scores. We discuss power in detail in Sections 4 and 5.

3 GENERATING KNOCKOFFS FROM LATENT VARIABLE MODELS

We first show how to generate knockoff samples from a Gaussian mixture model, which is already a new contribution, as a warm-up to our general procedure to generating knockoffs from latent variable models. This section relates just to the first step of the *Model-X* knockoff procedure. We will consider different classes of distributions used to model P^X; we do not consider the response Y for this step.

Gaussian Mixture Knockoffs   The setup:

- K is a categorical latent variable taking values in $\{1, \dots, l\}$, where $l \geq 2$ is the number of mixtures.
- $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are the mean and variances for component k, that are known or estimated.
- We observe a $\mathbb{R}^d$-valued random variable X with distribution

$$X|K = k \sim P^{X|K}(x|K = k) = \mathcal{N}(x; \mu_k, \Sigma_k)$$

For a multivariate Gaussian distribution, we can explicitly write a joint density $P^{X, \tilde{X}|K}$ on $\mathbb{R}^d \times \mathbb{R}^d$ that satisfies the exchangeability property for a given K:

$$P^{X, \tilde{X}|K}(x, \tilde{x}|K = k) =$$

$$\mathcal{N}\left(\begin{pmatrix} x \\ \tilde{x} \end{pmatrix}; \begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \Sigma_k & \Sigma_k - D_k \\ \Sigma_k - D_k & \Sigma_k \end{pmatrix} \right)$$

D_k is a diagonal matrix with the constraint that the joint covariance matrix is positive definite (Barber

et al., 2015). The Gaussian mixture knockoff sampling procedure consists of first sampling the mixture assignment variable from the posterior distribution $P^{K|X}$. The knockoff is then sampled from the conditional distribution of the knockoff given the true variable and the sampled mixture assignment defined as $Q^{\tilde{X}|X,K}(\tilde{x}|x, K = k)$. More explicitly:

$$K|X \sim P^{K|X}(k|X) = \frac{\lambda_k \mathcal{N}(X; \mu_k, \Sigma_k)}{\sum_{j=1}^{l} \lambda_j \mathcal{N}(X; \mu_j, \Sigma_j)}$$

$$\tilde{X}|X, K = k \sim Q^{\tilde{X}|X,K}(\tilde{x}|X, K = k) = \mathcal{N}(\tilde{x}; \tilde{\mu}_k, \tilde{\Sigma}_k)$$

$$\text{where} \begin{cases} \tilde{\mu}_k = D_k \Sigma_k^{-1} \mu_k + (I_d - D_k \Sigma_k^{-1})X \\ \tilde{\Sigma}_k = 2D_k - D_k \Sigma_k^{-1} D_k \end{cases}$$

$$\text{as } P^{X,\tilde{X}|K} = Q^{\tilde{X}|X,K} P^{X|K}.$$

Proposition 3.1. *The Gaussian mixture knockoff sampling procedure stated above is such that $\tilde{X}$ is a valid knockoff of X.*

The proof is deferred to Appendix A.2. Once we compute the necessary parameters $\tilde{\mu}_k, \tilde{\Sigma}_k$ of these conditional distributions, we can sample valid knockoffs for a Gaussian mixture model. Whenever we no longer know the ground truth for P^X, Gaussian mixtures can approximate arbitrarily well any continuous distribution (Bacharoglou, 2010), so we now have a very flexible model for the feature distribution P^X from which we know how to tractably sample knockoffs. In a real setting, the validity of our knockoffs will be limited by the quality of the approximation of our model for P^X – the same way in a parametric model of $Y|X$ the validity of the p-values depends on the validity of model and distributional assumptions (of the eventual random noise, on the asymptotic regime, etc). Next we show how to efficiently generate valid knockoffs for general Bayesian Networks, which can be used to accurately model P^X in many settings.

Exchangeable Conjugate Our knockoff sampling procedures will be based on the following idea. Instead of keeping long range dependencies across all features as SCIP does, we want to exploit the structure given by the local conditional distributions of a Bayesian network: sampling knockoffs becomes as hard as doing probabilistic inference on the Bayesian network, and sampling "local" knockoffs based on the local structure.

Definition 3.1. *Let $U \in \mathcal{U}$, $V \in \mathcal{V}$ be arbitrary multidimensional random variables. For any given conditional distribution $P^{U|V}(\cdot|v)$, we say that $Q^{\tilde{U}|U,V}(\cdot|u, v)$ is an* exchangeable conjugate conditional *if it is a probability kernel on $(\mathcal{U} \times \mathcal{V}) \times \mathcal{U}$ such that, for any fixed v, $\phi_v(u, \tilde{u}) := P^{U|V}(u|v)Q^{\tilde{U}|U,V}(\tilde{u}|u, v)$ satisfies the exchangeability property in $(u, \tilde{u})$.*

In other words, if U is sampled conditionally on V, and then $\tilde{U}$ is sampled from the probability kernel $Q^{\tilde{U}|U,V}$, then $\tilde{U}$ is a knockoff of U (conditionally on V). In some cases it is easy to find examples of such a conjugate conditional distribution. If U is univariate, then $Q^{\tilde{U}|U,V}(\cdot|u, v) = P^{U|V}(\cdot|v)$ is a valid conjugate conditional. More generally, this choice is valid if the coordinates of U are independent conditionally on V. We can also construct valid conjugates when the distribution of U conditionally on V is Gaussian (Candès et al., 2018) or a Markov chain (Sesia et al., 2017).

Sampling Knockoffs for Bayesian Networks We define a procedure for sampling knockoffs when we assume a generative model for P^X. X corresponds to the observed variables in a latent variable model which we assume can be represented as a Bayesian Network (BN): Gaussian Mixture Models, Hidden Markov Models (HMM) (Poritz, 1988), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Naive Bayes (Friedman et al., 1997) and many other Bayesian models fit this description. Consider a BN defined by the directed acyclic graph (DAG) (G, E) with cardinality $m = |G|$ where the index is a topological ordering. X_O and X_H refer respectively to the observed variables and the hidden ones: we have restated our initial set of features X as X_O, and we want to construct knockoffs by leveraging our knowledge about the DAG. The only restriction is that the observed variables must correspond to nodes in the graph without descendents. It is important to notice that here, the variables associated to a node of the DAG *can be multi-dimensional*, which sometimes simplifies the sampling procedure (an example are knockoffs for HMM, detailed in Appendix A.3).

We assume that we can sample from $P^{H|O}$ (for simplicity we refer to the random variables by their indices), the conditional probability of the hidden variables given the observed ones. Assume also that, for every node i, we can compute the local conditional probability $P^{i|MB(i)}$ of X_i given its Markov blanket $X_{MB(i)}$, and that we can compute a knockoff conjugate distribution $Q^{\tilde{i}|i,MB(i)}$ of $P^{i|MB(i)}$ and sample $\tilde{X}_i$ from it, which will be a "local" knockoff. We describe the knockoff generating procedure in Algorithm 1 and further discuss the assumptions on the inputs on Appendix A.3.

Theorem 3.2. *The distribution of the concatenated random variables $(X_O, \tilde{X}_O)$ satisfies the exchangeability property. That is, $\tilde{X}_O$ is a valid knockoff of X_O.*

The proof is in Appendix A.4. This sampling method is not equivalent to the Sequential Conditional Independent Pairs (SCIP) introduced in (Candès et al., 2018). Consider the simple case with one observed multi-dimensional variable X and one hidden variable H. SCIP would sample each coordinate $\tilde{X}_i$ of $\tilde{X}$ se-

Algorithm 1: Knockoff Sampling Procedure in BN

Input : Initial sample X_O, conditional distribution $P^{H|O}$, conjugate conditionals $Q^{\tilde{i}|i,MB(i)}$ for every node i.

Output : Knockoff sample $\tilde{X}_O$

1 Sample $X_H \sim P^{H|O}(\cdot|X_O)$;
2 **for** $i = 1$ **to** m **do**
3   Sample $\tilde{X}_i \sim$
     $Q^{\tilde{i}|i,MB(i)}(\cdot|\tilde{X}_{\{1:i-1\}\cap MB(i)}, X_{\{i+1:m\}\cap MB(i)}, X_i)$
4 **end**
5 **return** $\tilde{X}_O = (\tilde{X}_i)_{i\in O}$

quentially by computing (for each knockoff sample) conditional distributions of the joint probability distribution of $(X, \tilde{X}_{j\leq i})$, without leveraging our knowledge of the generative process. That is not computationally feasible in an arbitrary generative model. Another difference with respect to SCIP is that our sampling procedure is not creating a knockoff from the full set of hidden and observed variables: not all "local" knockoffs are retained in the final output. That is, $(\tilde{X}_O, \tilde{X}_H)$ is not a valid knockoff for (X_O, X_H). We refer to Appendix A.3 for more details on these differences between methods.

The HMM knockoff sampling procedure (Sesia et al., 2017) and the Gaussian mixture sampling above are special cases of Algorithm 1 (detailed in Appendix A.3). Naive Bayes would have a straightforward knockoff sampling scheme (as the observed nodes are independent conditionally on the hidden node): given the observed nodes, generate a sample for the hidden node (based on the posterior distribution), then sample independent knockoffs based on the sampled hidden node. We provide in Appendix A.3 an additional example of a DAG knockoff generating procedure based on the LDA model.

4 POWER ANALYSIS: EVALUATING IMPORTANCE SCORE METHODS

The previous section shows how to generate valid knockoffs $\tilde{X}$ for complex latent variable distributions. The next challenge is to compute importance scores for each original and knockoff feature; this was denoted as Z_j and $\tilde{Z}_j$ in Section. 2 (recall that the feature statistics is $W_j = Z_j - \tilde{Z}_j$). While the knockoff procedure guarantees that the FDR is controlled for any importance score, the power of the procedure—i.e. how many true non-null features are discovered—is highly dependent on the method for computing the importance score. In this section we discuss several importance score meth-
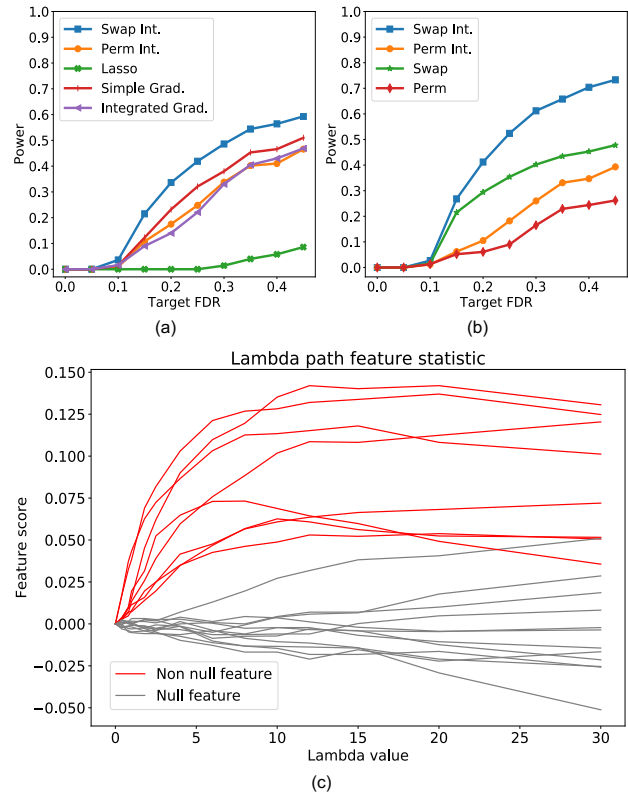

Figure 1: **Comparison Between Importance Score Methods** (a) Synthetic data where $Y|X$ is a polynomial. Using a linear LASSO regression model to generate importance scores results in near-zero power; our proposed Swap Integral statistic achieves higher power across all the target FDRs. (b) Swap Integral and Permutation Integral achieve higher power compared respectively to basic Swap and Permutation. (c) Example of lambda path for Swap method.

ods applicable to a large family of predictive models and also define new ones.

LASSO Coefficient-Difference (LCD) were used as feature statistics by Candès et al. (2018). However, for non-linear $Y|X$, using importance scores based on a linear model assumption will result in low power as LASSO is not able to fit the real input-output relationship (Figure 1(a)). Complex unknown non-linear input-output relationship is an important restriction if we want to use a parametric model to obtain importance scores through interpretable parameters.

Saliency Methods If $Y|X$ is fitted using a neural network, then one could compute saliency scores, a popular metric for importance scores (Lipton, 2016; Ghorbani et al., 2017). Saliency is computed for each feature of each example, and averaged across all the examples to derive a global importance score. In our

experiments, we use Gradient (Baehrens et al., 2010; Simonyan et al., 2013) and Integrated Gradients (Sundararajan et al., 2017) saliency methods.

Importance Scores Based on Accuracy Drops
After a model is trained on the dataset $([X, \tilde{X}], Y)$, for each column of $[X, \tilde{X}] \in \mathbb{R}^{N \times 2d}$, we create a new "fake" feature column resulting in the "fake" set of features $[X^f, \tilde{X}^f]$, for example by permuting across samples column-wise. Replacing each feature with its fake version one at a time, the accuracy drop can be used as the importance score Z_j for that feature.

Note that achieving a high performance of the classifier on the initial dataset is not necessary to generate these importance scores, as our analysis is based on the accuracy drops. Classifiers with low predictive power can still be useful at determining non-null features, and although they may not make many discoveries, FDR is still controlled.

Permutation and Swapping Originally introduced for Random Forests (Breiman, 2001), the **Permutation** method creates each column of $[X^f, \tilde{X}^f]$ by simply shuffling the entries of the corresponding column in $[X, \tilde{X}]$. Although the permutation method may be effective at creating a new "fake" feature independently from Y, it comes with an important issue: when evaluating the classifier on the dataset where just one of the features has been shuffled, for general distributions of the features, we may end up with fake data points that are completely off the original data distribution and as a result outside of the learned manifold of the predictive model. The classifier's predictions on these *off-distribution* regions of the input space where it hasn't been trained may be arbitrary. As a result, the performance can potentially drop across all features and their knockoffs, regardless of whether they actually had an impact on Y. A simple illustration of this phenomenon has been provided in Appendix E.

The main problem occurs with the non-null features, as an unexpectedly large drop in accuracy for the control knockoff feature would mask the non-null importance score, hence limiting the possibility that it is selected. When the original distribution P^X of the features is a mixture of Gaussians, permutation across any of the features will create data points that do not lie on any of the mixture components of P^X, and as the number of mixtures increases, the chance for a fake data point to be off-distribution increases. We confirm this phenomenon with synthetic data experiments Figure 3(a-c). Power for the Permutation method strictly decreases as the number of mixture components grows.

To tackle this problem, we introduce the **Swap** method for importance scores. Instead of shuffling, for each

original feature column, we use its knockoff corresponding column as its fake replacement and the other way around: $X^f = \tilde{X}$ and $\tilde{X}^f = X$. Keeping in mind the exchangeability condition when generating knockoffs, this method guarantees that replacing each feature with its fake will result in data points that still belong to the original data distribution, and therefore the problem with fake data points moving to regions of the space unknown to the predictive model is mitigated.

Path Integration Generates Importance Scores with Better Power For a given parameter $\lambda \geq 0$, we compute for each original feature the fake replacement $X_j^f(\lambda) = X_j + \lambda(\tilde{X}_j - X_j)$ (and reciprocally for each knockoff feature). The basic Swap (or Permutation) method corresponds to $\lambda = 1$, and we plot the feature statistics W as a function of the parameter λ, which we call *lambda path*. We give an example of one such path in Figure 1(c) for a simulation where the classifier is a neural network with 3 fully connected layers, and Y is categorical with 10 classes such that $Y|X$ corresponds to the level sets of a random polynomial of degree 3 of the non-null features, and X comes from a mixture of 10 Gaussians. Similarly, we can also take lambda path for the Permutation method, where we compute $X_j^f(\lambda)$ by taking $\tilde{X}_j$ as the shuffled feature X_j. However, our lambda path quickly becomes unstable as for $\lambda = 1$ we already get out of the data manifold.

Basic Swap and Permutation (which correspond to $\lambda = 1$) do not always give the best separation between the null features (grey) and the alternate features (red). This motivates taking the integral of the lambda path and use the area under the curve as the feature statistic W—we call this **Swap Integral** and, similarly, **Permutation Integral**. For all of our experiments, we integrate from $\lambda = 0$ to $\lambda = 10$. Figure 1(b) shows that, for the same data as used in Figure 1(c), Swap Integral and Permutation Integral methods are respectively much more powerful than the basic Swap and Permutation.

Our goal is to identify the classifier's decision boundaries within the data manifold (thus likely to identify relevant features). When considering a lambda path, all comes down to choosing an appropriate direction when translating each feature, and keeping track of the decision boundaries crossed. We may temporarily step out of the data manifold (indeed there is no guarantee that the manifold is convex, our Gaussian mixture model is especially well-suited for this) –and eventually will be out of it. The integral (up to $\lambda = 10$) assigns a higher importance to boundaries crossed early on. As the Swap Integral method "stays longer" in the data manifold early on ($\lambda = 1$ is still in it), it performs better than when choosing an arbitrary direction (such

as one given by a shuffled vector, i.e. the Permutation Integral). Knockoffs not only are useful for FDR control, they also contain information about the data manifold. The outcome of the following experiments whenever we use Swap Integral does not substantially change whenever λ stays in the range 5 to 15.

5 EXPERIMENTS

We start by precising that we will not be able to compare our DAG knockoff procedure (for the Gaussian Mixture Model) with the more general SCIP procedure, because the latter ~~very simple model~~

The Need for ~~resentative exam~~ ~~ply using the emp~~ ~~the knockoff—i.e.~~ ~~yields invalid knoc~~

We generate data ~~Gaussian random~~ ing P^X by a single Gaussian vector yields an empirical covariance matrix which is dominated by the diagonal (and therefore considers P^X as having independent coordinates). We generate a binary outcome Y in a logistic regression setting (such that the log odds of $Y = 1$ is a linear combination of the features of X). The linear combination is such that just the first 10 features of X are non-null (i.e. have non zero coefficient, red), the next 10 (null, black) features are correlated with the non-nulls, and the last 10 are independent nulls. During the knockoff procedure, the importance scores Z are the absolute value of the logistic regression coefficients of each feature and the feature statistics W are the difference of importance scores.

We run our knockoff procedure, modeling P^X as a mixture of three multivariate Gaussian variables fitted via EM (Figure 2(a)), and as a single Gaussian (Figure 2(b)). Importance scores Z_j of the non-null original features are high in both cases, and those of the correlated null original features are also quite high. Importance scores of valid knockoffs cancel out the correlated null original features' importance scores, and therefore the adaptively chosen threshold avoids selecting nulls (Figure 2(a)). However, when modeling by one mixture, the knockoffs corresponding to the null features (10-20) have importance scores close to 0 (Figure 2(b)). As a result these correlated null features are selected by the procedure and the FDR control is broken. After 40 repetitions of the procedure (randomly generating each run the parameters of P^X and sampling X), for a target FDR of 0.2, we get FDR control when we use our mixture model knockoff (em-
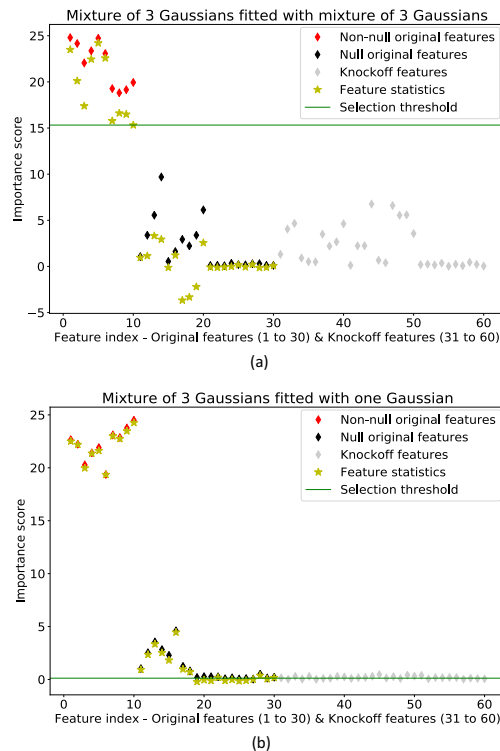

(a)

(b)

Figure 2: **FDR Control Fails for Non Valid Knockoffs:** We plot side by side the importance scores for each feature (original and knockoff, 60 in total) for one run of the simulation, and the feature statistics computed as the difference between importance scores of original vs. knockoff features (plotted in the first 30 indices). Selected features are those whose feature statistic W is above the threshold.

pirical FDR: 0.1823) and fail to control FDR whenever we fit the data with one single multivariate Gaussian (empirical FDR: 0.5565).

Feature Selection in Synthetic Data We generate P^X as mixtures of 5, 10 and 20 multivariate Gaussians, and Y as polynomial functions of a subset of the dimensions of X—i.e. these are the non-null features. From each P^X, we draw samples of X and Y, fit a mixture of Gaussians and run our knockoff procedure. Figure 3 (a-c) shows the power of five importance score methods, and Swap Integral consistently achieves the best power. In all of our experiments, the empirical FDR is controlled at the target level (Appendix Figure 5).

Feature Selection in Real Data Our real-world simulations are based on three datasets from the UK biobank data and the Bank Marketing and Polish Bankruptcy from the UCI repository (Dheeru and Karra Taniskidou, 2017). We provide more informa-
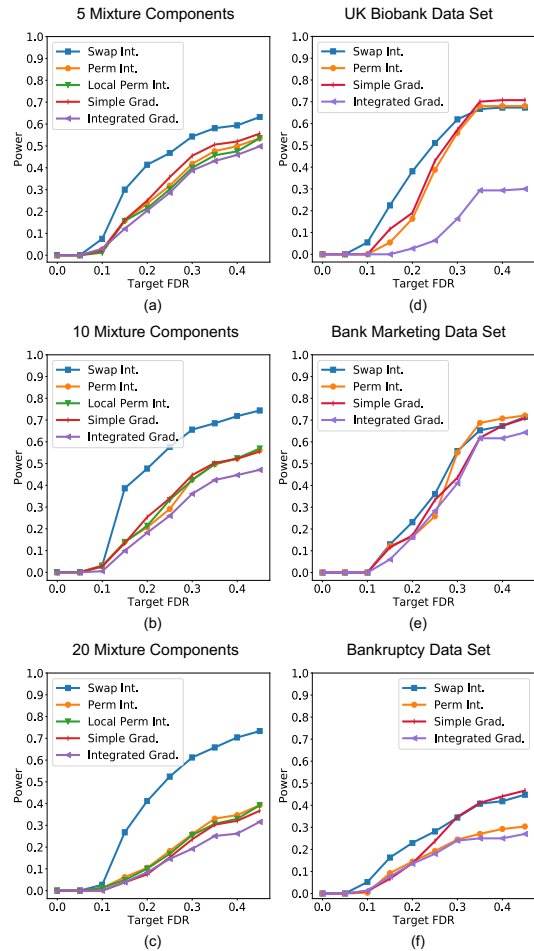
Figure 3: **Power Comparison Between Importance Score Methods** (a-c) As the number of mixture components increases, the power of Swap Integral method remains high while that of other methods decreases. (d-f) Power for Swap Integral method is as good as and sometimes better than other methods on real world data with synthetic labels.

tion in Appendix D. As our first experiment, in order to have control over the non-null features, we created synthetic labels for each dataset by randomly selecting a subset of the features to be alternatives (the rest are nulls) and use noisy polynomials on these features to generate labels. Details of the data generation are described in Appendix C. We fit each dataset with a Gaussian mixture model using the AIC method (Akaike, 1992). AIC automatically selected 9, 15, and 25 number of clusters for the UK Biobank, Bank Marketing and Bankruptcy datasets, respectively. Then we apply our Gaussian Mixture knockoff sampler to generate knockoffs for each dataset. Average results for repeating this procedure 100 times on each data set are reported in Figure 3. Across the range of target FDRs, the Swap Integral has at least as much power (and sometimes

much more) than the other feature statistic methods. It is important to notice that, as we work with a real dataset X that we cannot resample, our experiments cannot provide us with empirical FDR values in order to confirm FDR control. We expect that our model for P^X is accurate enough so that the robustness of the procedure accounts for model inaccuracies.

As a discovery experiment, we run the knockoff procedure for feature selection on each of the three data sets with the real labels –cancer in UK Biobank, telemarketing success in Marketing, and bankruptcy in Bankruptcy. In all cases we train a neural network binary classifier with 4 hidden layers of size 100. All selected features are described in Appendix D. With the real data, we do not have ground truth on which are the true alternate features. However, knockoff selected features make sense and some of the features agree with previously reported literature. With target FDR of 0.3 (it can be shown that for small target FDR values the knockoff procedure requires a minimum number of rejections, which is why we selected a fairly large target FDR), in UK Biobank cancer prediction, the knockoff procedure selected 14 out of the total 284 features. These 14 include *Number of smoked cigarettes daily* and *Wine intake*. Eight out of ten features in Bank marketing data set were selected. For Bankruptcy prediction, the knockoff procedure selected features including *gross profit* and *short-term liabilities*.

Discussion The knockoff procedure is a powerful framework for selecting important features while statistically controlling false discoveries. The two main challenges of applying knockoffs in practice are the difficulty in generating valid knockoffs and the choice of importance scores. This paper takes a step toward addressing both challenges. We develop a framework to efficiently sample valid knockoffs from Bayesian Networks. We show that even a simple application of this framework to generate knockoffs from Gaussian mixture models already enables us to apply our procedure with several real datasets, while being optimistic that FDR is controlled. We also systematically evaluate and compare the statistical power of several importance scores. We propose a new score, Swap Integral, which is stable and substantially more powerful. Swap Integral can be applied on top of any classification model, including neural networks and random forests. Note that Swap Integral is a post-processing on top of a single trained classifier and hence is computationally efficient. Our results enable knockoffs to be more practically useful and open many doors for future investigation.

Acknowledgments

References

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

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

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Adrian E Raftery. Choosing models for cross-classifications. *American sociological review*, 51(1): 145–146, 1986.

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

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

Matteo Sesia, Chiara Sabatti, and Emmanuel J Candès. Gene hunting with knockoffs for hidden markov models. *arXiv preprint arXiv:1706.04677*, 2017.

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

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

Athanassia Bacharoglou. Approximation of probability distributions by convex mixtures of gaussian measures. *Proceedings of the American Mathematical Society*, 138(7):2619–2628, 2010.

Alan B Poritz. Hidden markov models: A guided tour. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 7–13. IEEE, 1988.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29 (2-3):131–163, 1997.

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

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *arXiv preprint arXiv:1710.10547*, 2017.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 610–624. Springer, 1992.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank tele-marketing. *Decision Support Systems*, 62:22–31, 2014.

Maciej Zięba, Sebastian K Tomczak, and Jakub M Tomczak. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58:93–101, 2016.