Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Deep Knockoffs

Yaniv Romano , Matteo Sesia & Emmanuel Candès

To cite this article: Yaniv Romano , Matteo Sesia & Emmanuel Candès (2020) Deep Knockoffs, Journal of the American Statistical Association, 115:532, 1861-1872, DOI: 10.1080/01621459.2019.1660174

To link to this article: https://doi.org/10.1080/01621459.2019.1660174

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1861–1872: Theory and Methods
https://doi.org/10.1080/01621459.2019.1660174

Deep Knockoffs

Yaniv Romano, Matteo Sesia, and Emmanuel Candès

Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

This article introduces a machine for sampling approximate model-X knockoffs for arbitrary and unspecified data distributions using deep generative models. The main idea is to iteratively refine a knockoff sampling mechanism until a criterion measuring the validity of the produced knockoffs is optimized; this criterion is inspired by the popular maximum mean discrepancy in machine learning and can be thought of as measuring the distance to pairwise exchangeability between original and knockoff features. By building upon the existing model-X framework, we thus obtain a flexible and *model-free* statistical tool to perform controlled variable selection. Extensive numerical experiments and quantitative tests confirm the generality, effectiveness, and power of our deep knockoff machines. Finally, we apply this new method to a real study of mutations linked to changes in drug resistance in the human immunodeficiency virus. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2018
Accepted August 2019

KEYWORDS
False discovery rate; Generative models; Neural networks; Nonparametric methods; Variable selection

1. Introduction

1.1. Motivation

Model-X knockoffs (Candès et al. 2018) is a new statistical tool that allows the scientist to investigate the relationship between a response of interest and a large number of explanatory variables. In particular, model-X knockoffs can be used to identify a subset of important variables from a larger pool that could potentially explain a phenomenon under study while controlling the false discovery rate (Benjamini and Hochberg 1995). This methodology does not require any knowledge of how the response depends on the features, and the correctness of the inferences rests entirely on a precise description of the distribution of the explanatory variables, which are assumed to be random. This makes model-X knockoffs well-adapted to situations in which good models are available to describe the joint distribution of the features, as in genome-wide association studies (Sesia, Sabatti, and Candès 2018). To extend this approach to a broad set of applications, however, we need flexible tools to construct knockoff variables in the absence of reliable prior knowledge about the distribution of the covariates. Instead, we assume to have sufficient labeled or unlabeled samples to learn this distribution to a suitable level of approximation.

The goal of this article is simply stated: to make the knockoffs framework practically model-free and, therefore, widely applicable. This is achieved by exploiting recent progress in machine learning, which is repurposed to harness information from large unsupervised datasets and sample approximate model-X knockoffs. The outcome is a sensible set of tools for controlled variable selection that can help alleviate the irreproducibility issues afflicting many areas of science and data analysis (Ioannidis 2005; Gelman and Loken 2014; Baker 2016; Munafò et al. 2017).

1.2. A Preview of our Contribution

Given independent copies of $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ from some unknown distribution P_X, we seek to construct a random generator of valid knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ such that the joint law of $(X, \tilde{X})$ is invariant under the swapping of any X_j and $\tilde{X}_j$ for each $j \in \{1, \ldots, p\}$. Concretely, the machine takes X as input and generates $\tilde{X}$ through a mapping $f_\theta(X, V)$, where V is random noise and f_θ is a deep neural network. The parameters of the network are fitted on the data to optimize a loss function that quantifies the extent to which $\tilde{X}$ is a good knockoff copy of X. This goal is related to the classical problem of learning generative models; however, the challenge here is unusual since no sample from the target distribution $P_{\tilde{X}|X}$ is available. Fortunately, the existing methods of deep generative modeling can be suitably repurposed, as we shall see in Section 4. The lack of uniqueness of the target distribution should be resolved by making $\tilde{X}$ as different as possible from X, since a trivial copy would satisfy the required symmetry without being of any practical use. Our approach generalizes the solution in Candès et al. (2018), which relies on the simplifying assumption that X is multivariate Gaussian. In the context of deep generative models, the analogous idea consists of training a machine that optimizes the compatibility of the first two moments of $(X, \tilde{X})$ while keeping the strength of the pairwise correlations between X_j and $\tilde{X}_j$ under control. By including in the loss function an additional term that promotes the matching of higher moments, we will

CONTACT Matteo Sesia *msesia@stanford.edu* Department of Statistics, Sequoia Hall, 390 Jane Stanford Way, Stanford University, Stanford, CA 94305.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

show that one can move beyond the second-order approximation toward a model-free knockoff generator. The effectiveness of deep knockoff machines can be quantitatively measured using suitable goodness-of-fit diagnostics, as shown empirically by the results of our numerical experiments and data analysis. The algorithms described in this article have been implemented in Python and the corresponding software is available from *https://web.stanford.edu/group/candes/deep-knockoffs/*.

1.3. Related Work

The idea of using knockoffs as negative control variables originated in the context of linear regression with a fixed design matrix (Barber and Candès 2015). The generation of knockoffs beyond the settings considered in Candès et al. (2018) has also been tackled in Gimenez, Ghorbani, and Zou (2018), which extends the results of Sesia, Sabatti, and Candès (2018) to a broader class of Bayesian networks. Other recent advances include the work of Lu et al. (2018), Fan et al. (2018), and Zheng et al. (2018), while some interesting applications can be found in Xiao et al. (2017), Xie, Chen, and Shi (2018), and Gao et al. (2018). Very recently, deep generative models have independently been suggested as a procedure for sampling knockoffs in Jordon, Yoon, and van der Schaar (2019), through adversarial rather than moment matching networks. Even though the fundamental aims coincide and the solutions are related, our machine differs profoundly by design and it offers a more direct connection with existing work on the second-order knockoffs. Also, it is well known that adversarial networks are difficult to train (Arjovsky and Bottou 2017), while moment matching is a simpler task (Dziugaite, Roy, and Ghahramani 2015; Li, Swersky, and Zemel 2015). Since the approach of Jordon, Yoon, and van der Schaar (2019) requires simultaneously training four interacting neural networks, we expect that our machine should demand less tuning and be faster to learn. This may be a significant advantage since the ultimate goal is to make knockoffs accessible to researchers from different fields. A computationally lighter alternative is proposed in Liu and Zheng (2018), which relies on the variational autoencoder (Kingma and Welling 2013) to generate knockoff copies. Since our work was developed in parallel[1] to those of Jordon, Yoon, and van der Schaar (2019) and Liu and Zheng (2018), we do not include these recent proposals in our simulation studies. Instead, we compare our method to well-established alternatives.

2. Model-X Knockoffs

A random vector $\tilde{X} \in \mathbb{R}^p$ is said to be a knockoff copy of $X \in \mathbb{R}^p$ (Candès et al. 2018) if

$$(X, \tilde{X}) \overset{d}{=} (X, \tilde{X})_{\text{swap}(j)}, \quad \text{for each } j \in \{1, \dots, p\}. \tag{1}$$

Above, the symbol $\overset{d}{=}$ indicates equality in distribution and $(\cdot)_{\text{swap}(j)}$ is the operator swapping X_j with $\tilde{X}_j$; if $p = 2$ and $j = 1$, $(X_1, X_2, \tilde{X}_1, \tilde{X}_2)_{\text{swap}(j)}$ is equal to $(\tilde{X}_1, X_2, X_1, \tilde{X}_2)$. Knockoffs play a key role in the following variable selection problem.

[1]The results of this article were first discussed at the University of California, Los Angeles, during the Green Family Lectures on September 27, 2018.

Consider n observations $\{X^i, Y^i\}_{i=1}^n$, with $X^i \in \mathbb{R}^p$ drawn independently from a known P_X, and the label $Y^i \in \mathbb{R}$ from an unknown conditional distribution $P_{Y|X}$. The goal is to identify a subset of components of X that affect Y. One refers to X_j as unimportant if it is conditionally independent of the response Y once the value of the other $p - 1$ variables is known. The set of true null hypotheses $\mathcal{H}_0$ contains all variables that are unimportant. While searching for the largest subset $\hat{\mathcal{S}}$ of important variables, the false discovery rate should be controlled below a nominal level $q \in (0, 1)$, that is, $\mathbb{E}\left[(|\hat{\mathcal{S}} \cap \mathcal{H}_0|)/(|\hat{\mathcal{S}} \vee 1|)\right] \leq q$.

The approach of Candès et al. (2018) provably controls this error rate without placing any restrictions on $P_{Y|X}$, which can be arbitrary and completely unspecified. The first step consists of generating a knockoff copy $\tilde{X}$ for each available sample of X, such that both Equation (1) is satisfied and $Y \perp\!\!\!\perp \tilde{X} \mid X$. Some measures of feature importance Z_j and $\tilde{Z}_j$ are then evaluated for each X_j and $\tilde{X}_j$. For this purpose, almost any available method from statistics and machine learning can be applied to the vector of labels $\mathbf{Y}$ and the augmented data matrix $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, as long as the identity of the knockoffs is not revealed. Each pair is then combined through an antisymmetric function into the statistics W_j, for example, $W_j = Z_j - \tilde{Z}_j$, such that a large and positive value suggests evidence against the jth null hypothesis, while unimportant variables are equally likely to be positive or negative. Then, exact control of the false discovery rate below the nominal level q can be obtained by selecting $\hat{\mathcal{S}} = \{j : W_j \geq \tau_q\}$ (Barber and Candès 2015), where

$$\tau_q = \min\left\{t > 0 : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q\right\}.$$

The validity of this approach relies on our ability to generate $\tilde{X}$ satisfying Equation (1). Even though procedures to sample exact knockoffs have been derived for a few special classes of P_X (Candès et al. 2018; Sesia, Sabatti, and Candès 2018; Gimenez, Ghorbani, and Zou 2018), the general case remains challenging because Equation (1) is very stringent. For instance, obtaining independent samples from P_X or permuting the rows of $\mathbf{X}$ would only ensure that (X_1, X_2) is equal in distribution to $(\tilde{X}_1, \tilde{X}_2)$, while the analogous result would not hold between (X_1, X_2) and $(X_1, \tilde{X}_2)$. At the same time, the latter property is crucial since a null variable and its knockoff must explain on average the same fraction of the variance in Y. A practical approximate solution (Candès et al. 2018) is to relax (1) and match only the first two moments of the distributions. In this sense, $\tilde{X}$ is a *second-order knockoff* copy of X if $\mathbb{E}[X] = \mathbb{E}[\tilde{X}]$ and

$$\text{cov}\left[(X, \tilde{X})\right] = \begin{bmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{bmatrix}, \tag{2}$$

where Σ is the covariance matrix of X and s is any p-dimensional vector such that (2) is positive semidefinite. This weaker form of exchangeability is reminiscent of the notion of fixed-design knockoffs (Barber and Candès 2015), and it can be practically implemented by approximating the distribution of X as multivariate Gaussian (Candès et al. 2018). This often works well in practice, even though it is in principle insufficient to guarantee control of the false discovery rate under the general conditions of the model-X framework (Barber, Candès, and Samworth

2018). In this article, we build upon the existing work to obtain higher-order knockoffs that can achieve a better approximation of Equation (1) using modern techniques from the field of deep generative models.

3. Deep Generative Models

Given n independent p-dimensional samples $\{X^i\}_{i=1}^n$ from an unknown distribution P_X, one often seeks a generative model to synthesize new observations that could plausibly belong to the training set, while being sufficiently different to be nontrivial. Several solutions have been proposed, some of which are based on hidden Markov models (Baum and Petrie 1966), Gaussian mixture models (Nasrabadi 2007), or Boltzmann machines (Ackley, Hinton, and Sejnowski 1985). Recently, many traditional methods have been largely replaced by variational autoencoders (Kingma and Welling 2013) and generative adversarial networks (Goodfellow et al. 2014). These are based on a parametric function $f_\theta(V)$ that maps an input noise vector V to the domain of X. The parameters in θ represent a neural network and they need to be learned from the data. The function f_θ is deterministic for any fixed V and, with an appropriate choice of θ, it transforms the noise to obtain a variable approximately distributed as X.

Training deep generative models is difficult, and considerable effort has been dedicated to the development of practical algorithms that can find good solutions. Even though adversarial networks have enjoyed a great deal of success, they require solving a non-convex minimax optimization problem that is notoriously difficult (Arjovsky and Bottou 2017). This issue is mitigated in more recent alternatives such as moment-matching networks and related methods (Dziugaite, Roy, and Ghahramani 2015; Li, Swersky, and Zemel 2015; Li et al. 2017; Bińkowski et al. 2018; Srivastava et al. 2018). The rest of this section is dedicated to reviewing the basics of some of the latter approaches, upon which we will begin to develop a knockoff machine.

Given two sets of independent observations $\{X^i\}_{i=1}^n$ and $\{Z^i\}_{i=1}^n$, drawn from some unknown distributions P_X and P_Z, a generative model must verify whether $P_X = P_Z$. This problem has a long history in the statistics literature and many nonparametric tests have been proposed to address it (Bickel 1969; Friedman and Rafsky 1979; Schilling 1986; Henze 1988; Friedman 2004; Gretton et al. 2012; Székely and Rizzo 2013). The work of Gretton et al. (2012) introduced a test statistic called *maximum mean discrepancy*, whose desirable computational properties have inspired the development of moment matching networks (Li, Swersky, and Zemel 2015; Dziugaite, Roy, and Ghahramani 2015). The key idea is to quantify the discrepancy between the two distributions in terms of the largest difference in expectation between $\phi(X)$ and $\phi(Z)$, over functions ϕ mapping into the unit ball of a reproducing kernel Hilbert space (Gretton et al. 2012). This characterization can be made explicit with the kernel trick (Gretton et al. 2012), leading to the practical utilization described below.

Let X, X', Z, Z' be independent samples drawn from P_X and P_Z, and k be a kernel function. Then, we define the maximum mean discrepancy between P_X and P_Z as follows:

$$\mathcal{D}_{\text{MMD}}(P_X, P_Z) = \mathbb{E}_{X,X'}\left[k(X, X')\right] - 2\mathbb{E}_{X,Z}\left[k(X, Z)\right] + \mathbb{E}_{Z,Z'}\left[k(Z, Z')\right]. \tag{3}$$

If the characteristic kernel of a reproducing kernel Hilbert space (Gretton et al. 2012) is used, it can be shown that Equation (3) is equal to zero if and only if $P_X = P_Z$. Concretely, valid choices of k include the Gaussian kernel, $k(X, X') = \exp\{-\|X - X'\|_2^2/(2\xi^2)\}$, with any $\xi > 0$, and mixtures of such. The choice of kernel implicitly determines the feature mapping ϕ that defines the discrepancy measure; this can be computed explicitly for the Gaussian kernel (Cotter, Keshet, and Srebro 2011). In general, the maximum mean discrepancy is always nonnegative and it can be estimated from finite samples $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n\times p}$ in an unbiased fashion via

$$\widehat{\mathcal{D}}_{\text{MMD}}(\mathbf{X}, \mathbf{Z}) = \frac{1}{n(n-1)} \sum_{i=1, j\neq i}^{n} \left[k(X^i, X^j) + k(Z^i, Z^j)\right] - \frac{2}{n^2} \sum_{i=1,j=1}^{n} k(X^i, Z^j), \tag{4}$$

as discussed in Gretton et al. (2012). Since the expression in Equation (4) is easily computable and differentiable, it can serve as the objective function of a deep generative model (Li, Swersky, and Zemel 2015; Dziugaite, Roy, and Ghahramani 2015). The generator is then trained on $\mathbf{X}$ to produce samples $\mathbf{Z}$ that minimize (4), by applying the standard techniques of gradient descent. This idea can also be repurposed to develop a knockoff machine, as discussed in the next section.

4. Deep Knockoff Machines

4.1. Overview

A knockoff machine is defined as a random mapping f_θ that takes as input a random $X \in \mathbb{R}^p$, an independent noise vector $V \sim \mathcal{N}(0, I) \in \mathbb{R}^p$, and returns an approximate knockoff copy $\tilde{X} = f_\theta(X, V) \in \mathbb{R}^p$. The machine is characterized by a set of parameters θ and it should be designed such that the joint distribution of $(X, \tilde{X})$ deviates from Equation (1) as little as possible. If the original variables follow a multivariate Gaussian distribution, that is, $X \sim \mathcal{N}(0, \Sigma)$, a family of machines generating exact knockoffs is given by

$$f_\theta(X, V) = X - X\Sigma^{-1}\text{diag}\{s\} + \left(2\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\text{diag}\{s\}\right)^{1/2} V, \tag{5}$$

for any choice of the vector s that keeps the matrix multiplying V positive-definite (Candès et al. 2018). In practice, the value of s is typically determined by solving a semidefinite program (Candès et al. 2018); see Section 4.3. By contrast, the algorithm for sampling knockoff copies of hidden Markov models in Sesia, Sabatti, and Candès (2018) cannot be easily represented as an explicit function f_θ. This difficulty should be expected for various other choices of P_X, and an analytic derivation of f_θ seems intractable in general.

In order to develop a flexible machine for arbitrary and unknown distributions P_X, we assume f_θ to take the form of a deep neural network, as described in the supplementary material. The values of its parameters will be estimated from observations of X by solving a stochastic optimization problem. Our approach is sketched in Figure 1 and it can be summarized as follows. The machine is provided with n realizations of the

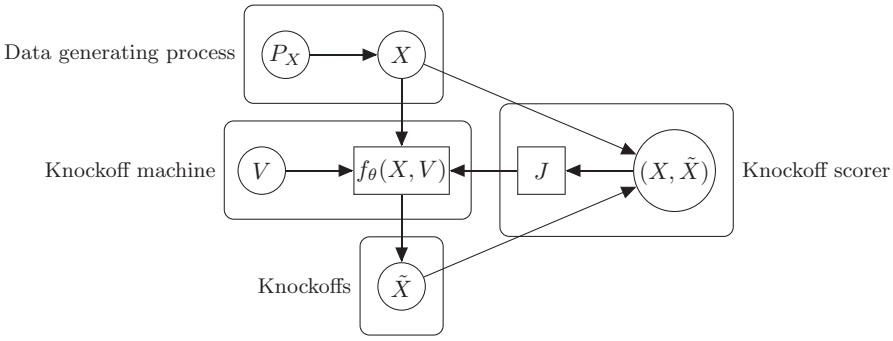

Figure 1. Schematic of the learning mechanism of a knockoff machine. The arrows indicate the flow of information between the data, the machine and the knockoff scoring function.

random vector X, independently sampled from an unknown P_X. For any fixed θ, each $\tilde{X}^i$ is computed as a function of the input X^i and the noise V^i, for $i \in \{1,\dots,n\}$. The noise V^i is independently resampled for each observation and each time the machine is called. A scoring function J examines the empirical distribution of $(X,\tilde{X})$ and quantifies its compliance with Equation (1). After each iteration, the parameters θ are updated in the attempt to improve future scores. Ideally, upon successful completion of this process, the machine should be ready to generate approximate knockoffs $\tilde{X}$ for new observations of X drawn from the same P_X. A specific scoring function that can generally lead to high-quality knockoffs will be defined below.

4.2. Second-Order Machines

We begin by describing the training of a special machine for expository purposes. Suppose that instead of requiring $(X,\tilde{X})$ to respect Equation (1), we would be satisfied with the second-order knockoffs. In order to incentivize the machine to produce $\tilde{X}$ such that the joint covariance matrix $\hat{G}$ of $(X,\tilde{X}) \in \mathbb{R}^{2p}$ obeys Equation (2), we consider a simple loss function that computes a differentiable measure of its compatibility with these requirements. After writing

$$\hat{G} = \begin{bmatrix} \hat{G}_{XX} & \hat{G}_{X\tilde{X}} \\ \hat{G}_{X\tilde{X}} & \hat{G}_{\tilde{X}\tilde{X}} \end{bmatrix}, \tag{6}$$

where $\hat{G}_{XX}, \hat{G}_{\tilde{X}\tilde{X}} \in \mathbb{R}^{p\times p}$ are the empirical covariance matrices of $X,\tilde{X}$, we define

$$J_{\text{second-order}}(\mathbf{X},\tilde{\mathbf{X}}) = \lambda_1 \frac{\|\hat{G}_{XX} - \hat{G}_{\tilde{X}\tilde{X}}\|_F^2}{\|\hat{G}_{XX}\|_F^2} + \lambda_2 \frac{\|M \circ (\hat{G}_{XX} - \hat{G}_{X\tilde{X}})\|_F^2}{\|\hat{G}_{XX}\|_F^2}. \tag{7}$$

The symbol $\circ$ indicates element-wise multiplication, while $M = E - I \in \mathbb{R}^{p\times p}$, with E being a matrix of ones and I the identity matrix. While this loss function encourages the matching of the second moments, we will also add $(\lambda_3/p)\cdot\|n^{-1}\sum_{i=1}^n (X^i - \tilde{X}^i)\|_2^2$ to Equation (7), in order to ensure that $\mathbb{E}[X] = \mathbb{E}[\tilde{X}]$. Smaller values of this loss suggest that $\tilde{X}$ is a better second-order approximate knockoff of X. Since J is smooth, the machine can be trained by stochastic gradient descent. For simplicity, $\lambda_1,\lambda_2,\lambda_3 = 1$ throughout this article.

As we mentioned earlier, knockoffs are not uniquely defined, and it is desirable to make $\tilde{X}$ as different as possible from X. A practical solution consists of adding a regularization term to the loss, targeting large pairwise empirical correlations between X and $\tilde{X}$:

$$J_{\text{decorrelation}}(\mathbf{X},\tilde{\mathbf{X}}) = \|\text{diag}(\hat{G}_{X\tilde{X}}) - 1 + s^*_{\text{SDP}}(\hat{G}_{XX})\|_2^2. \tag{8}$$

Above, $\hat{G}$ is defined as in Equation (6) and: $s^*_{\text{SDP}}(\Sigma) = \arg\min_{s\in[0,1]^p} \sum_{j=1}^p |1 - s_j|$ such that $2\Sigma \succeq \text{diag}(s) \succeq 0$. This semi-definite program is the same used in Candès et al. (2018) for the special case of $X \sim \mathcal{N}(0,\Sigma)$, to minimize the pairwise correlations between $\tilde{X}$ and X. Under the Gaussian assumption, the constraint $2\Sigma \succeq \text{diag}(s) \succeq 0$ is necessary and sufficient to ensure that the joint covariance matrix of $(X,\tilde{X})$ is positive semidefinite. Compared to the original method in Candès et al. (2018), the additional computational burden of fitting a neural network is significant. However, the tools developed in this section can be generalized beyond the second-order setting, as discussed next.

4.3. Higher-Order Machines

In order to build a general knockoff machine, one must precisely quantify and control the deviation from exchangeability: the difference in distribution between $(X,\tilde{X})$ and $(X,\tilde{X})_{\text{swap}(j)}$ for each $j \in \{1,\dots,p\}$. For this purpose, we deploy the maximum mean discrepancy from Section 3. In order to obtain an unbiased estimate, we randomly partition the data into $\mathbf{X}',\mathbf{X}'' \in \mathbb{R}^{n/2\times p}$ and define the corresponding output of the machine as $\tilde{\mathbf{X}}',\tilde{\mathbf{X}}''$. Then, it is natural to seek a machine that targets $\sum_{j=1}^p \widehat{\mathcal{D}}_{\text{MMD}}\left[(\mathbf{X}',\tilde{\mathbf{X}}'),(\mathbf{X}'',\tilde{\mathbf{X}}'')_{\text{swap}(j)}\right]$. Above, $\widehat{\mathcal{D}}_{\text{MMD}}$ stands for the empirical estimate in (4) with a Gaussian kernel. Intuitively, this is minimized in expectation if (1) is satisfied, as more precisely stated below. We refer to this solution as a higher-order knockoff machine because the expansion of the Gaussian kernel into a power series leads to a characterization of (3) in terms of the higher-moments of the two distributions (Cotter, Keshet, and Srebro 2011; Gretton et al. 2012). Our approach can thus be interpreted as a natural generalization of the method in Candès et al. (2018).

Since computing $\widehat{\mathcal{D}}_{\text{MMD}}$ for p swaps may be expensive, in practice we will only consider

$$J_{\text{MMD}}(\mathbf{X},\tilde{\mathbf{X}}) = \widehat{\mathcal{D}}_{\text{MMD}}\left[(\mathbf{X}',\tilde{\mathbf{X}}'),(\tilde{\mathbf{X}}'',\mathbf{X}'')\right] + \widehat{\mathcal{D}}_{\text{MMD}}\left[(\mathbf{X}',\tilde{\mathbf{X}}'),(\mathbf{X}'',\tilde{\mathbf{X}}'')_{\text{swap}(S)}\right], \tag{9}$$

where S indicates a uniform random subset of $\{1,\dots,p\}$ such that $j \in S$ with probability 1/2. The following result, whose proof is in the supplementary material, confirms that the objective in Equation (9) provides a sensible guideline for training knockoff machines.

Theorem 1. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a collection of independent observations drawn from P_X, and define $\tilde{\mathbf{X}}$ as the corresponding random output of a fixed machine f_θ. Then for J_{MMD} defined as in Equation (9), $\mathbb{E}\left[J_{\text{MMD}}(\mathbf{X},\tilde{\mathbf{X}})\right] \geq 0$. Moreover, equality holds if and only if the machine produces valid knockoffs for P_X. Above, the expectation is taken over $\mathbf{X}$, the noise in the knockoff machine, and the random swaps in the loss function.

With finitely many observations, stochastic gradient descent aims to minimize the expectation of Equation (9) conditional on the data. This involves a high-dimensional nonconvex optimization problem that is difficult to study theoretically. Nonetheless, effective algorithms exist and a weak form of convergence of stochastic gradient descent for our machine is established in the supplementary material. Therefore, these results provide a solid basis for our method.

The full objective function of a knockoff machine may also include the quantities from Equations (7) and (8), as a form of regularization, thus reading as follows:

$$J(\mathbf{X},\tilde{\mathbf{X}}) = \gamma J_{\text{MMD}}(\mathbf{X},\tilde{\mathbf{X}}) + \lambda J_{\text{second-order}}(\mathbf{X},\tilde{\mathbf{X}}) + \delta J_{\text{decorrelation}}(\mathbf{X},\tilde{\mathbf{X}}). \quad (10)$$

The second-order penalty may appear redundant because J_{MMD} already penalizes discrepancies in the covariance matrix, as well as in all other moments. However, setting $\lambda > 0$ explicitly leverages the second-order approximation upon which we aim to improve, and often helps in reducing the training time. With any fixed (γ,λ,δ), the machine can be fitted by stochastic gradient descent, as summarized in Algorithm 1. Further details regarding the implementation of our machines are in the supplementary material. For optimal performance, the hyperparameters should be tuned to the data distribution at hand. For this purpose, we discuss below some tools to measure goodness of fit.

Algorithm 1: Training a deep knockoff machine

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}$ – Training data.
(γ,λ,δ) – Hyperparameter of the loss function.
θ_1 – Initialization values for the weights and biases of the network.
(μ, T) – Learning rate and number of iterations.

for $t = 1 : T$ **do**

Sample the noise realizations: $V^i \sim \mathcal{N}(0, I)$, for all $1 \leq i \leq n$;

Randomly divide $\mathbf{X}$ into two disjoint mini-batches $\mathbf{X}', \mathbf{X}''$;

Pick a subset of swapping indices $S \subset \{1,\dots,p\}$ uniformly at random;

Generate the knockoffs as a deterministic function of θ: $\tilde{X}^i = f_{\theta_t}(X^i, V^i)$, for all $1 \leq i \leq n$;

Evaluate the objective function, using the fixed batches and swapping indices: $J_{\theta_t}(\mathbf{X},\tilde{\mathbf{X}}) = \gamma J_{\text{MMD}}(\mathbf{X},\tilde{\mathbf{X}}) + \lambda J_{\text{second-order}}(\mathbf{X},\tilde{\mathbf{X}}) + \delta J_{\text{decorrelation}}(\mathbf{X},\tilde{\mathbf{X}})$;

Compute the gradient of $J_{\theta_t}(\mathbf{X},\tilde{\mathbf{X}})$, which is now a deterministic function of θ;

Update the parameters: $\theta_{t+1} = \theta_t - \mu\nabla_{\theta_t} J_{\theta_t}(\mathbf{X},\tilde{\mathbf{X}})$;

end

Output: f_{θ_T} – A knockoff machine.

5. Robustness and Diagnostics

5.1. Measuring Goodness of Fit

The goodness of fit of a conditional model producing approximate knockoff copies $\tilde{X} \mid X$ can be informally described as the compatibility of the distribution of $(X,\tilde{X})$ with (1). By defining and evaluating different measures of discrepancy, the quality of our deep knockoff machines can be quantitatively compared to that of existing alternatives. For any P_X and $P_{\tilde{X}|X}$, one should verify whether $\mathcal{H}_0^{(j)} : P_{(X,\tilde{X})} = P_{(X,\tilde{X})_{\text{swap}(j)}}$, $\forall j \in \{1,\dots,p\}$. In order to reduce the number of comparisons, we will instead consider the following two hypotheses:

$$\mathcal{H}_0^{\text{full}} : P_{(X,\tilde{X})} = P_{(\tilde{X},X)}, \quad \mathcal{H}_0^{\text{partial}} : P_{(X,\tilde{X})} = P_{(X,\tilde{X})_{\text{swap}(S)}}, \quad (11)$$

where S is a random subset of $\{1,\dots,p\}$, chosen uniformly and independently of $(X,\tilde{X})$, such that $j \in S$ with probability 1/2. Either hypothesis can be separately investigated with a variety of existing two-sample tests. In order to study $\mathcal{H}_0^{\text{full}}$, we define $\mathbf{Z}_1$ and $\mathbf{Z}_2$ as two independent sets of n observations of $Z_1 = (X,\tilde{X})$ and $Z_2 = (\tilde{X},X)$. The analogous tests of $\mathcal{H}_0^{\text{partial}}$ can be performed by defining $\mathbf{Z}_2$ as samples of $(X,\tilde{X})_{\text{swap}(S)}$, and are omitted.

Covariance diagnostics. It is natural to begin with a comparison of the covariance matrices of Z_1 and Z_2, namely $G_1, G_2 \in \mathbb{R}^{2p\times 2p}$. For this purpose, we compute the following statistic meant to test the hypothesis that $G_1 = G_2$:

$$\widehat{\varphi}_{\text{COV}} = \frac{1}{n(n-1)} \sum_{i=1,j\neq i}^{n} \left[(Z_{1i}^\top Z_{1j})^2 + (Z_{2i}^\top Z_{2j})^2\right] - \frac{2}{n^2}\sum_{i=1,j=1}^{n} (Z_{1i}^\top Z_{2j})^2. \quad (12)$$

This quantity is an unbiased estimate of $\|G_1 - G_2\|_F^2 = \text{Tr}(G_1^\top G_1) + \text{Tr}(G_2^\top G_2) - 2\text{Tr}(G_1^\top G_2)$, if Z_1 and Z_2 have zero mean (Li and Chen 2012). In practice, Z_1 and Z_2 will be centered if this assumption does not hold. The asymptotic distribution of Equation (12) can be derived under mild conditions, thus yielding a nonparametric test of the null hypothesis that $G_1 = G_2$ (Li and Chen 2012). However, since our goal is to compare knockoffs generated by alternative algorithms, we will simply interpret larger values of (12) as evidence of a worse fit.

MMD diagnostics. Since $\widehat{\varphi}_{\text{COV}}$ does not capture the higher-order moments of $(X,\tilde{X})$, different diagnostics should be used

in order to have power against other alternatives. For example, the first null hypothesis in Equation (11) can be tested by computing:

$$\widehat{\varphi}_{\text{MMD}} = \widehat{\mathcal{D}}_{\text{MMD}}\left(\mathbf{Z}_1, \mathbf{Z}_2\right), \tag{13}$$

where the function $\widehat{\mathcal{D}}_{\text{MMD}}$ is defined as in Equation (4); see Gretton et al. (2012) for details. Since this is an unbiased estimate of the maximum mean discrepancy between the two distributions, large values can again be interpreted as evidence against the null. On the other hand, exact knockoffs will yield zero on average.

KNN diagnostics. The k-nearest neighbors test (Schilling 1986) can also be employed to obtain a non-parametric measure of goodness of fit. For simplicity, we consider here the special case of $k = 1$. For each sample $z_{li} \in \mathbf{Z}_l$, with $l \in \{1, 2\}$, we denote the nearest neighbor in Euclidean distance of z_{li} among $\mathbf{Z} = \mathbf{Z}_1 \cup \mathbf{Z}_2 \setminus \{z_{li}\}$ as $NN(z_{li})$. Then, we define $I_l(i)$ to be equal to one if $NN(z_{li}) \in \mathbf{Z}_l$ and zero otherwise, and compute the fraction of samples whose nearest neighbor happens to originate from the same distribution:

$$\widehat{\varphi}_{\text{KNN}} = \frac{1}{2n} \sum_{i=1}^{n} \left[I_1(i) + I_2(i)\right]. \tag{14}$$

In expectation, $\widehat{\varphi}_{\text{KNN}}$ is equal to $1/2$ if the two distributions are identical, while larger values provide evidence against the null (Schilling 1986).

Energy diagnostics. Finally, the hypotheses in Equation (11) can also be tested in terms of the energy distance (Székely and Rizzo 2013), defined as follows:

$$\begin{aligned}\mathcal{D}_{\text{Energy}}(P_{Z_1}, P_{Z_2}) = 2\mathbb{E}_{Z_1,Z_2}\|Z_1 - Z_2\|_2 - \mathbb{E}_{Z_1,Z_1'}\|Z_1 - Z_1'\|_2 \\ - \mathbb{E}_{Z_2,Z_2'}\|Z_2 - Z_2'\|_2,\end{aligned} \tag{15}$$

where $Z_1, Z_1', Z_2, and Z_2'$ are independent samples drawn from P_{Z_1} and P_{Z_2}, respectively. Assuming finite second moments, one can conclude that $\mathcal{D}_{\text{Energy}} \geq 0$, with equality if and only if Z_1 and Z_2 are identically distributed (Székely and Rizzo 2013). Therefore, we follow the approach of Székely and Rizzo (2013) and define the empirical estimator

$$\begin{aligned}\widehat{\mathcal{D}}_{\text{Energy}}\left(\mathbf{Z}_1, \mathbf{Z}_2\right) = \frac{2}{n^2} \sum_{i=1,j=1}^{n} \|Z_{1i} - Z_{2j}\|_2 \\ - \frac{1}{n^2} \sum_{i=1,j=1}^{n} \left[\|Z_{1i} - Z_{1j}\|_2 + \|Z_{2i} - Z_{2j}\|_2\right],\end{aligned}$$

and the test statistic

$$\widehat{\varphi}_{\text{Energy}} = \frac{n}{2}\left[\widehat{\mathcal{D}}_{\text{Energy}}\left(\mathbf{Z}_1, \mathbf{Z}_2\right) - \frac{2}{n^2} \sum_{i=1}^{n} \|Z_{1i} - Z_{2i}\|_2\right]. \tag{16}$$

If the two distributions are equal, this quantity converges to zero as n grows, under the assumption of finite second moments. Otherwise, if the two distributions differ, it grows to infinity. For our purpose, we interpret larger values of $\widehat{\varphi}_{\text{Energy}}$ as evidence of a poorer fit.

5.2. False Discovery Rate Under Model Misspecification

The quality of the knockoffs produced by our machines will be tested according to the measures of discrepancy defined above. However, even when $(X, \tilde{X})$ does not respect Equation (1), the false discovery rate may sometimes be controlled in practice. Since a scientist aiming to perform inference on real problems cannot blindly trust any statistical method, it is important to develop a richer set of validation tools. The strategy in Candès et al. (2018) consists of making numerical experiments that replicate the model misspecification present in the data of interest. The idea is to sample an artificial response Y from some known conditional likelihood given the real explanatory variables X. Meanwhile, approximate knockoff copies are generated using the best available algorithm. Since the true null hypotheses are known in this setting, the proportion of false discoveries can be evaluated after applying the knockoff filter. By repeating this a sufficient number of times, it is possible to verify whether the false discovery rate is contained. Such experiments help confirm whether the knockoffs can be applied because the distribution of $(X, \tilde{X})$ is the same as in the real data.

6. Numerical Experiments

6.1. Experimental Setup

The machine of Section 4 is implemented in Python, according to the technical details in the supplementary material. Here, we analyze the performance of our method in a variety of experiments for different P_X. In each case the machines are trained on $n = 10^4$ realizations of $X \in \mathbb{R}^p$, with $p = 100$. Stochastic gradient descent is applied using mini-batches of size $n/4$ and learning rate $\mu = 0.001$, for $T = 10^5$ gradients steps. A few values of $(\gamma, \lambda, \delta)$ in the proximity of $(1, 1, 1)$ are considered. The machine is typically not very sensitive to this choice, although we will discuss how different ratios work better with some distributions. Upon completion of training, the goodness of fit is quantified in terms of the metrics in Section 5.1: the matching of second moments (12), the maximum mean discrepancy (13), the k-nearest neighbors statistic (14) and the energy statistic (16). These measures are evaluated on 1000 previously unseen samples from the same P_X. The diagnostics obtained with deep machines are compared against those corresponding to other existing algorithms. We consider different choices of P_X: (i) a multivariate Gaussian; (ii) a multivariate Student's-t distribution; (iii) a "sparse Gaussian" model; (iv) a hidden Markov model; (v) a Gaussian mixture. In the interest of space, the results corresponding to (iv) and (v) are in the supplementary material. A natural benchmark in all scenarios is the second-order method from Candès et al. (2018), which we apply by relying on the empirical covariance matrix $\hat{\Sigma}$ computed on the same data used to train the deep machine. We will observe that our machines significantly improve on the second-order method when P_X deviates the most from a multivariate Gaussian, that is, in (ii) and (iii), while performing similarly in the other cases. Moreover, we also consider exact knockoff constructions with perfect oracle knowledge of P_X as ideal competitors.

Finally, variable selection is carried out in a controlled setting, on a response simulated from a known conditional

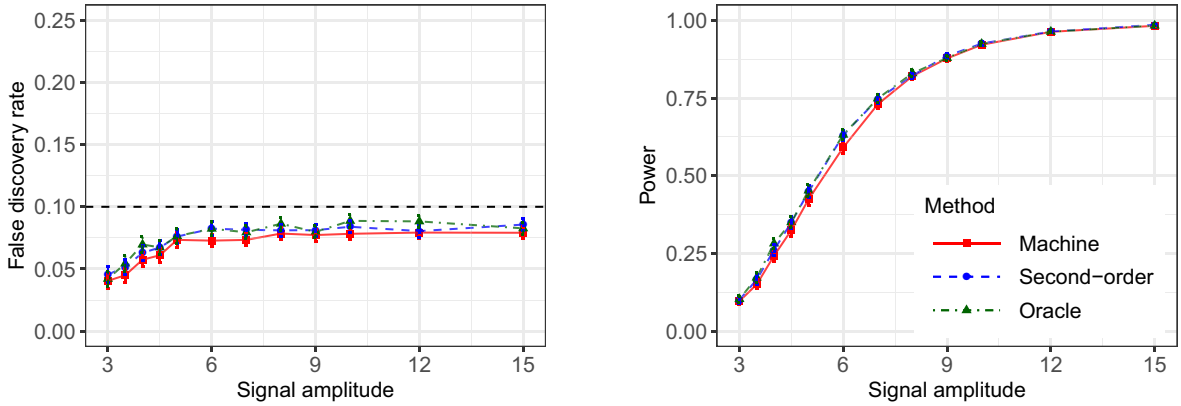

Figure 2. Experiments with multivariate Gaussian variables and simulated response. The performance of the machine is compared to that of second-order and oracle knockoffs. The results are averaged over 1000 independent experiments.

likelihood. For each $i \in \{1, \dots, m\}$, the response variable $Y^i \in \mathbb{R}$ is sampled according to $Y^i \sim \mathcal{N}(X^i\beta, 1)$, with $\beta \in \mathbb{R}^p$ containing 30 randomly chosen non-zero elements equal to $a/\sqrt{m}$. The experiments are repeated 1000 times, for different values of the signal amplitude a and the number of observations m. The importance measures are defined by fitting the elastic-net (Zou and Hastie 2005) on the augmented data matrix $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{m\times 2p}$ and $\mathbf{Y} \in \mathbb{R}^m$. More precisely, we compute $(\hat{\beta}, \tilde{\beta}) \in \mathbb{R}^{2p}$ as the solution of

$$\underset{(b,\tilde{b})}{\arg\min} \left\{ \frac{\|\mathbf{Y} - \mathbf{X}b - \tilde{\mathbf{X}}\tilde{b}\|_2^2}{m} + (1-\alpha)\frac{\tau}{2}\left(\|b\|_2^2 + \|\tilde{b}\|_2^2\right) + \alpha\tau\left(\|b\|_1 + \|\tilde{b}\|_1\right)\right\}, \tag{17}$$

with the value of τ tuned by 10-fold cross validation and some fixed $\alpha \in [0, 1]$. The knockoff filter is applied on the statistics $W_j = |\hat{\beta}_j| - |\tilde{\beta}_j|$, for all $1 \le j \le p$, at the nominal level $q = 0.1$. The power and the false discovery rate with knockoffs generated by different algorithms can be evaluated and contrasted, as a consequence of the exact knowledge of the ground truth. It is important to stress that these experiments and all the diagnostics described above only rely on new observations from P_X, generated independently of those used for training.

6.2. Multivariate Gaussian

The first example is that of a multivariate Gaussian distribution, for which the exact construction of knockoffs in Candès et al. (2018) provides the ideal benchmark. We consider P_X to be an autoregressive process of order one, such that $X \sim \mathcal{N}(0, \Sigma)$ and $\Sigma_{ij} = \rho^{|i-j|}$, with $\rho = 0.5$. A deep knockoff machine is trained with $(\gamma, \lambda, \delta) = (1, 1, 1)$. The controlled numerical experiments are carried out on synthetic datasets containing $m = 150$ samples, and setting $\alpha = 0.1$ in Equation (17). The results corresponding to the deep machine are shown in Figure 2 as a function of the signal amplitude. The performance is compared to that of the second-order method (Candès et al. 2018) and an oracle that constructs exact knockoffs by applying Equation (5) with the true covariance matrix Σ and the value of s in (5) that solves the semi-definite program from Section 4.3. The goodness of fit is further investigated in terms of the diagnostics in Figure 3 and 8(a), which show that the knockoffs generated by the oracle are perfectly exchangeable, while the deep machine and the second-order knockoffs are almost equivalent. Additional diagnostics are reported in the supplementary material.

6.3. Multivariate Student's t-distribution

We now consider a multivariate Student's t-distribution with $\nu = 3$ degrees of freedom, defined such that $X = \sqrt{(\nu-2)/\nu} \cdot Z/\sqrt{\Gamma}$, where $Z \sim \mathcal{N}(0, \Sigma)$ and Γ is independently drawn from a gamma distribution with shape and rate parameters both equal to $\nu/2$. The covariance matrix Σ is that of an autoregressive process of order one with $\rho = 0.5$. Each variable has unit variance, while moments of order ν or higher are not finite. The numerical experiments of Section 6.2 are carried out using $m = 200$ samples and setting $\alpha = 0$ in (17). The performance of the deep machine is only compared to that of the second-order method. An oracle is not considered here because it is not well known, although it can be derived. The deep machine is trained with $(\gamma, \lambda, \delta) = (1, 0.01, 0.01)$ because we expect that less weight should be given to the empirical covariance matrix, which is less reliable than those in the previous experiments. The results shown in Figure 4 indicate that the deep knockoffs control the false discovery rate while second-order knockoffs fail. The goodness-of-fit diagnostics reported in Figures 5 and 8b illustrate that the deep machine significantly outperforms the second-order knockoffs.

6.4. Sparse Gaussian variables

A second example is presented in which second-order knockoffs do not control the false discovery rate. The distribution here involves variables that are weakly correlated but highly dependent. In particular, we sample $\eta \sim \mathcal{N}(0, 1)$, while a random subset A of size L is independently chosen from $\{1, \dots, p\}$. Then, $\forall j \in \{1, \dots, p\}$, $X_j = \eta\sqrt{\binom{L}{p}/\binom{L-1}{p-1}}$ if $j \in A$ and $X_j = 0$ otherwise. Here, we choose $L = 30$. The covariance matrix Σ corresponding to this P_X is equal to $\Sigma_{ij} = 1$, if $i = j$, and $\Sigma_{ij} = (L-1)/(p-1)$ otherwise. Then, we perform the usual controlled numerical experiment on the machine trained with hyperparameters equal to $(\gamma, \lambda, \delta) = (1, 0.1, 1)$, using $m = 200$ samples and $\alpha = 0$ in (17). The hyperparameter $\lambda = 0.1$

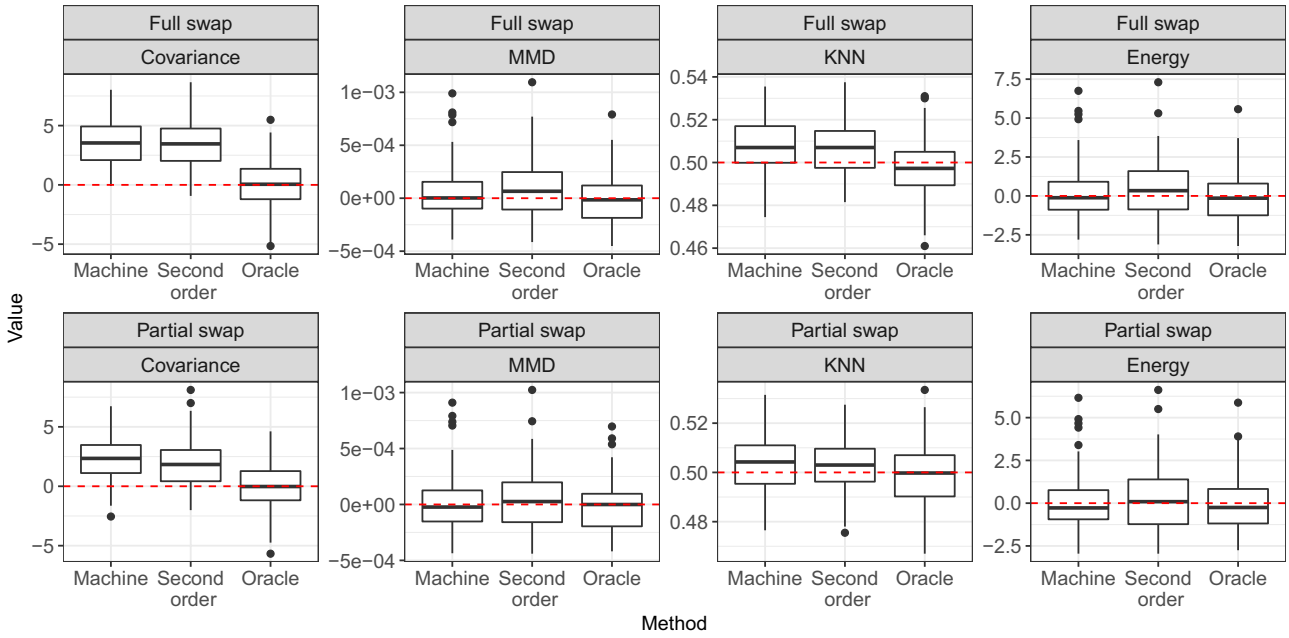

Figure 3. Boxplots comparing different goodness-of-fit diagnostics corresponding to alternative constructions of knockoffs for multivariate Gaussian variables, computed on 100 previously unseen independent datasets of size $n = 1000$. Lower values indicate better fit.

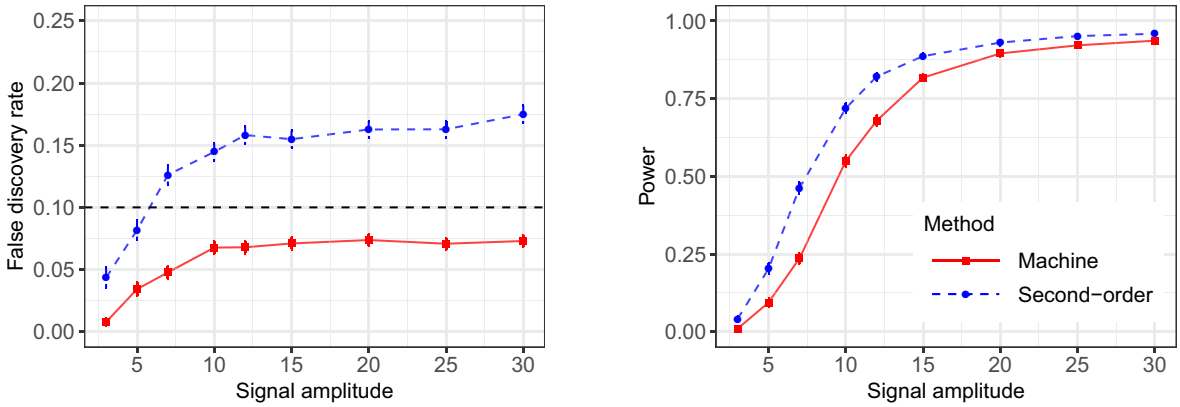

Figure 4. Experiments with a multivariate Student's t. The other details are as in Figure 2.

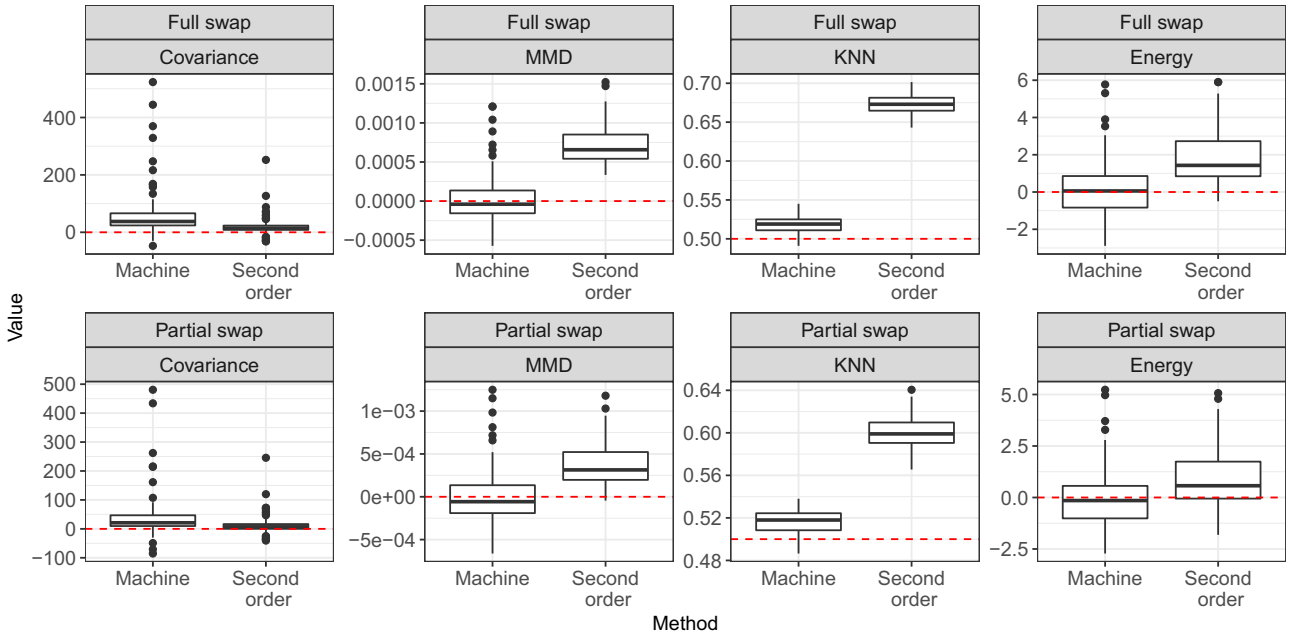

Figure 5. Boxplot comparing different knockoff diagnostics for variables sampled from a multivariate Student's t-distribution. The other details are as in Figure 3.

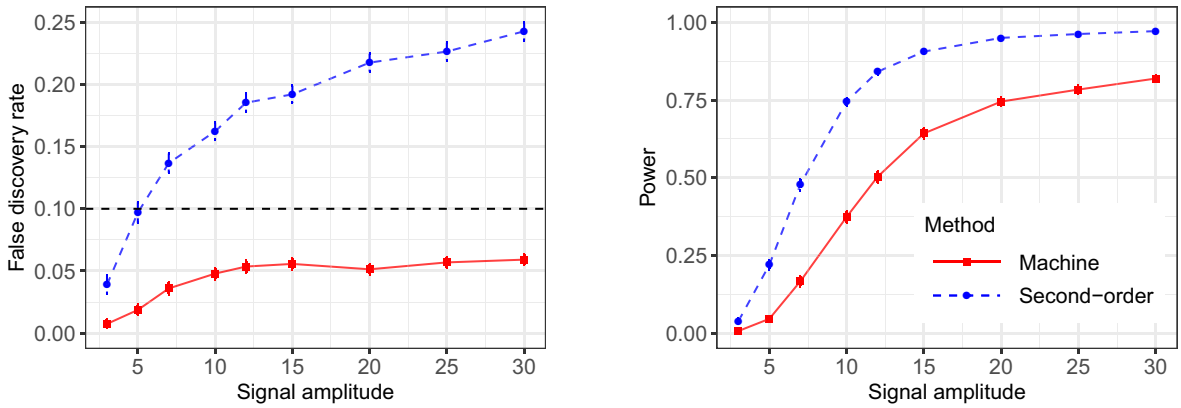

Figure 6. Experiments with sparse Gaussian variables. The other details are as in Figure 2.

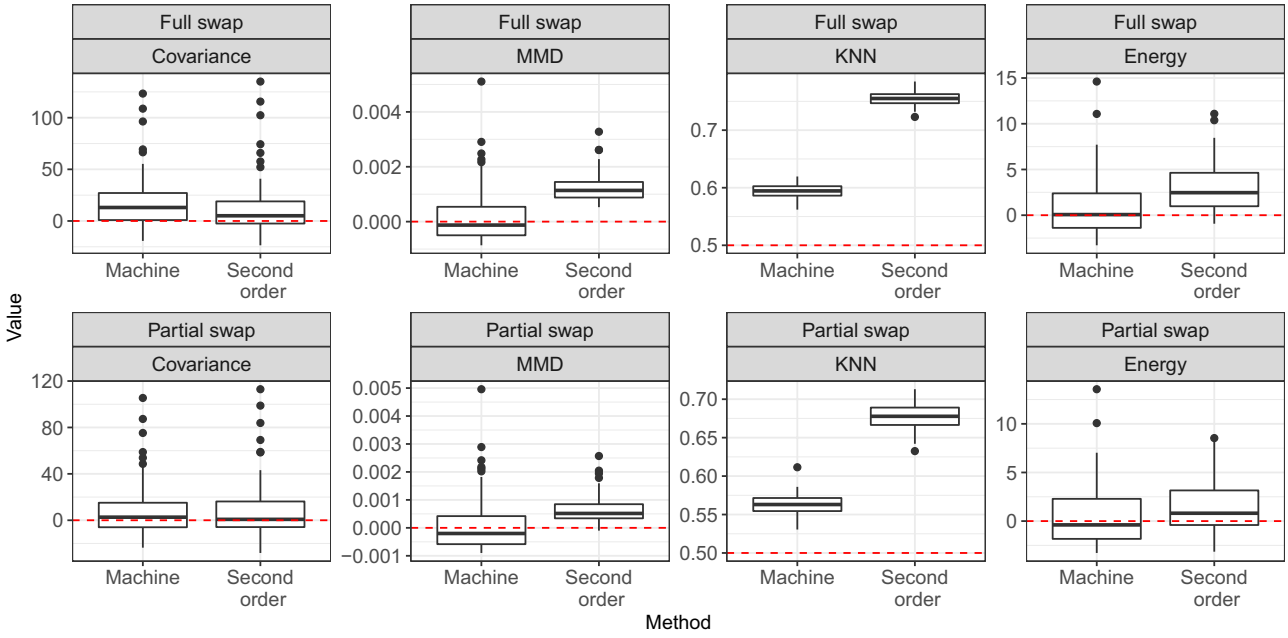

Figure 7. Boxplot comparing different knockoff diagnostics for variables sampled from a sparse multivariate Gaussian distribution. The other details are as in Figure 3.

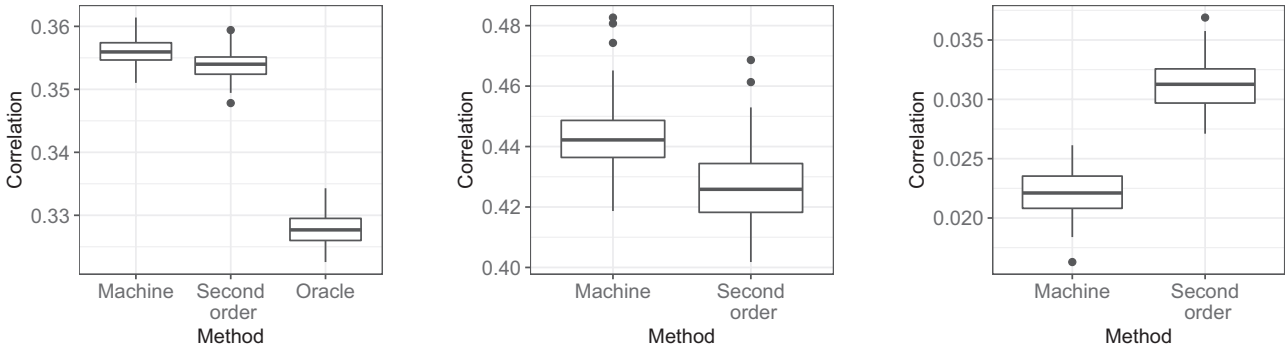

Figure 8. Boxplot comparing the average absolute pairwise correlation between variables and knockoffs for different P_X. Lower values tend to indicate more powerful knockoffs.

decreases the weight given to the empirical covariance matrix, as in the previous experiment. The performance of this machine is only compared to that of the second-order approximation, as shown in Figures 6–8(c). Even though our machine is not exact, its approximation is more accurate than that of the second-order method. This improvement is confirmed in Figure 6, illustrating that the deep machine leads to successful control of the false discovery rate, unlike the second-order knockoffs.

7. Application

7.1. Overview of the Data

We deploy our deep machine to a study of variations in drug resistance among human immunodeficiency viruses of type I in order to detect important mutations (Rhee et al. 2006). We choose this application for its importance and because the data are freely available from *https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/*. Moreover, an earlier release with fewer samples also appears in the first paper on knockoffs (Barber and Candès 2015). It should be acknowledged that it is not immediately clear whether the underlying assumptions of the model-X settings are really satisfied. In particular, we do not know how realistically the samples can be described as independent and identically distributed pairs (X, Y) drawn from some joint underlying distribution. Rigorously validating these assumptions would require expert domain knowledge and additional data. Therefore, we interpret this analysis as an illustration of how deep knockoff machines can be used in practice, without advancing any claim of new scientific findings. In any case, it is encouraging that many of the mutations discovered by our method are already known to be important, as discussed in Section 7.3 and in the supplementary material.

For simplicity, we focus on analyzing the resistance to one protease inhibitor drug, namely lopinavir. The response variable Y^i represents the log-fold increase in resistance measured in the ith virus. Having remote-file-name-inhibit-cacheved all samples containing missing values, we are left with $n = 1431$. Each of the $p = 150$ binary features X_j indicates the presence of a particular mutation. Half are chosen because they are previously known to be associated with changes in the drug resistance. The other half are chosen because they are the most frequently occurring mutations. If multiple mutations occur at the same position, the first two are treated as distinct while the others are ignored. The variables are standardized to have zero mean and unit variance, even though they have binary support. The machine used in Section 6.1 is slightly modified to produce binary output through a sigmoid activation. The hyperparameters in the loss function are $(\gamma, \lambda, \delta) = (1, 1, 1)$. The machine is trained after $T = 5 \times 10^4$ gradients steps and a learning rate $\mu = 0.01$.

The strategy adopted for the analysis of these data is different from that described in the simulations of Section 6. A deep knockoff machine is trained on the 150 mutation features corresponding to all 1431 subjects. Since the data is limited, we fit the machine on the same samples for which we need to generate the knockoff copies to perform variable selection. Therefore, it is possible that some overfitting will occur. In other words, even though the machine thus obtained may not be very accurate on new observations of X, the knockoffs produced on the training set will be nearly indistinguishable upon a finite-sample swap with the original variables. Overfitting knockoffs has been empirically observed to lead to a loss of power at worst, while the control of the Type I errors typically remains intact (Candès et al. 2018; Lu et al. 2018; Sesia, Sabatti, and Candès 2018). This claim is confirmed by the results of the numerical experiments presented below, although future research should investigate a theoretical explanation of this phenomenon. For now, we accept this limitation and proceed by verifying that the machine works for our purposes.

7.2. Numerical Experiments With Real Variables

The numerical experiments presented here are similar to those in Section 6, with the important difference that $\mathbf{X} \in \mathbb{R}^{1431\times 150}$ is held constant while we simulate a response variable for each observation. In theory, model-X knockoffs may not control the false discovery rate conditional on $\mathbf{X}$. However, it can be informative to apply and compare in this context the procedures described above. Since n is much greater than p and $\mathbf{X}$ is fixed, fixed-X knockoffs (Barber and Candès 2015) are a reasonable alternative to the deep machine and the second-order method. The results corresponding to the three competing approaches averaged over 1000 replications are shown in Figure 9 as a function of the signal amplitude. It is reassuring to observe that the second-order and the fixed-X knockoffs appear to control the false discovery rate and achieve similar power, while the deep machine outperforms both. Additional numerical experiments with these data can be found in the supplementary material.

7.3. Results

Finally, the knockoffs generated by the machine trained in Section 7.2 are used to select important features that contribute to explaining changes in the drug resistance of the viruses. The knockoff filter is applied using the same importance statistics

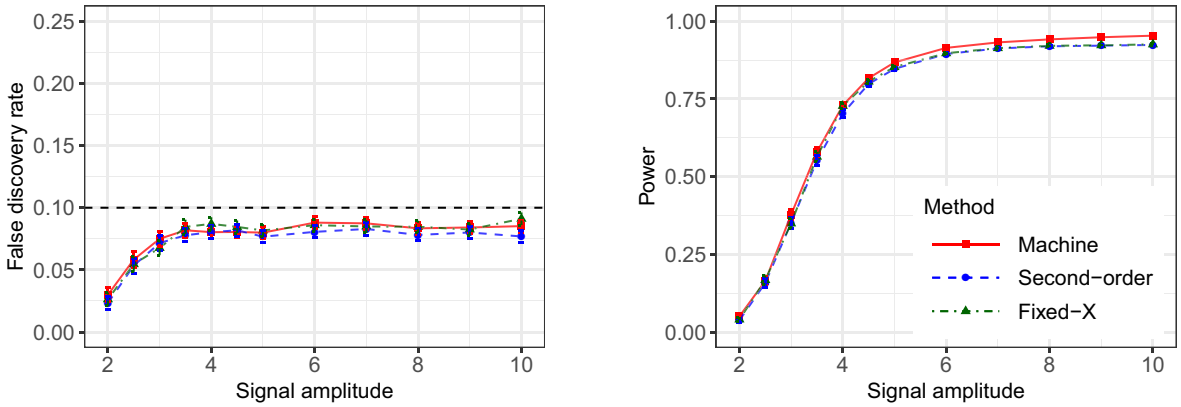

Figure 9. Numerical experiment with real human immunodeficiency virus mutation features and simulated response. The performance of the deep machine is compared to that of the second-order and fixed-X knockoff. The false discovery rate (left) and the power (right) are averaged over 1000 replications. Each replication is performed on the original $\mathbf{X}$.

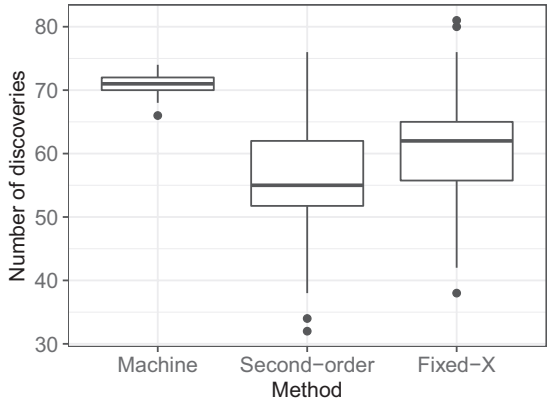

Figure 10. Boxplot of the number of drug-resistance mutations in the human immunodeficiency virus discovered using different knockoff generation methods. The variability in the results corresponds to 100 independent samples of the knockoff copies.

as above, setting $\alpha = 0.1$ in Equation (17). The nominal false discovery rate is $q = 0.1$. In order to investigate the stability of the findings obtained with this machine, the variable selection procedure is repeated 100 times, starting from a new independent realization of the knockoffs conditional on the data. The distribution of the number of discoveries on this dataset is displayed in Figure 10, along with the analogous quantity corresponding to the second-order knockoffs (Candès et al. 2018) and the randomized version of the fixed-X knockoffs (Barber and Candès 2015). The results indicate that the deep machine leads to more discoveries than the alternative approaches. This is in line with the numerical experiments presented above. It is interesting that the selections made with our machine are quite stable upon resampling of $\tilde{X} \mid X$, unlike those of other methods. This potentially significant advantage of deep knockoff machines should be investigated more rigorously in future work. The list of discovered mutations is in large part consistent with the prior knowledge on their importance, as shown in the supplementary material. According to the database on *https://hivdb.stanford.edu/dr-summary/comments/PI/*, many of our findings have been previously reported to have a major or accessory effect on changes in drug resistance.

8. Discussion

8.1. Summary

The deep machines presented in this paper extend the knockoff method to a vast range of problems. The idea of sampling knockoff copies by matching higher moments is a natural generalization of the existing second-order approximation; however, the inherent difficulties of this approach have prompted us to exploit the powerful new methods of deep learning. The numerical experiments and the data analysis described in this paper can be reproduced on a single graphics processing unit within a few hours. We believe that the computational cost will decrease as more experience is acquired, and applications on a larger scale should be pursued. The extensive numerical experiments show that our solution can match the performance of the available exact knockoff constructions for several data distributions, and greatly outperform the previous approximations in more complex cases. The diagnostics computed on independent test data confirm that the deep machines are correctly learning to generate valid knockoffs, without relying on prior knowledge. The encouraging outcomes of the data analysis motivate further applications.

There is a subtle but meaningful difference between the perspective taken by the existing theory of model-X knockoffs and the common practice on real data. In principle, finite-sample control of the false discovery rate is guaranteed when the knockoff copies are constructed with respect to the true P_X. However, knockoffs are often constructed using an estimated $\hat{P}_X$ obtained from the same samples used for variable selection, as discussed in Section 7. The interesting empirical observation is that when $\hat{P}_X$ overfits the training samples, knockoffs typically become more conservative rather than too liberal. To the best of our knowledge, this phenomenon still lacks a rigorous explanation. In any case, the numerical simulations of Section 6 show that our machines can learn how to generate valid knockoffs. In conclusion, we believe that this work is a valuable contribution because it allows the rich framework of knockoffs to be applied in very general settings. In fact, given sufficient data and adequate computing resources, deep knockoff machines can be trained on virtually any kind of features.

8.2. Future Work

There are several paths open for future research. For example, variations of our machines could be based on different scoring functions or regularization penalties. The machines described in this paper take an agnostic view of the data distribution, but there are many applications in which some prior knowledge of the structure of the variables is available. Exploiting this could improve the computational and statistical efficiency of our method, especially in high-dimensions where the maximum mean discrepancy may not be very powerful (Ramdas et al. 2015). For example, one may use prior knowledge to reduce the dimensions of the data prior to computing the maximum mean discrepancy (Li, Swersky, and Zemel 2015) or learn a suitable kernel from the data (Sutherland et al. 2016; Li et al. 2017).

A different project could involve the extension of our diagnostics using a wider selection of two-sample tests (Heller, Heller, and Gorfine 2012; Heller and Heller 2016), and a systematic study of their relative strengths. An extension of the theoretical results in Barber, Candès, and Samworth (2018) may also be valuable. Since alternative knockoff constructions based on different deep learning techniques have been independently proposed in parallel to the writing of this paper (Jordon, Yoon, and van der Schaar 2019; Liu and Zheng 2018), it is also up to future research to extensively compare their empirical performance.

Funding

Emmanuel J. Candès was partially supported by the Office of Naval Research (ONR) under grant N00014-16-1-2712, by the Army Research Office (ARO) under grant W911NF-17-1-0304, by the Math + X award from the Simons Foundation and by a generous gift from TwoSigma. Yaniv Romano was supported by the same Math + X award and ARO grant. Y. R. also thanks the Zuckerman Institute, ISEF Foundation and the

Viterbi Fellowship, Technion, for supporting this research. Matteo Sesia was supported by the same Math + X award and ONR grant. The authors thank Stephen Bates, Nikolaos Ignatiadis and Eugene Katsevich for their insightful comments on an earlier draft of this article.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985), "A Learning Algorithm for Boltzmann Machines," *Cognitive Science*, 9, 147–169. [1863]

Arjovsky, M., and Bottou, L. (2017), "Towards Principled Methods for Training Generative Adversarial Networks," arXiv no. 1701.04862. [1862,1863]

Baker, M. (2016), "1,500 Scientists Lift the Lid on Reproducibility," *Nature News*, 533, 452–454. [1861]

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [1862,1870,1871]

Barber, R. F., Candès, E. J., and Samworth, R. J. (2018), "Robust Inference With Knockoffs," arXiv no. 1801.03896. [1863,1871]

Baum, L. E., and Petrie, T. (1966), "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563. [1863]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 289–300. [1861]

Bickel, P. J. (1969), "A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case," *The Annals of Mathematical Statistics*, 40, 1–23. [1863]

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018), "Demystifying MMD GANs," arXiv no. 1801.01401. [1863]

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [1861, 1862,1863,1864,1866,1867,1870,1871]

Cotter, A., Keshet, J., and Srebro, N. (2011), "Explicit Approximations of the Gaussian Kernel," arXiv no. 1109.4603. [1863,1864]

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015), "Training Generative Neural Networks via Maximum Mean Discrepancy Optimization," arXiv no. 1505.03906. [1862,1863]

Fan, Y., Lv, J., Sharifvaghefi, M., and Uematsu, Y. (2018), "IPAD: Stable Interpretable Forecasting With Knockoffs Inference," arXiv no. 1809.05032. [1862]

Friedman, J. H. (2004), "On Multivariate Goodness-of-Fit and Two-Sample Testing," Technical Report, Stanford Linear Accelerator Center, Menlo Park, CA (US). [1863]

Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics*, 7, 697–717. [1863]

Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L. T. M., Herman, T., Giladi, N., Kalinin, A., Spino, C., Dauer, W., Hausdorff, J. M., and Dinov, I. D. (2018), "Model-Based and Model-Free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease," *Scientific Reports*, 8, 7129. [1862]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [1861]

Gimenez, J. R., Ghorbani, A., and Zou, J. (2018), "Knockoffs for the Mass: New Feature Importance Statistics With False Discovery Guarantees," arXiv no. 1807.06214. [1862]

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014), "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680. [1863]

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, 13, 723–773. [1863,1864,1866]

Heller, R., and Heller, Y. (2016), "Multivariate Tests of Association Based on Univariate Tests," in *Advances in Neural Information Processing Systems*, pp. 208–216. [1871]

Heller, R., Heller, Y., and Gorfine, M. (2012), "A Consistent Multivariate Test of Association Based on Ranks of Distances," *Biometrika*, 100, 503–510. [1871]

Henze, N. (1988), "A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences," *The Annals of Statistics*, 772–783. [1863]

Ioannidis, J. P. (2005), "Why Most Published Research Findings are False," *PLoS Medicine*, 2, e124. [1861]

Jordon, J., Yoon, J., and van der Schaar, M. (2019), "KnockoffGAN: Generating Knockoffs for Feature Selection Using Generative Adversarial Networks," in *International Conference on Learning Representations*. [1862, 1871]

Kingma, D. P., and Welling, M. (2013), "Auto-Encoding Variational Bayes," arXiv no. 1312.6114. [1862,1863]

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017), "MMD GAN: Towards Deeper Understanding of Moment Matching Network," in *Advances in Neural Information Processing Systems*, pp. 2203–2213. [1863,1871]

Li, J., and Chen, S. X. (2012), "Two Sample Tests for High-Dimensional Covariance Matrices," *The Annals of Statistics*, 40, 908–940. [1865]

Li, Y., Swersky, K., and Zemel, R. (2015), "Generative Moment Matching Networks," Lille, France: International Conference on Machine Learning, pp. 1718–1727. [1862,1863,1871]

Liu, Y., and Zheng, C. (2018), "Auto-Encoding Knockoff Generator for FDR Controlled Variable Selection," arXiv no. 1809.10765. [1862,1871]

Lu, Y. Y., Lv, J., Fan, Y., and Noble, W. S. (2018), "DeepPINK: Reproducible Feature Selection in Deep Neural Networks," arXiv no. 1809.01185. [1862,1870]

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017), "A Manifesto for Reproducible Science," *Nature Human Behaviour*, 1, 0021. [1861]

Nasrabadi, N. M. (2007), "Pattern Recognition and Machine Learning," *Journal of Electronic Imaging*, 16, 049901. [1863]

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015), "On the Decreasing Power of Kernel and Distance Based Nonparametric Hypothesis Tests in High Dimensions," Austin, TX: Twenty-Ninth AAAI Conference on Artificial Intelligence. [1871]

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006), "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance," *Proceedings of the National Academy of Sciences*, 103, 17355–17360. [1870]

Schilling, M. F. (1986), "Multivariate Two-Sample Tests Based on Nearest Neighbors," *Journal of the American Statistical Association*, 81, 799–806. [1863,1866]

Sesia, M., Sabatti, C., and Candès, E. (2018), "Gene Hunting With Hidden Markov Model Knockoffs," *Biometrika*, 106, 1–18. [1861,1862,1863, 1870]

Srivastava, A., Xu, K., Gutmann, M. U., and Sutton, C. (2018), "Ratio Matching MMD Nets: Low Dimensional Projections for Effective Deep Generative Models," arXiv no. 1806.00101. [1863]

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016), "Generative Models and Model Criticism Via Optimized Maximum Mean Discrepancy," arXiv no. 1611.04488. [1871]

Székely, G. J., and Rizzo, M. L. (2013), "Energy Statistics: A Class of Statistics Based on Distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272. [1863,1866]

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., and Liu, Y.-Y. (2017), "Mapping the Ecological Networks of Microbial Communities From Steady-State Data," bioRxiv no. 150649. [1862]

Xie, Y., Chen, N., and Shi, X. (2018), "False Discovery Rate Controlled Heterogeneous Treatment Effect Detection for Online Controlled Experiments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, pp. 876–885. [1862]

Zheng, Z., Zhou, J., Guo, X., and Li, D. (2018), "Recovering the Graphical Structures via Knockoffs," *Procedia Computer Science*, 129, 201–207. [1862]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1867]