The Annals of Statistics
2020, Vol. 48, No. 5, 3021–3042
https://doi.org/10.1214/19-AOS1920

RELAXING THE ASSUMPTIONS OF KNOCKOFFS BY CONDITIONING

BY DONGMING HUANG* AND LUCAS JANSON**

*Department of Statistics, Harvard University, *dhuang01@g.harvard.edu; **ljanson@fas.harvard.edu*

The recent paper Candès et al. (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** (2018) 551–577) introduced model-X knockoffs, a method for variable selection that provably and nonasymptotically controls the false discovery rate with no restrictions or assumptions on the dimensionality of the data or the conditional distribution of the response given the covariates. The one requirement for the procedure is that the covariate samples are drawn independently and identically from a precisely-known (but arbitrary) distribution. The present paper shows that the exact same guarantees can be made *without* knowing the covariate distribution fully, but instead knowing it only up to a parametric model with as many as $\Omega(n^* p)$ parameters, where p is the dimension and n^* is the number of covariate samples (which may exceed the usual sample size n of labeled samples when unlabeled samples are also available). The key is to treat the covariates as if they are drawn conditionally on their observed value for a sufficient statistic of the model. Although this idea is simple, even in Gaussian models conditioning on a sufficient statistic leads to a distribution supported on a set of zero Lebesgue measure, requiring techniques from topological measure theory to establish valid algorithms. We demonstrate how to do this for three models of interest, with simulations showing the new approach remains powerful under the weaker assumptions.

1. Introduction.

1.1. *Problem statement.* In this paper, we consider random variables $(Y, X_1, \ldots, X_p)$ where Y is a response or outcome variable, each X_j is a potential explanatory variable (also known as a covariate or feature) and p is the dimensionality, or number of covariates. For instance, Y could be the binary indicator of whether a patient has a disease or not, and X_j could be the number of minor alleles at a specific location (indexed by j) on the genome, also known as a single nucleotide polymorphism (SNP). A common question of interest is which of the X_j are important for determining Y, with importance defined in terms of conditional independence. That is, X_j is considered *un*important (or *null*) if

$$Y \perp\!\!\!\perp X_j \mid X_{-j},$$

where $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$; stated another way, X_j is unimportant exactly when Y's conditional distribution does not depend on X_j. Denote by $\mathcal{H}_0$ the set of all j such that X_j is unimportant. As discussed in Candès et al. [10], under very mild conditions the complement of the set of unimportant variables, that is, the *important* (or *nonnull*) variables, constitutes the Markov blanket S of Y, namely, the unique smallest set S such that $Y \perp\!\!\!\perp X_S \mid X_{-S}$. Note that when $Y \mid X_1, \ldots, X_p$ follows a generalized linear model (GLM) with no redundant covariates, the set of important variables exactly equals the set of variables with nonzero coefficients, as usual [10].

In our search for the Markov blanket, we usually cannot possibly hope for perfect recovery, so we instead attempt to maximize the number of important variables discovered while

Received March 2019; revised October 2019.
MSC2020 subject classifications. Primary 62G10; secondary 62B05, 62J02.
Key words and phrases. High-dimensional inference, knockoffs, model-X, sufficient statistic, false discovery rate (FDR), topological measure, graphical model.

probabilistically controlling the number of false discoveries. In this paper, as with most others in the knockoffs literature, we consider the false discovery rate (FDR) [6], although other error rates can also be controlled [21]. The FDR is defined for a (random) selected subset of variables $\hat{S}$ as

$$\mathrm{FDR} := \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|}\right],$$

that is, the expected fraction of discoveries that are not in the Markov blanket (false discoveries), where we use the convention that $0/0 = 0$. Controlling the FDR at, say, 10% is powerful as compared to controlling more classical error rates like the familywise error rate, while still being interpretable, allowing a statistician to report a conclusion such as "here is a set of covariates $\hat{S}$, 90% of which I expect to be important."

1.2. *Our contribution.* In our discussion of approaches to this problem, we will draw on a fundamental decomposition of the joint distribution $F_{Y,X}$ of $(Y, X_1, \ldots, X_p)$ into the product of the conditional distribution $F_{Y|X}$ of $Y \mid X_1, \ldots, X_p$ and the joint distribution F_X of $X_1, \ldots, X_p$. The canonical approach to inference, which we refer to as the "fixed-X" approach, assumes $F_{Y|X}$ is a member of a parametric family of conditional distributions (e.g., a GLM), while placing weak or no assumptions on F_X. In fact, the fixed-X approach usually treats the observed values of $X_{i,1}, \ldots, X_{i,p}$ for $i = 1, \ldots, n$ as fixed; that is, it performs inference *conditionally* on the observed values of $X_1, \ldots, X_p$ in the data, which also allows the covariate rows to be drawn from different distributions or even be deterministic (fixed). The approach proposed in Candès et al. [10], referred to therein as the "model-X" approach, assumes the observations $(Y_i, X_{i,1}, \ldots, X_{i,p}) \overset{\text{i.i.d.}}{\sim} F_{Y,X}$ and places no restrictions on F_X but assumes it is known exactly, while assuming nothing about $F_{Y|X}$. So, to summarize slightly imprecisely, the canonical, fixed-X approach to inference places all assumptions on $F_{Y|X}$ and none on F_X, while the model-X approach does the opposite by placing all assumptions on F_X and none on $F_{Y|X}$.

Note that both $F_{Y|X}$ and F_X are exponentially complex in p: in the simple case where each element of $(Y, X_1, \ldots, X_p)$ is categorical with k categories, that is, $(Y, X_1, \ldots, X_p) \in \{1, \ldots, k\}^{p+1}$, it is easily seen that a fully general model for $F_{Y|X}$ has $(k-1)k^p$ free parameters while F_X has only slightly fewer with $k^p - 1$. So both fixed-X and model-X approaches astronomically reduce an exponentially large (in p) space of distributions in order to make inference feasible, highlighting the importance of robustness, assumption-checking and domain knowledge for justifying the resulting inference; see Janson [20], Chapter 1, for a detailed discussion of the role of fixed-X and model-X assumptions in high-dimensional inference. With that said, one apparent advantage of the fixed-X approach is that it does not require *exact* knowledge of $F_{Y|X}$, while the model-X approach of [10] does require F_X be known exactly.

The present paper removes this apparent advantage by showing that model-X knockoffs can still provide powerful and exact, finite-sample inference even when the covariate distribution is only known up to a parameterized family of distributions (also known as a model), as opposed to known exactly. In fact, in Section 3 we will show three examples in which the number of parameters we allow for F_X's model is $\Omega(n^* p)$, where n^* is the total number of samples of X (including unlabeled samples), which is always at least as large as the number of labeled samples n, and can be much larger in some applications. This is much greater than the number of parameters allowed in the model for $F_{Y|X}$ in fixed-X inference (see Section 1.3). Table 1 provides a summarized comparison of the model flexibility allowed in the fixed-X and model-X approaches.

TABLE 1

Maximum complexity of models allowed by existing methods (see Section 1.3) and our proposal (see the list in Section 2.2 and also Section 2.3 for the explanation for $\Omega(n^* p)$) for controlled variable selection. Note that without assuming a model, $F_{Y|X}$ and F_X are of exponentially complex in p. Note also that fixed-X inferential guarantees are generally asymptotic in nature. The exception to this and the $o(n)$ scaling stated in the table is Gaussian linear regression, which allows $n = O(n)$ parameters and is nonasymptotic

	Model for $F_{Y\|X}$	Model for F_X
Fixed-X	$o(n)$ parameters	arbitrary
Model-X [10]	arbitrary	0 parameters
Model-X (this paper)	arbitrary	$\Omega(n^* p)$ parameters

Of course the above discussion and table refer only to the *mathematical* complexity of models allowed by the fixed-X and model-X approaches. An analyst's decision between them should depend on how well domain knowledge and/or auxiliary data support their (very different) assumptions. But in light of Table 1, it seems the conditional model-X approach is easiest to justify unless substantially more is known about $F_{Y|X}$ than F_X.

1.3. *Related work.* By far the most common fixed-X approaches to inference rely on GLMs with p parameters, reducing model complexity from exponential to linear in p. When p is smaller than the number of observations n, inference for GLMs other than Gaussian linear models relies on large-sample approximation by assuming at least $p/n \to 0$ [19, 30]. Note that the commonly studied problem of inference for a single parameter can generally be translated to FDR control using the Benjamini–Hochberg [6] or Benjamini–Yekutieli [7] procedures (see, e.g., Javanmard and Javadi [22]), so that it makes sense to compare such inference with our paper that is focused on multiple testing. In high dimensions, that is, when $p > n$, even reducing the complexity of $F_{Y|X}$ to p parameters with a GLM is insufficient for fixed-X inference, as GLMs become unidentifiable in this regime due to the design matrix columns being linearly dependent. Early solutions for fixed-X inference in high-dimensional GLMs relied on β-min conditions that lower-bound the magnitude of nonzero coefficients to obtain asymptotically-valid p-values for individual variables (see, e.g., Chatterjee and Lahiri [11]). More recent work removes the β-min condition in favor of strong sparsity assumptions on the coefficient vector, usually $o(\sqrt{n}/\log(p))$ nonzeros, with notable examples including the debiased Lasso (see, e.g., Zhang and Zhang [35], Javanmard and Montanari [23], van de Geer et al. [34]) and the extended score statistic (see, e.g., Belloni et al. [4, 5], Chernozhukov et al. [12], Ning and Liu [29]), both of which provide asymptotically-valid p-values for GLMs with some additional assumptions on the "compatibility" of the design matrix. In recent work that seems to straddle the fixed-X and model-X paradigms, Bradic [36] and Zhu et al. [37] compute asymptotically-valid p-values for the Gaussian linear model without any extra restrictions like sparsity or β-min on $F_{Y|X}$, but with added assumptions on F_X about the sparsity of conditional linear dependence among covariates.

Another branch of recent research called post-selection inference can be viewed as a different approach to high-dimensional inference: it aims to test random hypotheses selected by a high-dimensional regression and provide valid p-values by conditioning on the selection event (see, e.g., Fithian et al. [15], Lee et al. [25] for foundational contributions and Candès et al. [10], Appendix A, for more about the difference between post-selection inference and our approach).

The method of knockoffs was first introduced by Barber and Candès [1] for low-dimensional homoscedastic linear regression with fixed design. The model-X knockoffs framework proposed by Candès et al. [10] read this idea from a different perspective, providing valid finite-sample inference with no assumptions on $F_{Y|X}$ but assuming full knowledge of F_X. Exact knockoff generation methods have been found for F_X following a multivariate Gaussian [10], a Markov chain or hidden Markov model [33], a graphical model [3] and certain latent variable models [17]. In the case that F_X is only known approximately, the robustness of model-X knockoffs is studied by Barber et al. [2]. When F_X is completely unknown some recent works have proposed methods to generate approximate knockoffs [24, 27, 31] which have shown promising empirical results, particularly in low-dimensional problems, but come with no theoretical guarantees. In contrast, the current paper proposes to construct valid knockoffs that provide exact finite sample error control.

This paper is based on the idea of performing inference conditional on a sufficient statistic for F_X's model so as to make that inference parameter-free. In low-dimensional inference, likely the simplest example of such an idea is a permutation test for independence, which can be thought of as a randomization test performed conditional on the order statistics of an observed i.i.d. vector of scalar X with unknown distribution (the order statistics are sufficient for the family of all one-dimensional distributions). Although permutation tests can only test marginal independence, not conditional independence as addressed in the present paper, Rosenbaum [32] constructs a conditional permutation test that does test conditional independence assuming a logistic regression model for $X_j \mid X_{\text{-}j}$, and allows the parameters of the logistic regression model to be unknown by conditioning on that model's sufficient statistic. However, that sufficient statistic is composed of inner products between the vector of observed X_j's and each of the vectors of observed values of the other covariates $X_{\text{-}j}$, precluding inference except in the case of covariates with a very small set of discrete values, and almost entirely precluding inference in a high-dimensional setting. A different conditional permutation test was recently proposed by [8] to test conditional independence in the model-X framework, but while their conditioning improves robustness, they still require the same assumptions as the original conditional randomization test [10], namely, that $X_j \mid X_{\text{-}j}$ is known exactly. To our knowledge, the present paper is the first to use the idea of conditioning on sufficient statistics for high-dimensional inference, enabling powerful and exact FDR-controlled variable selection under arguably weaker assumptions than any existing work.

1.4. *Outline.* The rest of the paper is structured as follows: Section 2 describes the main result and the proposed method of conditional knockoffs to generalize model-X knockoffs to the case when F_X is known only up to a distributional family, as opposed to exactly. Section 3 applies conditional knockoffs to three different models for F_X, and provides explicit algorithms for constructing valid knockoffs. Simulations are also presented, showing that conditional knockoffs often loses almost no power in exchange for its increased generality over model-X knockoffs with exactly-known F_X. Finally, Section 4 provides some synthesis of the ideas in this paper and directions for future work.

2. Main idea and general principles. Before going into more detail, we introduce some notation. Suppose we are given i.i.d. row vectors $(Y_i, X_{i,1}, \dots, X_{i,p}) \in \mathbb{R}^{p+1}$ for $i = 1, \dots, n$. We then stack these vectors into a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ whose ith row is denoted by $\boldsymbol{x}_i^\top = (X_{i,1}, \dots, X_{i,p}) \in \mathbb{R}^p$, and a column vector $\boldsymbol{y} \in \mathbb{R}^n$ whose ith entry is Y_i. We are about to define model-X knockoffs $(\tilde{X}_{i,1}, \dots, \tilde{X}_{i,p})$, and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$ will analogously denote these row vectors stacked to form a knockoff matrix. A square bracket around matrices, such as $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$, denotes the horizontal concatenation of these matrices. We use $[p]$ for $\{1, 2, \dots, p\}$, and $i : j$ for $\{i, i+1, \dots, j\}$ for any $i \le j$; for a set $A \subseteq [p]$, let $\boldsymbol{X}_A$ denote the matrix

with columns given by the columns of $\boldsymbol{X}$ whose indices are in A, and for singleton sets we streamline notation by writing $\boldsymbol{X}_j$ instead of $\boldsymbol{X}_{\{j\}}$. For sets $A_1, \ldots, A_m$, denote by $\prod_{j=1}^m A_j$ their Cartesian product. For two disjoint sets A and B, we denote their union by $A \uplus B$. We will denote by $\mathbb{N}$ the set of strictly positive integers.

2.1. *Model-X knockoffs*. We begin with a short review of model-X knockoffs [10]. The authors define model-X knockoffs for a random vector $X \in \mathbb{R}^p$ of covariates as being a random vector $\tilde{X} \in \mathbb{R}^p$ such that for any set $A \subseteq [p]$

$$\tilde{X} \perp\!\!\!\perp Y \mid X \quad \text{and} \quad (X, \tilde{X})_{\text{swap}(A)} \overset{\mathcal{D}}{=} (X, \tilde{X}), \tag{2.1}$$

where the swap(A) subscript on a $2p$-dimensional vector (or matrix with $2p$ columns) denotes that vector (matrix) with the jth and $(j+p)$th entries (columns) swapped, for all $j \in A$. To use knockoffs for variable selection, suppose some statistics Z_j and $\tilde{Z}_j$ are used to measure the importance of X_j and $\tilde{X}_j$, respectively, in the conditional distribution $Y \mid X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p$, with

$$(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}),$$

for some function z such that swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ swaps the components Z_j and $\tilde{Z}_j$, that is, for any $A \subseteq [p]$,

$$z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}(A)}, \boldsymbol{y}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})_{\text{swap}(A)}.$$

For example, $z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$ could perform a cross-validated Lasso regression of $\boldsymbol{y}$ on $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ and return the absolute values of the $2p$-dimensional fitted coefficient vector. More generally, the Z_j can be almost any measure of variable importance one can think of, including measures derived from arbitrarily-complex machine learning methods or from Bayesian inference, and this flexibility allows model-X knockoffs to be powerful even when $F_{Y|X}$ is quite complex.

The pairs $(Z_j, \tilde{Z}_j)$ of variable importance measures are then plugged into scalar-valued antisymmetric functions f_j to produce $W_j = f_j(Z_j, \tilde{Z}_j)$, which measures the *relative* importance of X_j to $\tilde{X}_j$. Viewed as a function of all the data, $W_j = w_j([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$ can be shown to satisfy the *flip-sign* property, which dictates that for any $A \subseteq [p]$,

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

Taking Z_j and $\tilde{Z}_j$ as the absolute values of Lasso coefficients as in the above example, one might choose $W_j = Z_j - \tilde{Z}_j$, referred to in Candès et al. [10] as the *Lasso coefficient-difference* (LCD) statistic. Finally, given a target FDR level q, the knockoff filter selects the variables $\hat{S} = \{j : W_j \geq T\}$ where T is either the *knockoff threshold* T_0 or the *knockoff+ threshold* T_+:

$$T_0 = \min\left\{t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\},$$

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

Candès et al. [10], Theorem 3.4, prove that $\hat{S}$ with T_+ exactly (nonasymptotically) controls the FDR at level q, and that $\hat{S}$ with T_0 exactly controls a modified FDR, $\mathbb{E}[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| + 1/q}]$, at level q. The key to the proof of exact control is the aforementioned flip-sign property of the W_j,

and that property follows from the following crucial property of model-X knockoffs: for any subset $A \subseteq \mathcal{H}_0$,

$$([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(A)}, \boldsymbol{y}) \overset{\mathcal{D}}{=} ([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y}),$$

which is proved in Candès et al. [10], Lemma 3.2, to hold for knockoffs satisfying equation (2.1).

The proofs of exact control required just one assumption, that one could construct knockoffs satisfying equation (2.1). To satisfy that assumption, Candès et al. [10] assumes throughout that F_X is known exactly. We will relax this assumption, but first slightly generalize the definition of valid knockoffs.

DEFINITION 2.1 (Model-X knockoff matrix). The random matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ is a *model-X knockoff matrix* for the random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ if for any subset $A \subseteq [p]$,

$$\tilde{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{X} \quad \text{and} \quad [\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}]. \tag{2.2}$$

Note that equation (2.2) is more general than equation (2.1), and indeed (2.1) implies (2.2) as long as the rows of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ are independent. However, the proof of Candès et al.'s [10] crucial Lemma 3.2 and, ultimately, FDR control in the form of their Theorem 3.4 used only equation (2.2). Therefore, Definition 2.1 is the "correct" definition, since the ability to generate knockoffs satisfying Definition 2.1 is all that is needed for the theoretical guarantees of knockoffs in Candès et al. [10] to hold, and it is well defined for any matrix $\boldsymbol{X}$, even when the rows are not independent. We will use this general definition because although we also assume samples are drawn i.i.d. from a distribution, those samples will no longer be independent when we condition on a sufficient statistic for the model for F_X. Hereafter, *model-X knockoffs* and *knockoffs* will always refer to model-X knockoff matrices as defined by Definition 2.1 unless otherwise specified.

For completeness, we restate the FDR control theorem in Candès et al. [10].

THEOREM 2.1. *Suppose $\tilde{\boldsymbol{X}}$ is a knockoff matrix for $\boldsymbol{X}$ and the statistics W_j's satisfy the flip-sign property. For any $q \in [0, 1]$, if $\hat{S}$ is selected by the knockoff method with threshold T being either T_+ or T_0, then*

$$\mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{\max(|\hat{S}|, 1)}\right] \le q \quad \textit{for } T_+; \qquad \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| + 1/q}\right] \le q \quad \textit{for } T_0.$$

It is worth mentioning that if $\tilde{\boldsymbol{X}}_j$ is identical to $\boldsymbol{X}_j$, then $W_j = 0$ and j cannot be selected by the knockoff filter. Formally, we call such a column in the knockoff matrix *trivial*.

2.2. *Conditional knockoffs.* The main idea of this paper is that if F_X is known only up to a parametric model, and that parametric model has sufficient statistic (for n i.i.d. observations drawn from F_X) given by $T(\boldsymbol{X})$, then by definition of sufficiency the distribution of $\boldsymbol{X} \mid T(\boldsymbol{X})$ does not depend on the model parameters and is thus known exactly a priori. To leverage this for knockoffs, consider the following definition.

DEFINITION 2.2 (Conditional model-X knockoff matrix). The random matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ is a *conditional model-X knockoff matrix* for the random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ if there is a statistic $T(\boldsymbol{X})$ such that for any subset $A \subseteq [p]$,

$$\tilde{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{X} \quad \text{and} \quad [\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}] \,\big|\, T(\boldsymbol{X}). \tag{2.3}$$

By the law of total probability, (2.3) implies (2.2), thus conditional model-X knockoffs are also model-X knockoffs.

PROPOSITION 2.2. *If $\tilde{\boldsymbol{X}}$ is a conditional model-X knockoff matrix for $\boldsymbol{X}$, then it is also a model-X knockoff matrix.*

Proposition 2.2 says that all the guarantees of model-X knockoffs (i.e., Theorem 2.1), such as exact FDR control and the flexibility in measuring variable importance, immediately hold when $\tilde{\boldsymbol{X}}$ is a *conditional* model-X knockoff matrix. Definition 2.2 is especially useful when the distribution of $\boldsymbol{X}$ is known to be in a model $G_{\boldsymbol{\Theta}} = \{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with parameter space $\boldsymbol{\Theta}$, and $T(\boldsymbol{X})$ is a sufficient statistic for $G_{\boldsymbol{\Theta}}$, because then the distribution of $\boldsymbol{X} \mid T(\boldsymbol{X})$ is known exactly even though the unconditional distribution of $\boldsymbol{X}$ is not. Exact knowledge of the distribution of $\boldsymbol{X} \mid T(\boldsymbol{X})$ in principle allows us to construct knockoffs, similar to how exact knowledge of the unconditional distribution of $\boldsymbol{X}$ has enabled all previous knockoff construction algorithms. As a simple example, when G_Θ is the set of all p-dimensional distributions with mutually-independent entries, the set of order statistics for each column of $\boldsymbol{X}$ constitutes a sufficient statistic $T(\boldsymbol{X})$, and a conditional knockoff matrix $\tilde{\boldsymbol{X}}$ can be generated by randomly and independently permuting each column of $\boldsymbol{X}$. Unfortunately, for more interesting models that allow for dependence among the covariates, even for canonical $G_{\boldsymbol{\Theta}}$ like multivariate Gaussian, the distribution of $\boldsymbol{X} \mid T(\boldsymbol{X})$ is often much more complex than those for which knockoff constructions already exist. Using novel methodological and theoretical tools, in Section 3 we provide efficient and exact algorithms for constructing nontrivial conditional knockoffs when F_X comes from each of the following three models:

1. *Low-dimensional Gaussian*:

$$F_X \in \{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}, \boldsymbol{\Sigma} \succ \mathbf{0}\},$$

when $n > 2p$. In this case, the number of model parameters is $p + \frac{p(p+1)}{2} = \Omega(p^2)$, and also $\Omega(np)$ in the most challenging case when $p = \Omega(n)$.

2. *Gaussian graphical model*:

$$F_X \in \{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}, \boldsymbol{\Sigma} \succ \mathbf{0}, \\ (\boldsymbol{\Sigma}^{-1})_{j,k} = 0 \text{ for all } (j,k) \notin E\}$$

for some known sparsity pattern E. For example, $\boldsymbol{\Sigma}^{-1}$ could be banded with bandwidth as large as $n/8 - 1$, provided $n/8 \le p$, allowing a number of parameters as large as $p + (\frac{np}{8} - \frac{n(n-8)}{128}) = \Omega(np)$. Note that p is not explicitly constrained, so this model allows both low- and high-dimensional data sets.

3. *Discrete graphical model*:

$$F_X \in \Bigg\{\text{distribution on } \prod_{j=1}^{p}[K_j] : X_j \perp\!\!\!\perp X_{[p]\setminus N_E(j)} \,\Big|\, X_{N_E(j)\setminus\{j\}} \\ \text{for all } (j,k) \notin E\Bigg\}$$

for some known positive integers $K_1, \ldots, K_p$ and known sparsity pattern E, where $N_E(j)$ is the closed neighborhood of j. For example, X could be a K-state (nonstationary) Markov chain whose $K - 1 + (p-1)K(K-1)$ parameters are the probability mass function of X_1 and the transition matrices $\mathbb{P}(X_j \mid X_{j-1})$ for each $j \in \{2, \ldots, p\}$, where K can be as large as

$\sqrt{\frac{n-2}{2}}$, allowing a number of parameters as large as $\sqrt{\frac{n-2}{2}} - 1 + (p-1)(\sqrt{\frac{n-2}{2}})(\sqrt{\frac{n-2}{2}} - 1) = \Omega(np)$. Again, p is not explicitly constrained, so this model allows both low- and high-dimensional data sets.

REMARK 1. It is worth mentioning that conditioning may shrink the set of nonnull hypotheses. For instance, if $\mathcal{H}_0 = \varnothing$ and $T(\boldsymbol{X})$ is chosen to be $\boldsymbol{X}$, then all variables are automatically null conditional on $T(\boldsymbol{X})$, and thus conditional knockoffs cannot select any nonnull variables. For a detailed discussion, see the Supplementary Material Appendix C [18].

REMARK 2. Any algorithm that generates conditional knockoffs given one sufficient statistic $T(\boldsymbol{X})$ (i.e., satisfying equation (2.3) for $T(\boldsymbol{X})$) by definition is also a valid algorithm for generating conditional knockoffs given any sufficient statistic $S(\boldsymbol{X})$ that is a function of $T(\boldsymbol{X})$. This means that any valid conditional knockoff algorithm satisfies equation (2.3) for the minimal sufficient statistic, since by definition a minimal sufficient statistic is a function of any other sufficient statistic. So we could say that the minimal sufficient statistic is in some sense the optimal one to condition on, in that the choice to condition on the minimal sufficient statistic allows for the most general set of conditional knockoff algorithms of any sufficient statistic one could choose to condition on for a given model.

2.3. *Integrating unlabeled data.* In addition to the n labeled pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$, we might also have unlabeled data $\{\boldsymbol{x}_i^{(u)}\}_{i=1}^{n^{(u)}}$, that is, covariate samples without corresponding responses/labels. This extra data can be integrated seamlessly into the construction of conditional knockoffs: stack the labeled covariate matrix $\boldsymbol{X}$ on top of the unlabeled covariate matrix $\boldsymbol{X}^{(u)}$ to get $\boldsymbol{X}^* \in \mathbb{R}^{n^* \times p}$, where $n^* = n + n^{(u)}$, then construct conditional knockoffs $\tilde{\boldsymbol{X}}^*$ for $\boldsymbol{X}^*$, and finally take $\tilde{\boldsymbol{X}}$ to be the first n rows of $\tilde{\boldsymbol{X}}^*$.

PROPOSITION 2.3. *Suppose the rows of $\boldsymbol{X}^*$ are i.i.d. covariate vectors and $\boldsymbol{X}$ is the matrix composed of the first n rows of $\boldsymbol{X}^*$. Let $\boldsymbol{y}$ be the response vector for $\boldsymbol{X}$. If for some statistic $T(\boldsymbol{X}^*)$ and any set $A \subseteq [p]$,*

$$\tilde{\boldsymbol{X}}^* \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{X}^* \quad \text{and} \quad [\boldsymbol{X}^*, \tilde{\boldsymbol{X}}^*]_{\text{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}^*, \tilde{\boldsymbol{X}}^*] \mid T(\boldsymbol{X}^*),$$

then if $\tilde{\boldsymbol{X}}$ is the matrix composed of the first n rows of $\tilde{\boldsymbol{X}}^*$, then $\tilde{\boldsymbol{X}}$ is a model-X knockoff matrix for $\boldsymbol{X}$.

Note that by taking $T(\boldsymbol{X}^*)$ to be constant, the same result holds unconditionally: if $\tilde{\boldsymbol{X}}^* \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{X}^*$ and $[\boldsymbol{X}^*, \tilde{\boldsymbol{X}}^*]_{\text{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}^*, \tilde{\boldsymbol{X}}^*]$ for any $A \subseteq [p]$, then $\tilde{\boldsymbol{X}}$ is a valid knockoff matrix for $\boldsymbol{X}$. Thus constructing knockoffs for $\boldsymbol{X}^*$, conditional or otherwise, produces valid knockoffs for $\boldsymbol{X}$ automatically. Of course, if F_X is known and the rows of $\boldsymbol{X}^*$ are i.i.d., it is natural to construct each row of $\tilde{\boldsymbol{X}}^*$ independently, in which case the presence of $\boldsymbol{X}^{(u)}$ changes nothing about the construction of the relevant knockoffs $\tilde{\boldsymbol{X}}$. But as seen in Section 2.2, when F_X is not known exactly the flexibility with which we can model it depends on the sample size, with the number of parameters allowed to be as large as $\Omega(np)$ in all the models in this paper. What Proposition 2.3 shows is that n can be replaced with n^*, which can dramatically increase the modeling flexibility allowed by conditional knockoffs, especially in high dimensions. For example, our conditional knockoffs construction in Section 3.1 for arbitrary multivariate Gaussian distributions naively requires $n > 2p$, but we now see it actually just requires $n^* > 2p$, which is much easier to satisfy when $n^{(u)}$ is large, as it often is in, for instance, genomics or economics applications. Even when n alone is large enough to construct nontrivial knockoffs for a desired model, constructing conditional knockoffs with unlabeled data as described in this section will tend to increase power.

3. Conditional knockoffs for three models of interest. In this section, we provide efficient algorithms to generate exact conditional model-X knockoffs under three different models for F_X, as well as numerical simulations comparing the variable selection power of the knockoffs thus constructed with those constructed by existing algorithms that require F_X be known exactly.

All proofs are deferred to the Supplementary Material Appendix A [18]. Any sampling described in the algorithms is conducted independently of all previous sampling in the same algorithm, unless stated otherwise. All simulations use a Gaussian linear model for the response: $Y_i \mid \boldsymbol{x}_i \sim \mathcal{N}(\frac{1}{\sqrt{n}}\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1)$ where $\boldsymbol{\beta}$ has 60 nonzero entries with random signs and equal amplitudes. Note the sparsity and magnitude equalities are simply chosen for convenience—we present additional simulations varying these choices in the Supplementary Material Appendix D.2 [18]. We remind the reader that, although we use linear regression as an illustrative example in the simulations, our methods apply to more general regressions, and all the same simulations are also rerun with a nonlinear model (logistic regression) with similar results, presented in the Supplementary Material Appendix D.1 [18]. We use the LCD knockoff statistic with tuning parameter chosen by 10-fold cross-validation and the knockoff+ threshold with target FDR $q = 20\%$; see Section 2.1 for details. Only power curves (power $= \mathbb{E}[\frac{|S\cap\hat{S}|}{|S|}]$) are shown because the FDR is always controlled (both theoretically and empirically). The procedure we compare to, unconditional knockoffs, refers to model-X knockoffs where F_X is taken to be known exactly (knockoff statistics and thresholds are chosen identically).

3.1. *Low-dimensional multivariate Gaussian model.* Despite the focus in variable selection on high-dimensional problems, we start with a low-dimensional example as it represents an interesting and instructive case. Suppose that

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.1}$$

for some unknown $\boldsymbol{\mu}$ and positive definite $\boldsymbol{\Sigma}$. Let $\hat{\boldsymbol{\mu}} := \boldsymbol{X}^\top \mathbf{1}_n/n$ denote the vector of column means of $\boldsymbol{X}$, and let $\hat{\boldsymbol{\Sigma}} := (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)^\top(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)/n$ be the empirical covariance matrix of $\boldsymbol{X}$. Then $T(\boldsymbol{X}) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ constitutes a (minimal, complete) sufficient statistic for the model (3.1) for $\boldsymbol{X}$.

3.1.1. *Generating conditional knockoffs.* When $n > 2p$, we can construct knockoffs for $\boldsymbol{X}$ conditional on $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ via Algorithm 3.1.

Algorithm 3.1 Conditional knockoffs for low-dimensional Gaussian models

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$.
Require: $n > 2p$.
1: Find $\boldsymbol{s} \in \mathbb{R}^p$ such that $\mathbf{0}_{p\times p} \prec \operatorname{diag}\{\boldsymbol{s}\} \prec 2\hat{\boldsymbol{\Sigma}}$.
2: Compute the Cholesky decomposition of $n(2\operatorname{diag}\{\boldsymbol{s}\} - \operatorname{diag}\{\boldsymbol{s}\}\hat{\boldsymbol{\Sigma}}^{-1}\operatorname{diag}\{\boldsymbol{s}\})$ as $\boldsymbol{L}^\top\boldsymbol{L}$.
3: Generate $\boldsymbol{W}$ a $n \times p$ matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ and independent of $\boldsymbol{X}$ and compute the Gram–Schmidt orthonormalization $[\underbrace{\boldsymbol{Q}}_{n\times(p+1)}, \underbrace{\boldsymbol{U}}_{n\times p}]$ of the columns of $[\mathbf{1}_n, \boldsymbol{X}, \boldsymbol{W}]$.
4: Set $\tilde{\boldsymbol{X}} = \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}^{-1}\operatorname{diag}\{\boldsymbol{s}\}) + \boldsymbol{U}\boldsymbol{L}$. (3.2)
5: **return** $\tilde{\boldsymbol{X}}$.

In Algorithm 3.1, $n > 2p$ is needed because in Line 3 the $n \times (2p+1)$ matrix $[\mathbf{1}_n, \boldsymbol{X}, \boldsymbol{W}]$ must have at least as many rows as columns to be a valid input to the Gram–Schmidt orthonormalization algorithm. The astute reader may notice a strong similarity between equation (3.2) and the fixed-X knockoff construction in Barber and Candès [1], equation (1.4). Indeed nearly the same tools can be used to find a suitable $\boldsymbol{s}$; in the Supplementary Material Appendix B.1 [18] we slightly adapt three methods from Barber and Candès [1] and Candès et al. [10] for computing suitable $\boldsymbol{s}$. The computational complexity of Algorithm 3.1 depends on the method used to find $\boldsymbol{s}$, with the fastest option requiring $O(np^2)$ time.

The differences between equation (3.2) and the fixed-X knockoff construction are the additional accounting for the mean by adding/subtracting $\hat{\boldsymbol{\mu}}$, the lack of requiring that $\boldsymbol{X}$ have normalized columns, the "$\prec$" relationships (as opposed to "$\preceq$"), and most importantly the requirement that $\boldsymbol{U}$ be random. Indeed, as can be seen in the proof of Theorem 3.1, the precise uniform distribution of $\boldsymbol{U}$ is crucial. And it bears repeating that unlike fixed-X knockoffs, Algorithm 3.1 produces valid *model-X* knockoffs, and hence permits importance statistics without the "sufficiency property" and applies to *any* $F_{Y|X}$, not just homoscedastic linear regression.

THEOREM 3.1. *Algorithm* 3.1 *generates valid knockoffs for model* (3.1).

The challenge in proving Theorem 3.1 is that the conditional distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \mid T(\boldsymbol{X})$ is supported on an uncountable subset of zero Lebesgue measure, and its distribution is only defined through the distribution of $\boldsymbol{X} \mid T(\boldsymbol{X})$ and the conditional distribution of $\tilde{\boldsymbol{X}} \mid \boldsymbol{X}$. Although $\boldsymbol{X} \mid T(\boldsymbol{X})$ and $\tilde{\boldsymbol{X}} \mid \boldsymbol{X}$ are both conditionally uniform on their respective supports, and the latter's normalizing constant does not depend on $\boldsymbol{X}$, these facts alone are not sufficient to conclude that $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \mid T(\boldsymbol{X})$ is uniform on its support (see Appendix A.2.1 in the Supplementary Material [18] for a simple counterexample), which is what we need to prove. Although these distributions on zero-Lebesgue-measure manifolds can be characterized using geometric measure theory (as in, e.g., Diaconis et al. [13]), we bypass this approach by directly using the concept of invariant measures from topological measure theory. Since these tools are new to the knockoffs literature and their use may be of independent interest, we include below a brief sketch of the proof of Theorem 3.1, deferring the full proof to Appendix A.2.2 in the Supplementary Material [18].

The proof of Theorem 3.1 follows three steps: we first show that the conditional distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \mid T(\boldsymbol{X})$ is invariant on its support to multiplication by elements of the topological group of orthonormal matrices that have $\mathbf{1}_n$ as a fixed point, and then show that the conditional distribution remains invariant (on the same support) after swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$. Finally, we show that the invariant measure on the support of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \mid T(\boldsymbol{X})$ is unique. These three steps combined show that the distributions before and after swapping are the same, and hence $\tilde{\boldsymbol{X}}$ is a valid conditional knockoff matrix for $\boldsymbol{X}$.

A useful consequence of Theorem 3.1 is the double robustness property that if knockoffs are constructed by Algorithm 3.1 and knockoff statistics are used which obey the sufficiency property of Barber and Candès [1] (i.e., the knockoff statistics only depend on $\boldsymbol{y}$ and $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ through $[\mathbf{1}_n, \boldsymbol{X}, \tilde{\boldsymbol{X}}]^\top \boldsymbol{y}$ and $[\mathbf{1}_n, \boldsymbol{X}, \tilde{\boldsymbol{X}}]^\top [\mathbf{1}_n, \boldsymbol{X}, \tilde{\boldsymbol{X}}]$), then the resulting variable selection controls the FDR exactly as long *as at least one of* the following holds:

- $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, both unknown (*regardless of* $F_{Y|X}$), or
- $y_i \mid \boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ for some $\boldsymbol{\beta}$ and σ^2, both unknown (*regardless of* F_X).

In the Supplementary Material Appendix B.1 [18] we extend Algorithm 3.1 to the case when the mean is known (Algorithm B.1) or a subset of columns of $\boldsymbol{X}$ are additionally conditioned on (Algorithm B.2). Both extensions may be of independent interest, but will also be used as subroutines when generating knockoffs for Gaussian graphical models in Section 3.2.

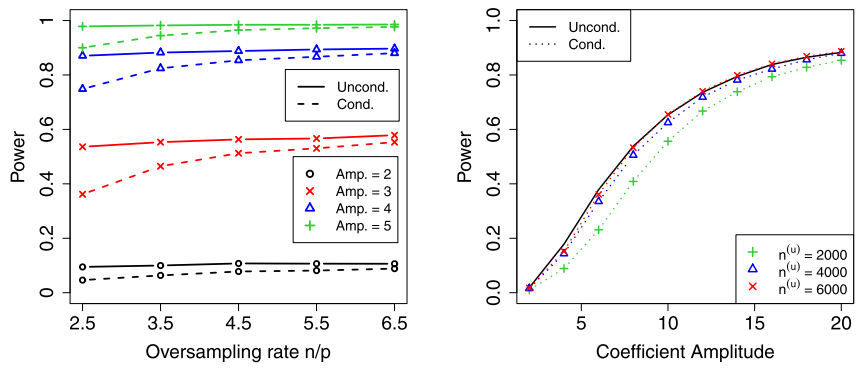


FIG. 1. *Power curves of conditional and unconditional knockoffs for an* AR(1) *model with* $p = 1000$ (a) *as* n/p *varies for various coefficient amplitudes and* (b) *as the coefficient amplitude varies for various values of* $n^{(u)}$, *with* $n = 300$ *fixed. Standard errors are all below* 0.008.

3.1.2. *Numerical examples.* We present two simulations comparing the power of conditional knockoffs to the analogous unconditional construction that uses the exactly-known F_X. We remind the reader that the simulation setting is at the beginning of Section 3. The vector $\boldsymbol{s}$ in Algorithm 3.1 is computed using the SDP method of equation (B.1) in the Supplementary Material [18], and the analogous vector for the unconditional construction is chosen by the analogous SDP method [10]. Although in both examples $n^* > 2p$, the number of unknown parameters in the Gaussian model for F_X is $p + \frac{p(p+1)}{2} > 500{,}000$, vastly larger than any of the sample sizes.

Figure 1(a) fixes $p = 1000$ and plots the difference in power between unconditional and conditional knockoffs as $n > 2p$ increases for a few different signal amplitudes. The power of the conditional and unconditional constructions is quite close except when $n = 2.5p$ is just above its threshold of $2p$, and even then the power of the conditional construction is respectable.

Figure 1(b) shows how unlabeled samples improve the power of conditional knockoffs. The model is the same as the first example but the labeled sample size is fixed at $n = 300$ and we vary the number of unlabeled samples. Again, the power of the conditional and unconditional constructions is extremely close except when $n^* = 2.3p$ is just above its threshold, and again even in that setting the power of the conditional construction is respectable. Note that unlabeled samples here have enabled the *low-dimensional* Gaussian construction to apply in a high-dimensional setting with $n < p$, since $n^* > 2p$.

3.2. *Gaussian graphical model.* Ignoring unlabeled data, the method of the previous subsection is constrained to low-dimensional (or perhaps more accurately, medium-dimensional, since it allows $p = \Omega(n)$) settings and cannot be immediately extended to high dimensions. In many applications, however, particularly in high dimensions, the covariates are modeled as multivariate Gaussian with *sparse* precision matrix $\boldsymbol{\Sigma}^{-1}$, and when the sparsity pattern is known a priori, we can condition on much less. For instance, time series models such as autoregressive models assume a banded precision matrix with known bandwidth, and the model used in this subsection would also allow for nonstationarity. Spatial models often assume a (known) neighborhood structure such that the only nonzero precision matrix entries are index pairs corresponding to spatial neighbors.

Precisely, suppose $\boldsymbol{X}$'s rows $\boldsymbol{x}_i^\top$ are i.i.d. draws from a distribution known to be in the model

$$\{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \succ \mathbf{0}, (\boldsymbol{\Sigma}^{-1})_{j,k} = 0 \text{ for all } j \neq k \text{ and } (j,k) \notin E\}, \tag{3.3}$$

where $E \subseteq [p] \times [p]$ is some symmetric set of integer pairs (i.e., $(j,k) \in E \Rightarrow (k,j) \in E$) with no self-loops. Then the undirected graph $G := ([p], E)$ defines a Gaussian graphical model with vertex set $[p]$ and edge set E. For any $j \in [p]$, define $I_j = \{k : (j,k) \in E\}$ for the vertices that are adjacent to j. We will use the terms "vertex" ($j \in [p]$) and "variable" (X_j) interchangeably. $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}_E$ together constitute a sufficient statistic, where $\hat{\boldsymbol{\Sigma}}_E := \{\hat{\boldsymbol{\Sigma}}_{j,k} : j = k \text{ or } (j,k) \in E\}$. We will show in this section how to generate conditional knockoffs, and we will characterize the sparsity patterns E for which we can generate knockoffs with $\tilde{\boldsymbol{X}}_j \neq \boldsymbol{X}_j$ for all $j \in [p]$.

REMARK 3. More generally, sparsity in the precision matrix, but with *unknown* sparsity pattern, is a common assumption in Gaussian graphical models which are used to model many types of data in high dimensions such as gene expressions. Although the construction in this section no longer holds exactly when the sparsity pattern is unknown, approximate knockoffs could still be constructed by first using a method for estimating the sparsity pattern [9], Chapter 13, and then treating it as known. Note that we only require the edge set E to contain all non-zero entries of $\boldsymbol{\Sigma}^{-1}$, which is no harder than the exact identification of the nonzero entries.

3.2.1. *Generating conditional knockoffs by blocking*. First, consider the ideal case when the graph G separates into disjoint connected components whose respective vertex sets are $V_1, \dots, V_\ell$. Then X can be divided into independent subvectors, $X_{V_1}, \dots, X_{V_\ell}$, and if each $|V_k| < n/2$, we can construct low-dimensional conditional knockoffs separately and independently for each $\boldsymbol{X}_{V_k}$ as in Section 3.1. Moving to the general case when G is connected, we can do something intuitively similar by conditioning on a subset of variables in addition to $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}_E$. If there is a subset of vertices B such that the subgraph G_B induced by deleting B separates into small disjoint connected components, then we should be able to construct conditional knockoffs as above for $\boldsymbol{X}_{B^c}$ by conditioning on $\boldsymbol{X}_B$. We think of the variables in B as being *blocked* to separate the graph into small disjoint parts, hence we refer to this B as a *blocking set*.

The following definition formalizes when we can apply the above procedure, and Algorithm 3.2 states that procedure precisely.

DEFINITION 3.1. A graph G is *n-separated* by a set $B \subset [p]$ if the subgraph G_B induced by deleting all vertices in B has connected components whose respective vertex sets we denote by $V_1, \dots, V_\ell$ such that for all $k \in [\ell]$,

$$2|V_k| + |I_{V_k} \cap B| < n,$$

where $I_{V_k} := \bigcup_{j \in V_k} I_j$ is the neighborhood of V_k in G.

Algorithm 3.2 Conditional knockoffs for Gaussian graphical models

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $G = ([p], E)$, $B \in [p]$.

Require: For some $n' \leq n$, G is n'-separated by B into connected component vertex sets $V_1, \dots, V_\ell$.

1: **for** $k = 1, \dots, \ell$ **do**

2: Construct partial low-dimensional knockoffs $\tilde{\boldsymbol{X}}_{V_k}$ for $\boldsymbol{X}_{V_k}$ conditional on $\boldsymbol{X}_{I_{V_k} \cap B}$ via Algorithm B.2 (a slight modification of Algorithm 3.1).

3: **end for**

4: Set $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$.

5: **return** $\tilde{\boldsymbol{X}}$.

Note that when the V_k separated X into independent subvectors, we only needed $2|V_k| < n$; now that they only represent *conditionally* independent subvectors, we must also account for V_k's neighbors in B that we condition on, resulting in the requirement that $2|V_k| + |I_{V_k} \cap B| < n$.

Algorithm 3.2 constructs knockoffs for the model (3.3) by first conditioning on $\boldsymbol{X}_B$ and then running a slight modification of Algorithm 3.1 (Algorithm B.2 in the Supplementary Material Appendix B.1.3 [18]) on the variables/columns V_k corresponding to the induced subgraphs. The computational complexity of Algorithm 3.2 is $O(n \sum_{k=1}^{\ell}(|I_{V_k} \cap B|^2|V_k| + |V_k|^2))$, which is upper bounded by the simpler expression $O(np(n' + \max_{k\in[\ell]} |I_{V_k} \cap B|^2))$ (both complexities assume the most efficient construction of $\boldsymbol{s}$ is used as a primitive in Algorithm B.2).

THEOREM 3.2. *Algorithm 3.2 generates valid knockoffs for model* (3.3).

Algorithm 3.2 raises two key issues: how to find a suitable blocking set B, and how to address the fact that $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$ are trivial knockoffs, so using conditional knockoffs from Algorithm 3.2 will have no power to select any of the variables in B.

Algorithm 3.3 provides a simple greedy way to find a suitable B or, given an initial blocking set B, can also be used to shrink B (see Proposition B.3). The algorithm visits every vertex in G once in the order π and decides whether each vertex it visits is blocked or *free* (not blocked). Meanwhile, it constructs a graph $\bar{G}$ from G, which gets expanded every time a vertex j is determined to be free: all pairs of j's neighbors in $\bar{G}$ get connected (if not already) and a new vertex $\tilde{j}$ that has the same neighborhood as j in $\bar{G}$ is added to the graph. A vertex is blocked if, when it is visited, its degree in $\bar{G}$ is greater than $n' - 3$.

PROPOSITION 3.3. *If B is the blocking set determined by Algorithm 3.3 with input (π, n'), then G is n-separated by B for any $n \geq n'$.*

Algorithm 3.3 is meant to be intuitive but a more efficient implementation is given in the Supplementary Material Appendix B.2 [18]. Algorithm 3.3 can also be made even greedier by choosing the next j at each step as the unvisited vertex in $[p]$ with the smallest degree in $\bar{G}$ (breaking ties at random), instead of following the ordering π. The algorithm also takes an input n', which one may prefer to choose smaller than n for computational or statistical efficiency, as we investigate in Section 3.2.2 (smaller n' will mean smaller V_k to generate knockoffs for in Line 2 of Algorithm 3.2). The flexibility in both π and n' is mainly motivated by the second aforementioned issue of trivial knockoffs $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$, addressed next.

Algorithm 3.3 Greedy search for a blocking set

Input: π a permutation of $[p]$, $G = ([p], E)$, n'.

1: Initialize a graph $\bar{G} = G$, and $B = \varnothing$.
2: **for** $t = 1, \ldots, p$ **do**
3: Let $j = \pi_t$, and $\bar{I}_j$ be the neighborhood of j in the graph $\bar{G}$.
4: **if** $n' \geq 3 + |\bar{I}_j|$ **then**
5: Add edges between all pairs of vertices in $\bar{I}_j$.
6: Add a vertex $\tilde{j}$ to $\bar{G}$ and add edges between $\tilde{j}$ and all vertices in $\bar{I}_j$.
7: **else**
8: $B \leftarrow B \cup \{j\}$.
9: **end if**
10: **end for**
11: **return** B.

Algorithm 3.4 Conditional knockoffs for Gaussian graphical models with data splitting

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $G = ([p], E)$, $B_1, \dots, B_m \subset [p]$, $n_1, \dots, n_m \in \mathbb{N}$
Require: $\bigcup_{i=1}^m B_i^c = [p]$, G is n_i-separated by B_i for all $i = 1, \dots, m$, and $\sum_{i=1}^m n_i = n$.
1: Partition the rows of $\boldsymbol{X}$ into submatrices $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(m)}$ with each $\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i\times p}$.
2: **for** $i = 1, \dots, m$ **do**
3: Run Algorithm 3.2 on $\boldsymbol{X}^{(i)}$ with blocking set B_i to obtain $\tilde{\boldsymbol{X}}^{(i)}$.
4: **end for**
5: **return** $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}^{(1)}; \dots; \tilde{\boldsymbol{X}}^{(m)}]$ (the row-concatenation of the $\tilde{\boldsymbol{X}}^{(i)}$'s).

An intuitive solution to prevent the trivial knockoffs $\tilde{\boldsymbol{X}}_B$ in Algorithm 3.2 is to split the rows of $\boldsymbol{X}$ in half and run Algorithm 3.2 on each half with disjoint blocking sets B_1 and B_2 such that G is $n/2$-separated by both blocking sets. Then the knockoffs for variables in B_1 will be trivial for half the rows of $\tilde{\boldsymbol{X}}$ and those for variables in B_2 will be trivial for the other half of the rows of $\tilde{\boldsymbol{X}}$, but since B_1 and B_2 are disjoint, no variables will have entirely trivial knockoffs. Even though some knockoff variables are trivial for half their rows, we find the power loss for these variables to be surprisingly small, see the simulations in Section 3.2.2.

This data-splitting idea is generalized in Algorithm 3.4 to splitting the rows of $\boldsymbol{X}$ into m folds and running Algorithm 3.2 on each fold with a different input B.

In Algorithm 3.4, since $\bigcup_{i=1}^m B_i^c = [p]$, for each $j \in [p]$ there is at least one i such that $j \notin B_i$, and thus $\tilde{\boldsymbol{X}}_j \neq \boldsymbol{X}_j$. Before characterizing when it is possible to find such B_i, we formalize the requirements of Algorithm 3.4 into a definition.

DEFINITION 3.2. $G = ([p], E)$ is (m, n)-*coverable* if there exist $B_1, \dots, B_m$ subsets of $[p]$ and integers $n_1 \dots, n_m$ such that $\bigcup_{i=1}^m B_i^c = [p]$, G is n_i-separated by B_i for all $i = 1, \dots, m$, and $\sum_{i=1}^m n_i \le n$.

The following common graph structures are (m, n)-coverable:

- If the largest connected component of G is not larger than $(n-1)/2$, G is $(1, n)$-coverable.
- If G is a Markov chain of order r (making the model a time-inhomogeneous AR(r) model), that is, $E = \{(i, j) : 1 \le |i - j| \le r\}$, and $n \ge 2 + 8r$, then G is $(2, n)$-coverable.
- If G is a *m-colorable* (also known as *m-partite*), that is, the vertices can be divided into m disjoint sets such that the vertices in each subset are not adjacent, and $n \ge m(3 + \max_j |I_j|)$, then G is (m, n)-coverable. For example,
  - A tree ($m = 2$) in which the maximal number of children of any vertex is no more than $(n-8)/2$,
  - A circle with p even ($m = 2$) and $n \ge 10$, or with p odd ($m = 3$) and $n \ge 15$,
  - A finite subset of the d-dimensional lattice $\mathbb{Z}^d$ where vertices separated by distance 1 are adjacent ($m = 2$) and $n \ge 6 + 4d$.

For simple graphs such as those listed above, finding appropriate blocking sets B_i can be done by inspection; see the Supplementary Material Appendix B.2.3 [18]. More generally, determining (m, n)-coverability for an arbitrary graph or, given an (m, n)-coverable graph, determining blocking sets B_i's that are optimal in some sense (e.g., minimizing $|\bigcup_{i\le m} B_i|$) are beyond the scope of this work. However, in Algorithm B.5 in the Supplementary Material Appendix B.2 [18], we provide a randomized greedy search for suitable B_i's that be applied in practice when the graph structure is too complex to find such B_i's by inspection.

3.2.2. *Numerical examples*. We present two simulations comparing the power of Algorithm 3.4 with its unconditional counterpart, one a time-varying AR(1) model and the other

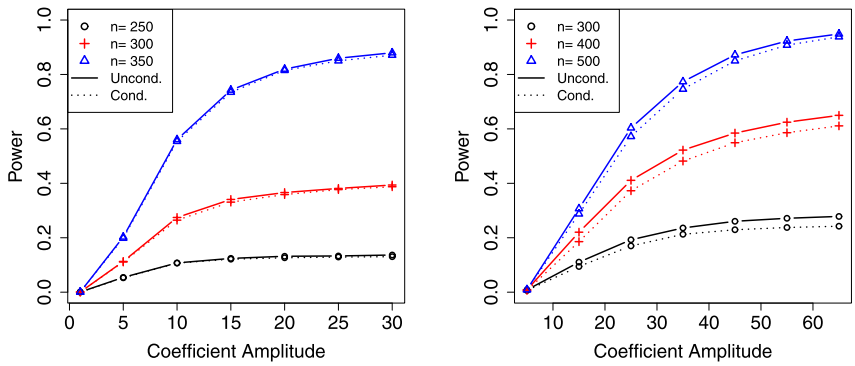


FIG. 2. *Power curves of conditional and unconditional knockoffs for* $p = 2000$ *and a range of* n *for* (a) *an* AR(1) *model and* (b) *an* AR(10) *model. Standard errors are all below* 0.008.

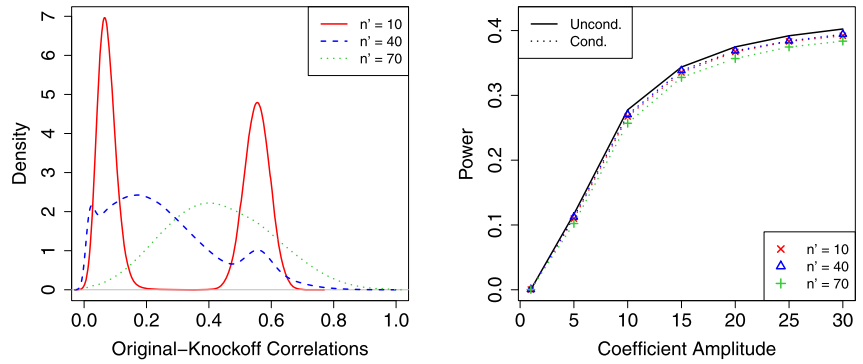
a time-varying AR(10). Line 2 of Algorithm 3.2 uses Algorithm 3.1 with the vector $\boldsymbol{s}$ computed using the SDP method of equation (B.1), and the unconditional construction also uses the SDP method [10]. Algorithm 3.4 was run with $m = 2$ and B_1 and B_2 chosen by fixing n' (specified in the following paragraphs) and running Algorithm 3.3 twice with two different π's. The first run used the original variable ordering for π, and the second run used ordered B_1 followed by the ordered remaining variables. This is a nonrandomized version of Algorithm B.5, which works well for AR models because of their graph structure. We remind the reader that the simulation setting is at the beginning of Section 3.

In Figure 2(a), the $\boldsymbol{x}_i \in \mathbb{R}^{2000}$ are i.i.d. AR(1) with autocorrelation coefficient 0.3 (although the autocorrelation coefficient does not vary with time, this is not assumed by Algorithm 3.4). We chose $n' = 40$, resulting in 210 variables that are each blocked in half the samples. The number of unknown parameters is $3p - 1 = 5999$ while the sample sizes simulated are much smaller, $n \leq 350$, yet the power of conditional knockoffs is nearly indistinguishable from that of unconditional knockoffs which uses the exactly-known distribution of X.

In Figure 2(b), the $\boldsymbol{x}_i \in \mathbb{R}^{2000}$ are time-varying AR(10); specifically, $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the renormalization of $\boldsymbol{\Sigma}^0$ to have 1's on the diagonal, and $(\boldsymbol{\Sigma}^0)^{-1}_{j,k} = 1_{\{j=k\}} - 0.05 \cdot 1_{\{1 \leq |j-k| \leq 10\}}$. We chose $n' = 50$, resulting in 1660 variables that are each blocked in half the samples. The number of unknown parameters is $2p + 10p - 10 \times 11/2 = 23{,}945$ while the sample sizes are again much smaller, $n \leq 500$, and the power difference between conditional and unconditional knockoffs remains very slight.

Note that the simulation in Figure 2(a) blocked on just roughly 10% of its variables (i.e., $|B_1 \cup B_2|/p \approx 10\%$), and since the signals are uniformly distributed, one might worry that in specific applications where the blocked variables and signals happened to align, the power loss might be much worse. But Figure 2(b)'s simulation blocked on over 80% of its variables and still suffered very little power loss compared to unconditional knockoffs, suggesting that even the blocking of signal variables has only a small effect on power thanks to the data splitting in Algorithm 3.4.

Finally, we examine the sensitivity of the power of conditional knockoffs to the choice of n' in Algorithm 3.3 for choosing the B_i. In the case of AR(1) with $n = 300$ and $p = 2000$, Figure 3(a) shows the averaged density of original-knockoff correlations $\tilde{\rho}_j = \boldsymbol{X}_j^\top \tilde{\boldsymbol{X}}_j / (\|\boldsymbol{X}_j\| \|\tilde{\boldsymbol{X}}_j\|)$ for three different choices of n', and Figure 3(b) shows the corresponding power curves. Recall that smaller n' means blocking on more variables but generating better knockoffs for the non-blocked variables in each step i of Algorithm 3.4. Figure 3(a) shows quite different correlation profiles for different n', with $n' = 40$ seeming to provide the



(a) (b)

FIG. 3. *Sensitivity of conditional knockoffs to the choice of n' for an* AR(1) *model with $n = 300$ and $p = 2000$.* (a) *Histograms of the original-knockoff correlations and* (b) *power curves. Standard errors in* (b) *are all below* 0.004.

density with mass most concentrated to the left. Indeed Figure 3(b) shows $n' = 40$ is most powerful, but only by a small margin—the power is quite insensitive to the choice of n'. In applications, the choice of n' may rely on an approximate version of Figure 3(a) obtained by simulating $\boldsymbol{X}$ from an estimated model.

In the Supplementary Material [18], we provide additional experiments that compare the performance of conditional knockoffs that are generated using different sufficient statistics (Appendix D.3) and examine the scenario where a superset of the edge set E is unknown and is instead estimated using the data (Appendix D.4).

3.3. *Discrete graphical model.* We now turn to applying conditional knockoffs to discrete models for X. Such models are used, for example, for survey responses, general binary covariates and single nucleotide polymorphisms (mutation counts at loci along the genome) in genomics. Many discrete models assume some form of local dependence, for instance in time or space. We will show how to construct conditional knockoffs when that local dependence is modeled by (undirected) graphical models (see, e.g., Edwards [14], Chapter 2), for example, Ising models, Potts models and Markov chains.

A random vector X is Markov with respect to a graph $G = ([p], E)$ if for any two disjoint subsets $A, A' \subset [p]$ and a *cut set* $B \subset [p]$ such that every path from A to A' passes through B, it holds that $X_A \perp\!\!\!\perp X_{A'} \mid X_B$. Denote by I_j the vertices adjacent to j in G (excluding j itself). X being Markov implies the *local Markov property* that $X_j \perp\!\!\!\perp X_{(\{j\}\cup I_j)^c} \mid X_{I_j}$.

In this section, we assume X is locally Markov with respect to a known graph G and each variable X_j takes $K_j \geq 2$ discrete values (for simplicity label these values $[K_j] = \{1, \ldots, K_j\}$). Although the algorithms in this section can be applied when K_j is infinite, we assume for simplicity that K_j is finite. Formally, we assume

$$F_X \in \left\{ \text{distribution on } \prod_{j=1}^{p} [K_j] \text{ satisfying the local Markov property w.r.t. } G \right\}. \tag{3.4}$$

3.3.1. *Generating conditional knockoffs by blocking.* Our algorithm for generating conditional knockoffs for discrete graphical models uses again the ideas of blocking and data splitting in Section 3.2. However, unlike Section 3.2 which built upon the low-dimensional construction of Section 3.1, there is no known efficient algorithm for constructing conditional knockoffs for general discrete models in low dimensions. As such, instead of blocking to isolate small graph components, we now block to isolate *individual* vertices, and as such need to be more careful with data splitting to ensure the resulting knockoffs remain powerful.

Suppose B is a cut set such that every path connecting *any* two different vertices in B^c passes through B; call such a set a *global cut set* with respect to G. The local Markov property implies the elements of X_{B^c} are conditionally independent given X_B:

$$\mathbb{P}(X_{B^c} \mid X_B) = \prod_{j \in B^c} \mathbb{P}(X_j \mid X_B) = \prod_{j \in B^c} \mathbb{P}(X_j \mid X_{I_j}),$$

where we used the fact that for any $j \in B^c$, $I_j \subseteq B$ and $X_j \perp\!\!\!\perp X_{B \setminus I_j} \mid X_{I_j}$. For any $A \subseteq [p]$ and $k_1, \ldots, k_p$, denote by $\boldsymbol{k}_A$ the vector of k_j's for $j \in A$ and by $[\boldsymbol{K}_A]$ the Cartesian product $\prod_{j \in A}[K_j]$. Then the conditional probability $\mathbb{P}(X_j \mid X_{I_j})$ can be written as

$$\prod_{k_j \in [K_j], \boldsymbol{k}_{I_j} \in [\boldsymbol{K}_{I_j}]} \theta_j(k_j, \boldsymbol{k}_{I_j})^{1_{\{X_j = k_j, X_{I_j} = \boldsymbol{k}_{I_j}\}}},$$

with parameters $\theta_j(k_j, \boldsymbol{k}_{I_j}) \in [0, 1]$ for all k_j, $\boldsymbol{k}_{I_j}$, with the convention that $0^0 := 1$. Let $\psi_B(X_B)$ be the probability mass function for X_B, the joint distribution for n i.i.d. samples from the graphical model is then

$$\prod_{i=1}^{n} \psi_B(X_{i,B}) \prod_{j \in B^c} \Bigg(\prod_{k_j \in [K_j], \boldsymbol{k}_{I_j} \in [\boldsymbol{K}_{I_j}]} \theta_j(k_j, \boldsymbol{k}_{I_j})^{N_j(k_j, \boldsymbol{k}_{I_j})} \Bigg),$$

where $N_j(k_j, \boldsymbol{k}_{I_j}) = \sum_{i=1}^{n} 1_{\{X_{i,j} = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}}$. Let $T_B(\boldsymbol{X})$ be the statistic that includes $\boldsymbol{X}_B$ and the counts $N_j(k_j, \boldsymbol{k}_{I_j})$ for all $j \in B^c$ and all possible $(k_j, \boldsymbol{k}_{I_j})$. Then $T_B(\boldsymbol{X})$ is a sufficient statistic for model (3.4). Conditional on $T_B(\boldsymbol{X})$, the random vectors $\{\boldsymbol{X}_j, j \in B^c\}$ are independent and each $\boldsymbol{X}_j$ is uniformly distributed on all $\boldsymbol{w} \in [K_j]^n$ such that $\sum_{i=1}^{n} 1_{\{w_i = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}} = N_j(k_j, \boldsymbol{k}_{I_j})$ for any $(k_j, \boldsymbol{k}_{I_j})$. Algorithm 3.5 generates knockoffs conditional on $T_B(\boldsymbol{X})$ by, for each j, uniformly permuting subsets of entries of $\boldsymbol{X}_j$ to produce $\tilde{\boldsymbol{X}}_j$. The subsets of entries are defined by blocks of identical rows of $\boldsymbol{X}_{I_j}$ so that $\sum_{i=1}^{n} 1_{\{\tilde{X}_{i,j} = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}} = N_j(k_j, \boldsymbol{k}_{I_j})$, as required.

The computational complexity of Algorithm 3.5 is (shown in the Supplementary Material Appendix B.3 [18]) $O(\sum_{j \in B^c}(n + \min(\prod_{\ell \in I_j} K_\ell, n|I_j|)))$. If $n > \max_{j \in B^c} \prod_{\ell \in I_j} K_\ell$, as needed to guarantee nontrivial knockoffs for all $j \in B^c$ are generated with positive probability, then the complexity can be simplified to $O(n(p - |B|))$. In general, Algorithm 3.5's computational complexity is bounded by the simple expression $O(np\bar{d})$, where $\bar{d}$ is the average degree in B^c.

THEOREM 3.4. *Algorithm 3.5 generates valid knockoffs for model (3.4).*

Algorithm 3.5 Conditional knockoffs for discrete graphical models

Input: $\boldsymbol{X} \in \mathbb{N}^{n \times p}$, $G = ([p], E)$, $B \in [p]$.
Require: B is a global cut set of G.

```
for j in [p] \ B do
    Initialize X̃_j to X_j.
    for k_{I_j} ∈ [K_{I_j}] do
        Uniformly randomly permute the entries of X̃_j whose corresponding rows of X_{I_j} equal k_{I_j}.
    end for
end for
Set X̃_B = X_B.
return X̃ = [X̃_1, ..., X̃_p].
```

Algorithm 3.6 Conditional knockoffs for discrete graphical models with data splitting

Input: $\boldsymbol{X} \in \mathbb{N}^{n\times p}$, $G = ([p], E)$, $B_1, \dots, B_m \subset [p]$, $n_1, \dots, n_m \in \mathbb{N}$.
Require: $[p] = \bigcup_{i=1}^m B_i^c$ and each B_i is a global cut set.
1: Partition the rows of $\boldsymbol{X}$ into submatrices $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(m)}$ with each $\boldsymbol{X}^{(i)} \in \mathbb{N}^{n_i\times p}$.
2: **for** $i = 1, \dots, m$ **do**
3: Run Algorithm 3.5 or B.6 on $\boldsymbol{X}^{(i)}$ with B_i to obtain $\tilde{\boldsymbol{X}}^{(i)}$.
4: **end for**
5: **return** $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}^{(1)}; \dots; \tilde{\boldsymbol{X}}^{(m)}]$ (row-concatenation of $\tilde{\boldsymbol{X}}^{(i)}$'s).

As with Algorithm 3.2, in Algorithm 3.5 variables in B are blocked and their knockoffs are trivial: $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$. One way to mitigate this drawback is to, after running Algorithm 3.5, expand the graph to include the generated knockoff variables and then conduct a second knockoff generation with the expanded graph. We elaborate on this idea and present Algorithm B.6, a modified version of Algorithm 3.5, in the Supplementary Material Appendix B.4 [18].

Another systematic way to address this issue is to take the same approach as Algorithm 3.4 by splitting the data and running Algorithm 3.5 (or Algorithm B.6) on each split with different B's; see Algorithm 3.6.

If $n_i > \max_{j\in B_i^c} \prod_{\ell\in I_j} K_\ell$ for all $i \le m$ and all the model parameters $\theta_j(k_j, \boldsymbol{k}_{I_j})$ are positive, then Algorithm 3.6 produces nontrivial knockoffs for all j with positive probability. Note that in the continuous case, similar mild conditions guarantee that Algorithm 3.4 produces nontrivial knockoffs for all j with *probability 1*. This is unachievable in general in the discrete case no matter how the sufficient statistic is chosen, as there is always a positive probability (for every j) that the sufficient statistic takes a value such that $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j$ is uniquely determined given that sufficient statistic (e.g., if $\boldsymbol{X}_{i,j} = 1$ for all i).

One way to ensure $B_1, \dots, B_m$ satisfy the requirements of Algorithm 3.6 is if assigning each B_i^c a different color produces a proper coloring of G, that is, no adjacent vertices have the same color. The end of Section 3.2.1 listed some common graph structures with known chromatic numbers (the chromatic number of a graph G is the minimal m such that G is m-colorable), which subsume many common models including Ising models and Potts models. Although not specified in Section 3.2.1, a Markov chain of order $m-1$ is m-colorable and a planar graph (map) is 4-colorable. Also, for any graph of maximal degree d, a $(d+1)$-coloring can be found in $O(dp)$ time by greedy coloring [26], Chapter 2. In general, both finding the chromatic number and finding a corresponding coloring of a graph G are NP-hard [16], but there exist efficient algorithms that in practice are able to color graphs with a near-optimal number of colors (see Malaguti and Toth [28] for a survey).

3.3.2. *Refined constructions for Markov chains.* For Markov chains, we develop two alternative conditional knockoff constructions that take advantage of the Markovian structure. Let $N^{(j)}_{k_{j-1},k_j} = \sum_{i=1}^n 1_{\{X_{i,j-1}=k_{j-1}, X_{i,j}=k_j\}}$. Then all the $N^{(j)}_{k_{j-1},k_j}$'s together form a sufficient statistic, which we denote by $T(\boldsymbol{X})$. As opposed to the statistics $N_j(k_j, \boldsymbol{k}_{\{j-1,j+1\}})$'s used in Section 3.3.1, $T(\boldsymbol{X})$ is minimal, and thus we expect that generating knockoffs conditional on it will be more powerful than knockoffs generated conditional on a non-minimal statistic and will dominate Algorithm 3.6 when G is a Markov chain. However, we found the difference in power to be negligible in every simulation we tried, and so we defer these algorithms to the Supplementary Material Appendix B.4 [18].

3.3.3. *Numerical examples.* We present two simulations, comparing the power of Algorithm 3.6 with its unconditional counterpart for discrete Markov chains [33] and for Ising models [3]. We remind the reader that the simulation setting is at the beginning of Section 3.

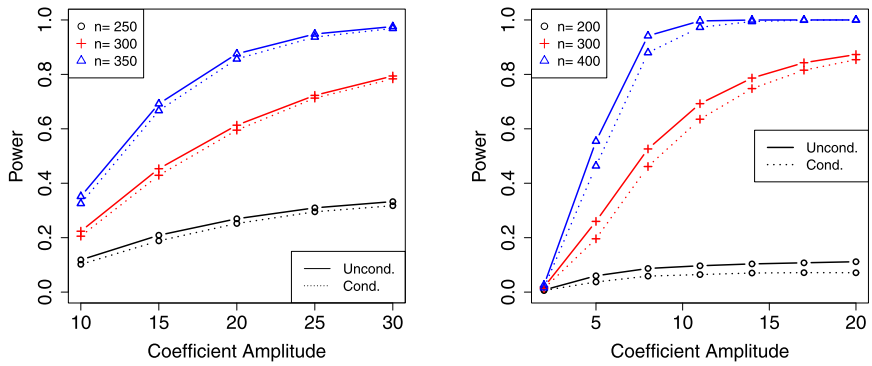


FIG. 4. *Power curves of conditional and unconditional knockoffs with a range of n for* (a) *a Markov chain of length* $p = 1000$ *and* (b) *an Ising model of size* 32×32. *Standard errors are all below* 0.008.

In Figure 4(a), the $\boldsymbol{x}_i \in \{0,1\}^{1000}$ are i.i.d. from an inhomogeneous binary Markov chain. The initial distribution is $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = 0.5$, and the transition probabilities $\mathbb{P}(X_j = 0 \mid X_{j-1} = 1) = Q_{10}^{(j)}$, $\mathbb{P}(X_j = 1 \mid X_{j-1} = 0) = Q_{01}^{(j)}$ are randomly generated as $Q_{10}^{(j)} = U_1^{(j)}/(0.4 + U_1^{(j)} + U_2^{(j)})$ and $Q_{01}^{(j)} = U_3^{(j)}/(0.4 + U_3^{(j)} + U_4^{(j)})$, where $U_i^{(j)} \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$ but held fixed across all replications. We implemented Algorithm 3.6 with B_1 as the even variables and B_2 as the odds, with $n_1 = n_2 = n/2$, and used Algorithm B.6 (with $Q = 2$) in Line 3. The number of unknown parameters in the model is $2p - 1 = 1999$ and all plotted power curves have $n \leq 350$. Despite the high-dimensionality, conditional knockoffs are nearly as powerful as the unconditional SCIP procedure of Sesia et al. [33] which requires knowing the exact distribution of X.

In Figure 4(b), the $\boldsymbol{x}_i \in \{-1,+1\}^{32\times 32}$ are i.i.d. draws from an Ising model given by (see the Supplementary Material Appendix B.3 [18] for the sampling details):

$$\mathbb{P}(X = \boldsymbol{x}) \propto \exp\Bigg(\sum_{(s,t)\in E} \theta_{s,t} x_s x_t + \sum_{s \in V} h_s x_s \Bigg), \quad \boldsymbol{x} \in \{-1,+1\}^V, \tag{3.5}$$

where the vertex set $V = [32] \times [32]$ and the edge set E is all the pairs (s,t) such that $\|s - t\|_1 = 1$. We take $\theta_{s,t} = 0.2$ and $h_s = 0$. Model (3.5) has $2 \times 32 \times 31 + 32^2 = 3008$ parameters, again far larger than any of the sample sizes simulated, yet conditional knockoffs are still nearly as powerful as their unconditional counterparts. The conditional knockoffs are generated by Algorithm 3.6 with two-fold data-splitting ($m = 2$, vertices are colored by the parity of the sum of their coordinates) and no graph-expanding. Although it is possible to use graph-expanding, the power improvement is negligible because the sample size is quite small relative to the size of the neighborhoods in the expanded graph, resulting in the second round of knockoffs being nearly identical to their original counterparts.

4. Discussion. This paper introduced a way to use knockoffs to perform variable selection with exact FDR control under much weaker assumptions than made in Candès et al. [10], while retaining nearly as high power in simulations. In fact, our method controls the FDR under arguably weaker assumptions than *any* existing method (see Section 1.2). The key idea is simple, to generate knockoffs conditional on a sufficient statistic, but finding and proving valid algorithms for doing so required surprisingly sophisticated tools. One particularly appealing property of conditional knockoffs is how it directly leverages unlabeled data

for improved power. We conclude with a number of open research questions raised by this paper.

Algorithmic: Perhaps the most obvious question is how to construct conditional knockoffs for models not addressed in this paper. Even for the models in this paper, what is the best way to choose the tuning parameters (e.g., $\boldsymbol{s}$ in Algorithm 3.1, or the blocks B_i in Algorithms 3.4 and 3.6)?

Robustness: Can techniques like those in Barber et al. [2] be used to quantify the robustness of conditional knockoffs to model misspecification? Empirical evidence for such robustness is provided in the Supplementary Material Appendix D.2 [18]. Also, it is worth pointing out that there are models for which no "small" sufficient statistic exists, that is, every sufficient statistic $T(\boldsymbol{X})$ has the property that $\boldsymbol{X}_j \mid \boldsymbol{X}_{-j}$, $T(\boldsymbol{X})$ is a point mass at $\boldsymbol{X}_j$, which forces the conditional knockoffs $\tilde{\boldsymbol{X}}_j$ to be trivial. In such models where the proposal of this paper can only produce trivial knockoffs, could postulating a distribution and generating knockoffs conditional on *some* (not-sufficient) statistic still improve robustness to the parameter values in the model, relative to generating knockoffs for the same distribution but unconditionally? See Sesia et al. [8] for a positive example for the related conditional randomization test.

Power: In this paper, we always used unconditional knockoffs as a power benchmark for conditional knockoffs, as it seems intuitive that conditioning on less should result in higher power. Can this be formalized, and/or can the cost of conditioning in terms of power be quantified? Combining this with the previous paragraph, we expect there to be a *power–robustness tradeoff* that can be navigated by conditioning on more or less when generating knockoffs.

Conditioning: There are reasons other than robustness that one might wish to generate knockoffs conditional on a statistic. For instance, if a model for $\boldsymbol{X}$ needs to be checked by observing a statistic of $\boldsymbol{X}$, generating knockoffs conditional on that statistic would guarantee a form of post-selection inference after model selection. Or when data contains variables that confound the variables of interest, it may be desirable to generate knockoffs conditional on those confounders (e.g., by Algorithm B.2) in order to control for them. Also, can the conditioning tools and ideas in this paper be used to relax the assumptions of the conditional randomization test, generalizing Rosenbaum [32]?

Acknowledgments. D. H. would like to thank Yu Zhao for advice on topological measure theory. L. J. would like to thank Emmanuel Candès, Rina Barber, Natesh Pillai, Pierre Jacob and Joe Blitzstein for helpful discussions regarding this project. The authors also thank the Editors and the three referees for their constructive comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Relaxing the assumptions of knockoffs by conditioning" (DOI: 10.1214/19-AOS1920SUPP; .pdf). Appendix A proves the theoretical results in the paper. Appendix B provides the algorithmic details. Appendix C discusses the hypotheses being tested and Appendix D provides additional simulations.

REFERENCES

[1] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

[2] BARBER, R. F., CANDÈS, E. J. and SAMWORTH, R. J. (2020). Robust inference with knockoffs. *Ann. Statist.* **48** 1409–1431.

[3] BATES, S., CANDÈS, E., JANSON, L. and WANG, W. (2020). Metropolized knockoff sampling. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1729163
[4] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044
[5] BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097 https://doi.org/10.1093/biomet/asu056
[6] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392
[7] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998
[8] BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2018). The conditional permutation test. Preprint. Available at arXiv:1807.05405.
[9] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9
[10] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265
[11] CHATTERJEE, A. and LAHIRI, S. N. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41** 1232–1259. MR3113809 https://doi.org/10.1214/13-AOS1106
[12] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Ann. Rev. Econ.* **7** 649–688. https://doi.org/10.1146/annurev-economics-012315-015826
[13] DIACONIS, P., HOLMES, S. and SHAHSHAHANI, M. (2013). Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications*: *A Festschrift in Honor of Morris L. Eaton*. *Inst. Math. Stat.* (*IMS*) *Collect.* **10** 102–125. IMS, Beachwood, OH. MR3586941
[14] EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1880319 https://doi.org/10.1007/978-1-4612-0493-0
[15] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.
[16] GAREY, M. R. and JOHNSON, D. S. (1979). *Computers and Intractability*: *A Guide to the Theory of NP-Completeness*. W. H. Freeman and Co., San Francisco, CA. MR0519066
[17] GIMENEZ, J. R., GHORBANI, A. and ZOU, J. (2019). Knockoffs for the mass: New feature importance statistics with false discovery guarantees. In *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 2125–2133.
[18] HUANG, D. and JANSON, L. (2020). Supplement to "Relaxing the assumptions of knockoffs by conditioning." https://doi.org/10.1214/19-AOS1920SUPP.
[19] HUBER, P. J. (1972). The 1972 Wald lecture. Robust statistics: A review. *Ann. Math. Stat.* **43** 1041–1067. MR0314180 https://doi.org/10.1214/aoms/1177692459
[20] JANSON, L. (2017). A model-free approach to high-dimensional inference. Ph.D. thesis, Stanford Univ.
[21] JANSON, L. and SU, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975. MR3486422 https://doi.org/10.1214/16-EJS1129
[22] JAVANMARD, A. and JAVADI, H. (2019). False discovery rate control via debiased lasso. *Electron. J. Stat.* **13** 1212–1253. MR3935848 https://doi.org/10.1214/19-ejs1554
[23] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[24] 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*.
[25] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371
[26] LEWIS, R. M. R. (2016). *A Guide to Graph Colouring*: *Algorithms and Applications*. Springer, Cham. MR3408102 https://doi.org/10.1007/978-3-319-25730-3
[27] LIU, Y. and ZHENG, C. (2019). Deep latent variable models for generating knockoffs. *Stat* **8** e260. MR4072122 https://doi.org/10.1002/sta4.260
[28] MALAGUTI, E. and TOTH, P. (2010). A survey on vertex coloring problems. *Int. Trans. Oper. Res.* **17** 1–34. MR2598219 https://doi.org/10.1111/j.1475-3995.2009.00696.x

[29] NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

[30] PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499 https://doi.org/10.1214/aos/1176349744

[31] ROMANO, Y., SESIA, M. and CANDÈS, E. J. (2019). Deep knockoffs. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2019.1660174

[32] ROSENBAUM, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. MR0763575

[33] SESIA, M., SABATTI, C. and CANDÈS, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. MR3912377 https://doi.org/10.1093/biomet/asy033

[34] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[35] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[36] ZHU, Y. and BRADIC, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319

[37] ZHU, Y. and BRADIC, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electron. J. Stat.* **12** 3312–3364. MR3861831 https://doi.org/10.1214/18-EJS1443