Relaxing the Assumptions of Knockoffs by Conditioning

Dongming Huang and Lucas Janson

Department of Statistics, Harvard University

Abstract

The recent paper Candès et al. (2018) introduced model-X knockoffs, a method for variable selection that provably and non-asymptotically 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.

Keywords. High-dimensional inference, knockoffs, model-X, sufficient statistic, false discovery rate (FDR), topological measure, graphical model

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. (2018), under very mild conditions the complement of the set of unimportant variables, i.e., the *important* (or *non-null*) 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 (Candès et al., 2018).

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 probabilistically controlling the number of false discoveries. In this paper, as with most others in the knockoffs literature,[1] we consider the false discovery rate (FDR) (Benjamini and Hochberg, 1995), defined for a (random) selected subset of variables $\hat{S}$ as

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

i.e., 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. (2018), referred to therein as the 'model-X' approach, assumes the observations $(Y_i, X_{i,1}, \ldots, X_{i,p}) \overset{i.i.d.}{\sim} F_{Y,X}$ 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, i.e., $(Y, X_1, \ldots, X_p) \in \{1, \ldots, k\}^{p+1}$, 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

[1] Janson and Su (2016) show how the last step of knockoffs can easily be modified to control other error rates such as the k-familywise error rate.

importance of robustness, assumption-checking, and domain knowledge for justifying the resulting inference; see Janson (2017, Chapter 1) for a detailed discussion of the role of fixed-X and model-X[1] 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 Candès et al. (2018) 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 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.

	Model for $F_{Y\|X}$	Model for F_X
Fixed-X	$o(n)$ parameters[2,3]	arbitrary
Model-X (Candès et al., 2018)	arbitrary	0 parameters
Model-X (this paper)	arbitrary	$\Omega(n^*p)$ parameters

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 similar complexity (exponentially large in p).

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 $p/n \to 0$ [Huber1973, Portnoy1985]. Note that the commonly studied problem of inference for a single parameter can generally be translated to FDR control using the Benjamini–Hochberg (Benjamini and Hochberg, 1995) or Benjamini–Yekutieli (Benjamini and Yekutieli, 2001) procedures (see, e.g., Javanmard and Javadi (2018)), so that it makes sense to compare such inference with our paper that is focused on multiple testing. In high dimensions, i.e., when $p > n$, even

[1]Therein referred to as 'model-based' and 'model-free', respectively.

[2]In the exceptional case of Gaussian linear regression, n parameters are allowed.

[3]Except for Gaussian linear regression, fixed-X inferential guarantees are only asymptotic.

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 (2013)). 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 (2014); Javanmard and Montanari (2014); van de Geer et al. (2014)) and the extended score statistic (see, e.g., Belloni et al. (2014, 2015); Chernozhukov et al. (2015); Ning and Liu (2017)), 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, Zhu and Bradic (2018) and Zhu et al. (2018) 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. (2014); Lee et al. (2016) for foundational contributions, and Candès et al. (2018, 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 (2015) for low-dimensional homoscedastic linear regression with fixed design. The model-X knockoffs framework proposed by Candès et al. (2018) 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 (Candès et al., 2018), a Markov chain or hidden Markov models (Sesia et al., 2018), a graphical model (Bates et al., 2019), and certain latent variable models (Gimenez et al., 2018). In the case that F_X is only known approximately, the robustness of model-X knockoffs is studied by Barber et al. (2018). When F_X is completely unknown some recent works have proposed methods to generate approximate knockoffs (Jordon et al., 2019; Romano et al., 2018; Liu and Zheng, 2018) 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 (1984) 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.[1] A different conditional permutation test was recently proposed by Berrett et al. (2018) 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 (Candès et al., 2018), namely, that $X_j \mid X_{-j}$ 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 $(Y_i, X_{i,1}, \ldots, X_{i,p}) \in \mathbb{R}^{p+1}$ for $i = 1, \ldots, n$. We then stack these vectors into a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ whose ith row is denoted by $\boldsymbol{x}_i^\top = (X_{i,1}, \ldots, X_{i,p}) \in \mathbb{R}^p$, and a column vector $\boldsymbol{y} \in \mathbb{R}^n$ whose ith entry is Y_i. We are about to define model-X knockoffs $(\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,p})$, and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$ will analogously denote these row vectors stacked to form a knockoff design matrix. A square bracket around matrices, such as $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$, denotes the horizontal concatenation of these matrices. We use $[p]$ for $\{1, 2, \ldots, p\}$, and $i : j$ for $\{i, i+1, \ldots, j\}$ for any $i \leq j$; 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 (Candès et al., 2018). 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 $A \subseteq [p]$

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

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

[1] See the paragraph preceding Rosenbaum (1984, Theorem 1) for a description of the test's limitations.

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 $\tilde{Z}_j$, i.e., 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. (2018) 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_+:

$$T_0 = \min\left\{t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}, \qquad T_+ = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}.$$

Candès et al. (2018, Theorem 3.4) proves that $\hat{S}$ with T_+ exactly (non-asymptotically) controls the FDR at level q, and that $\hat{S}$ with T_0 exactly controls a modified FDR, $\mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| + 1/q}\right]$, 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 $A \subseteq \mathcal{H}_0$,

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

which is proved in Candès et al. (2018, 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. (2018) 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 $A \subseteq [p]$,

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

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. (2018)'s 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. (2018) 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. (2018).

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*

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

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 $\boldsymbol{X} \,|\, T(\boldsymbol{X})$ 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 $A \subseteq [p]$,

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

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 more generally 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**:

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

   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**:

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

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

3. **Discrete graphical model**:

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

   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 (non-stationary) 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 $\sqrt{\frac{n-2}{2}} - 1 + (p-1)\left(\sqrt{\frac{n-2}{2}}\right)\left(\sqrt{\frac{n-2}{2}} - 1\right) = \Omega(np)$. 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 $\mathcal{H}_0 = \emptyset$ 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 Appendix C.

[1] Here we assume $n/8 \leq p$.

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 $\{\boldsymbol{x}_i^{(u)}\}_{i=1}^{n^{(u)}}$, i.e., 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 $A \subseteq [p]$,*

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

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 $\tilde{\boldsymbol{X}}^* \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{X}^*$ 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 Appendix A. 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 non-zero 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 Appendix D.2. 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 Appendix D.1. 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 $= \mathbb{E}\left[\frac{|S \cap \hat{S}|}{|S|}\right]$) 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

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

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 1.

In Algorithm 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 (2015, Equation (1.4)). Indeed nearly the same tools can be used to find a suitable $\boldsymbol{s}$; in Appendix B.1 we slightly adapt three methods from Barber and Candès (2015) and Candès et al. (2018) for computing suitable $\boldsymbol{s}$. The computational complexity of Algorithm 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

Algorithm 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 $n\left(2\operatorname{diag}\{\boldsymbol{s}\} - \operatorname{diag}\{\boldsymbol{s}\}\hat{\boldsymbol{\Sigma}}^{-1}\operatorname{diag}\{\boldsymbol{s}\}\right)$ 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

$$\tilde{\boldsymbol{X}} = \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}^{-1}\operatorname{diag}\{\boldsymbol{s}\}) + \boldsymbol{U}\boldsymbol{L}. \tag{3.2}$$

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

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 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 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 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. (2013)), we bypass this approach by directly using the concept of invariant measures from topological measure theory; see Appendix A.2.2.

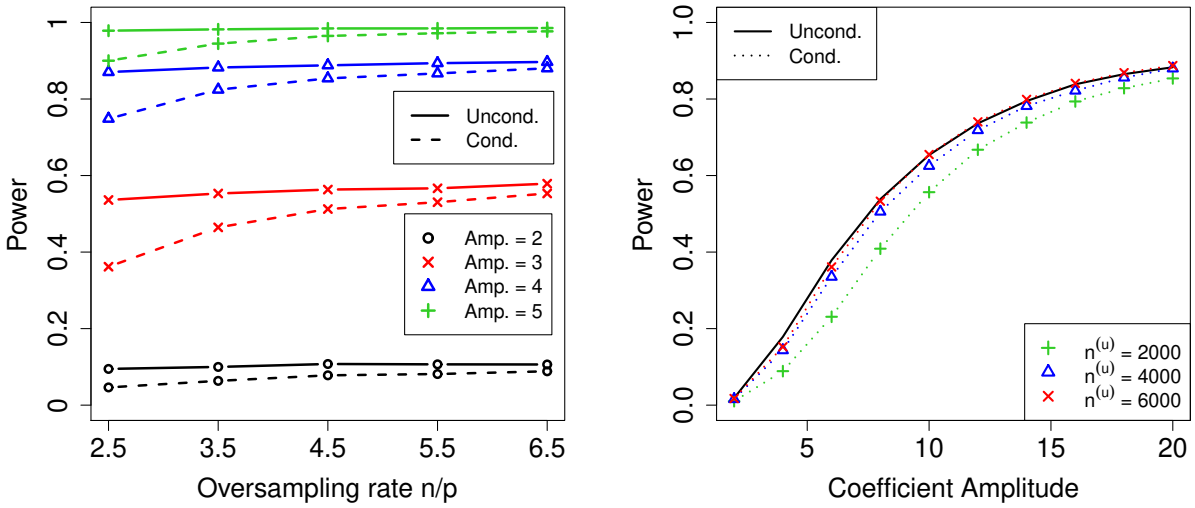
A useful consequence of Theorem 3.1 is the double robustness property that if knockoffs are constructed by Algorithm 1 and knockoff statistics are used which obey the sufficiency property of Barber and Candès (2015) (that is, the knockoff statistics only depend on $\boldsymbol{y}$ and $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ 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:

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

In Appendix B.1 we extend Algorithm 1 to the case when the mean is known (Algorithm 7) or a subset of columns of $\boldsymbol{X}$ are additionally conditioned on (Algorithm 8). 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.

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 1 is computed using the SDP method of Equation (B.1), and the analogous vector for the unconditional construction is chosen by the analogous SDP method (Candès et al., 2018). 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.



(a) (b)

Figure 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.

Figure 1a 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 1b 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 $\mathbf{\Sigma}^{-1}$, 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

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

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 $\hat{\mathbf{\Sigma}}_E$ together constitute a sufficient statistic, where $\hat{\mathbf{\Sigma}}_E := \left\{\hat{\mathbf{\Sigma}}_{j,k} : j = k \text{ or } (j,k) \in E\right\}$. 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 (Bühlmann and van de Geer, 2011, Chapter 13) and then treating it as known. Note that we only require the edge set E to contain all non-zero entries of $\mathbf{\Sigma}^{-1}$, which is no harder than the exact identification of the non-zero 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 $V_1, \ldots, V_\ell$. Then X can be divided into independent subvectors, $X_{V_1}, \ldots, X_{V_\ell}$, 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 $\hat{\mathbf{\Sigma}}_E$. 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 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 $V_1, \ldots, V_\ell$

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.

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 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 $V_1, \ldots, V_\ell$.

1: **for** $k = 1, \ldots, \ell$ **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 8 (a slight modification of Algorithm 1).

3: **end for**

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

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

Algorithm 2 constructs knockoffs for the model (3.3) by first conditioning on $\boldsymbol{X}_B$ and then running a slight modification of Algorithm 1 (Algorithm 8 in Appendix B.1.3) on the variables/columns V_k corresponding to the induced subgraphs. The computational complexity of Algorithm 2 is $O\left(n \sum_{k=1}^{\ell} \left(|I_{V_k} \cap B|^2 |V_k| + |V_k|^2\right)\right)$, which is upper-bounded by $O\left(\ell n n'^2 + np \max_{k \in [\ell]} |I_{V_k} \cap B|^2\right)$ (both complexities assume the most efficient construction of $\boldsymbol{s}$ is used as a primitive in Algorithm 8).

Theorem 3.2. *Algorithm 2 generates valid knockoffs for model* (3.3).

Algorithm 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 2 will have no power to select any of the variables in B.

Algorithm 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 with input (π, n'), then G is n-separated by B for any $n \geq n'$.*

Algorithm 3 is meant to be intuitive but a more efficient implementation is given in Appendix B.2. Algorithm 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

Algorithm 3 Greedy Search for a Blocking Set

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

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

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 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.

An intuitive solution to prevent the trivial knockoffs $\tilde{\boldsymbol{X}}_B$ in Algorithm 2 is to split the rows of $\boldsymbol{X}$ in half and run Algorithm 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 4 to splitting the rows of $\boldsymbol{X}$ into m folds and running Algorithm 2 on each fold with a different input B.

Algorithm 4 Conditional Knockoffs for Gaussian Graphical Models with Data Splitting

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $G = ([p], E)$, $B_1, \ldots, B_m \subset [p]$, $n_1, \ldots, n_m \in \mathbb{N}$

Require: $\bigcup_{i=1}^{m} B_i^c = [p]$, G is n_i-separated by B_i for all $i = 1, \ldots, m$, and $\sum_{i=1}^{m} n_i = n$.

Partition the rows of $\boldsymbol{X}$ into submatrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(m)}$ with each $\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i \times p}$.
for $i = 1, \ldots, m$ **do**
  Run Algorithm 2 on $\boldsymbol{X}^{(i)}$ with blocking set B_i to obtain $\tilde{\boldsymbol{X}}^{(i)}$.
end for
return $\tilde{\boldsymbol{X}} = \left[\tilde{\boldsymbol{X}}^{(1)}; \ldots; \tilde{\boldsymbol{X}}^{(m)}\right]$ (the row-concatenation of the $\tilde{\boldsymbol{X}}^{(i)}$'s).

In Algorithm 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 4 into a definition.

Definition 3.2. $G = ([p], E)$ is (m, n)-*coverable* if there exist $B_1, \ldots, B_m$ subsets of $[p]$ and integers

$n_1 \ldots, n_m$ such that $\bigcup_{i=1}^{m} B_i^c = [p]$, G is n_i-separated by B_i for all $i = 1, \ldots, m$, and $\sum_{i=1}^{m} n_i \leq n$.

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-inhomogenous AR(r) model), i.e., $E = \{(i, j) : 1 \leq |i - j| \leq r\}$, and $n \geq 2 + 8r$, then G is $(2, n)$-coverable.
- If G is a *m-colorable* (also known as *m-partite*), i.e., the vertices can be divided into m disjoint sets such that the vertices in each subset are not adjacent, and $n \geq m(3 + \max_j |I_j|)$, 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 $n \geq 10$, or with p odd ($m = 3$) and $n \geq 15$,
  - A finite subset of the d-dimensional lattice $\mathbb{Z}^d$ where vertices separated by distance 1 are adjacent ($m = 2$) and $n \geq 6 + 4d$.

For simple graphs such as those listed above, finding appropriate blocking sets B_i can be done by inspection; see Appendix B.2.3. 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 $\left| \bigcup_{i \leq m} B_i \right|$) are beyond the scope of this work. However, in Algorithm 11 in Appendix B.2, 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.

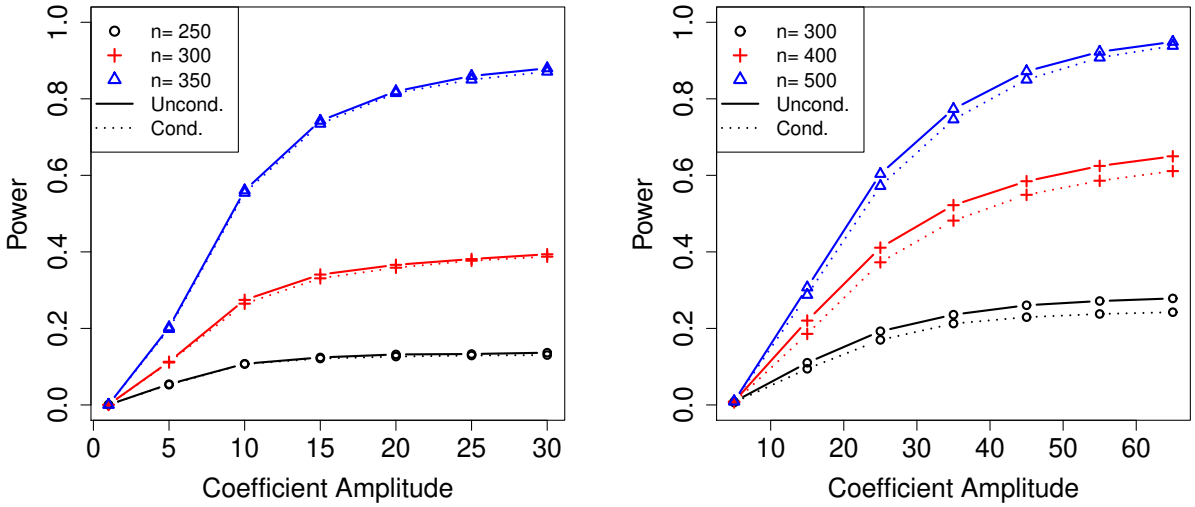
3.2.2 Numerical Examples

We present two simulations comparing the power of Algorithm 4 with its unconditional counterpart, one a time-varying AR(1) model and the other a time-varying AR(10). Line 2 of Algorithm 2 uses Algorithm 1 with the vector $\boldsymbol{s}$ computed using the SDP method of Equation (B.1), and the unconditional construction also uses the SDP method (Candès et al., 2018). Algorithm 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 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.[1] We remind the reader that the simulation setting is at the beginning of Section 3.

In Figure 2a, 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 4). We chose $n' = 40$, resulting in 210 variables that are each blocked in half the samples. The number of unknown parameters is $3p - 1 = 5,999$ 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.

[1] This is a nonrandomized version of Algorithm 11, which works well for AR models because of their graph structure.

In Figure 2b, the $\boldsymbol{x}_i \in \mathbb{R}^{2000}$ are time-varying AR(10); specifically, $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the renormalization of $\boldsymbol{\Sigma}^0$ to have 1's on the diagonal, and $(\boldsymbol{\Sigma}^0)^{-1}_{j,k} = \mathbf{1}_{\{j=k\}} - 0.05 \cdot \mathbf{1}_{\{1 \le |j-k| \le 10\}}$. We chose $n' = 50$, resulting in $1,660$ 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, $n \le 500$, and the power difference between conditional and unconditional knockoffs remains very slight.



(a) (b)

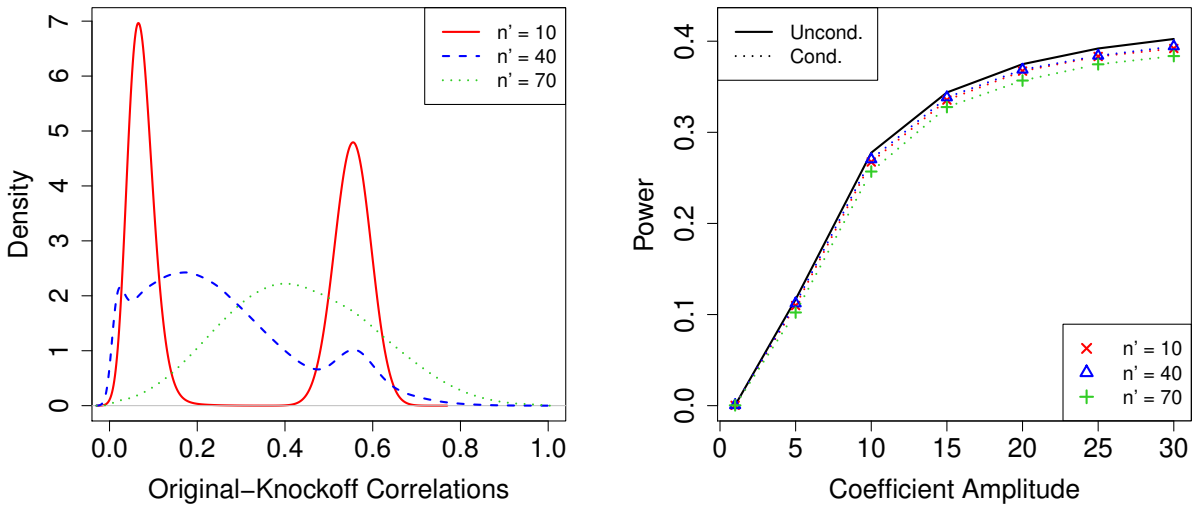
Figure 2: Power curves of conditional and unconditional knockoffs for $p = 2000$ and a range of n for (a) an AR(1) model and (b) an AR(10) model. Standard errors are all below 0.008.

Note that the simulation in Figure 2a 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 2b'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 4.

Finally, we examine the sensitivity of the power of conditional knockoffs to the choice of n' in Algorithm 3 for choosing the B_i. In the case of AR(1) with $n = 300$ and $p = 2000$, Figure 3a shows the averaged density[1] 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 3b 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 4. Figure 3a shows quite different correlation profiles for different n', with $n' = 40$ seeming to provide the density with mass most concentrated to the left. Indeed Figure 3b 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 3a obtained by simulating $\boldsymbol{X}$ from an estimated model.

In Appendix D, we provide additional experiments that compare the performance of conditional

[1] 3200 independent simulations were averaged and the kernel density estimate used a Gaussian kernel with a bandwidth of 0.01.



(a) (b)

Figure 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.

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 (2000, 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:

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

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 $\mathbb{P}\left(X_j \,\middle|\, X_{I_j}\right)$ can be written as

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

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

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

where $N_j(k_j, \boldsymbol{k}_{I_j}) = \sum_{i=1}^{n} \mathbf{1}_{\{X_{i,j} = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}}$. 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 $\sum_{i=1}^{n} \mathbf{1}_{\{w_i = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}} = N_j(k_j, \boldsymbol{k}_{I_j})$ for any $(k_j, \boldsymbol{k}_{I_j})$. Algorithm 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 $\sum_{i=1}^{n} \mathbf{1}_{\{\tilde{\boldsymbol{X}}_{i,j} = k_j, \boldsymbol{X}_{i,I_j} = \boldsymbol{k}_{I_j}\}} = N_j(k_j, \boldsymbol{k}_{I_j})$, as required.

The computational complexity of Algorithm 5 is $O\left(\sum_{j \in B^c} (n + \min(\prod_{\ell \in I_j} K_\ell, n|I_j|))\right)$, which is shown in Appendix B.3. 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 $O\left(n(p - |B|)\right)$. In general, Algorithm 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 5 generates valid knockoffs for model* (3.4).

Algorithm 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.
1: **for** j in $[p] \setminus B$ **do**
2: Initialize $\tilde{\boldsymbol{X}}_j$ to $\boldsymbol{X}_j$.
3: **for** $\boldsymbol{k}_{I_j} \in [\boldsymbol{K}_{I_j}]$ **do**
4: Uniformly randomly permute the entries of $\tilde{\boldsymbol{X}}_j$ whose corresponding rows of $\boldsymbol{X}_{I_j}$ equal $\boldsymbol{k}_{I_j}$.
5: **end for**
6: **end for**
7: Set $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$.
8: **return** $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_p]$.

As with Algorithm 2, in Algorithm 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 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 12, a modified version of Algorithm 5, in Appendix B.4.

Another systematic way to address this issue is to take the same approach as Algorithm 4 by splitting the data and running Algorithm 5 (or Algorithm 12) on each split with different B's; see Algorithm 6.

Algorithm 6 Conditional Knockoffs for Discrete Graphical Models with Data Splitting

Input: $\boldsymbol{X} \in \mathbb{N}^{n \times p}$, $G = ([p], E)$, $B_1, \ldots, B_m \subset [p]$, $n_1, \ldots, n_m \in \mathbb{N}$.
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 $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(m)}$ with each $\boldsymbol{X}^{(i)} \in \mathbb{N}^{n_i \times p}$.
2: **for** $i = 1, \ldots, m$ **do**
3: Run Algorithm 5 or 12 on $\boldsymbol{X}^{(i)}$ with B_i to obtain $\tilde{\boldsymbol{X}}^{(i)}$.
4: **end for**
5: **return** $\tilde{\boldsymbol{X}} = \left[\tilde{\boldsymbol{X}}^{(1)}; \ldots; \tilde{\boldsymbol{X}}^{(m)}\right]$ (row-concatenation of $\tilde{\boldsymbol{X}}^{(i)}$'s).

If $n_i > \max_{j \in B_i^c} \prod_{\ell \in I_j} K_\ell$ for all $i \leq m$ and all the model parameters $\theta_j(k_j, \boldsymbol{k}_{I_j})$ are positive, then Algorithm 6 produces nontrivial knockoffs for all j with positive probability. Note that in the continuous case, similarly mild conditions guarantee that Algorithm 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 $B_1, \ldots, B_m$ satisfy the requirements of Algorithm 6 is if assigning each B_i^c a different color produces a proper coloring of G.[1] The end of Section 3.2.1 listed some common graph structures with known chromatic numbers,[2] which subsume many common models including

[1] A coloring of G is *proper* if no adjacent vertices have the same color.
[2] The chromatic number of a graph G is the minimal m such that G is m-colorable.

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 (Lewis, 2016, Chapter 2). In general, both finding the chromatic number and finding a corresponding coloring of a graph G are NP-hard (Garey and Johnson, 1979), but there exist efficient algorithms that in practice are able to color graphs with a near-optimal number of colors (see Malaguti and Toth (2010) 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. Although we generally expect these constructions to dominate Algorithm 6 when G is a Markov chain, we found the difference in power to be negligible in every simulation we tried, and so we defer these algorithms to Appendix B.4 and only provide a brief summary here.

Suppose the components of X follow a K-state discrete Markov chain, and let $\pi_k^{(1)} = \mathbb{P}(X_1 = k)$ and $\pi_{k,k'}^{(j)} = \mathbb{P}(X_j = k' \mid X_{j-1} = k)$ be the model parameters. Then the joint distribution for n i.i.d. samples is,

$$\mathbb{P}(\boldsymbol{X}) = \prod_{k=1}^{K} (\pi_k^{(1)})^{\sum_{k'=1}^{K} N_{k,k'}^{(2)}} \prod_{j=2}^{p} \prod_{k=1}^{K} \prod_{k'=1}^{K} (\pi_{k,k'}^{(j)})^{N_{k,k'}^{(j)}},$$

where $N_{k,k'}^{(j)} = \sum_{i=1}^{n} \mathbf{1}_{\{X_{i,j-1}=k, X_{i,j}=k'\}}$. So all the $N_{k,k'}^{(j)}$'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. Conditional on $T(\boldsymbol{X})$, the columns of $\boldsymbol{X}$ still comprise a Markov chain whose distribution can be used to generate knockoffs in two possible ways:

1. *The sequential conditional independent pairs (SCIP)* algorithm (Candès et al., 2018; Sesia et al., 2018) has computational complexity exponential in n, but by splitting the samples into small folds and generating conditional knockoffs separately for each fold, n is artificially reduced and the computation made tractable.

2. *Refined blocking* modifies Algorithm 5 by first drawing a new contingency table that is exchangeable with the the three-way contingency table for $(\boldsymbol{X}_{j-1}, \boldsymbol{X}_j, \boldsymbol{X}_{j+1})$ and then sampling $\tilde{\boldsymbol{X}}_j$ given the new contingency table.

3.3.3 Numerical Examples

We present two simulations, comparing the power of Algorithm 6 with its unconditional counterpart for discrete Markov chains (Sesia et al., 2018) and for Ising models (Bates et al., 2019). We remind the reader that the simulation setting is at the beginning of Section 3.

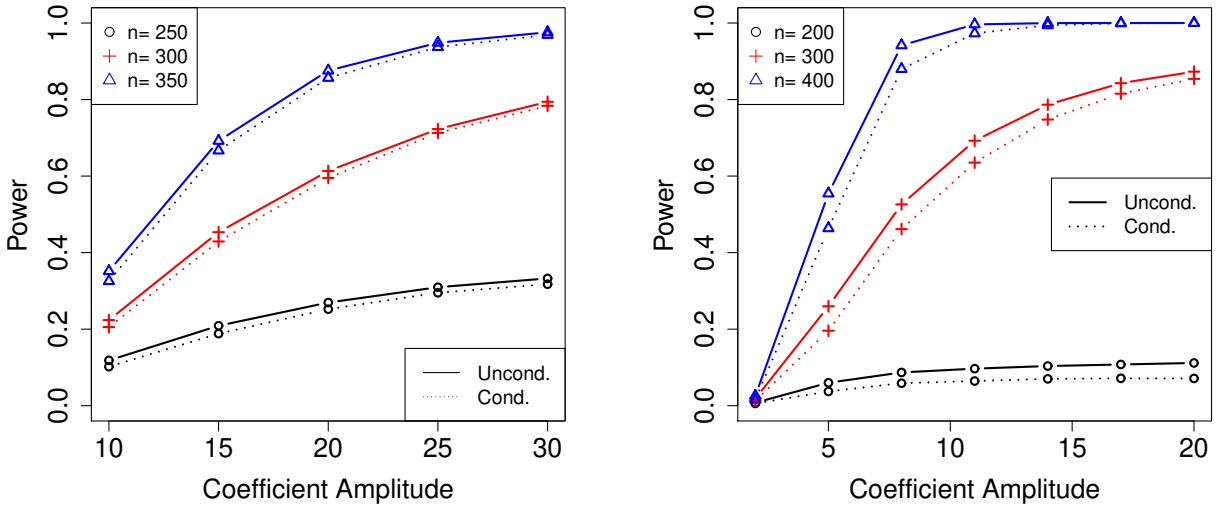
In Figure 4a, the $\boldsymbol{x}_i \in \{0,1\}^{1000}$ are i.i.d. from an inhomogeneous binary Markov chain with $p = 1000$. The initial distribution is $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = .5$, and the transition probabilities

$$\mathbb{P}(X_j = 0 | X_{j-1} = 1) = Q_{10}^{(j)}, \quad \mathbb{P}(X_j = 1 | X_{j-1} = 0) = Q_{01}^{(j)}$$

are randomly generated as

$$Q_{10}^{(j)} = \frac{U_1^{(j)}}{0.4 + U_1^{(j)} + U_2^{(j)}}, \qquad Q_{01}^{(j)} = \frac{U_3^{(j)}}{0.4 + U_3^{(j)} + U_4^{(j)}},$$

where $U_i^{(j)} \overset{i.i.d.}{\sim} \text{Unif}([0,1])$ but held fixed across all replications. We implemented Algorithm 6 with B_1 as the even variables and B_2 as the odds, with $n_1 = n_2 = n/2$, and used Algorithm 12 (with $Q = 2$) in Line 3. The number of unknown parameters in the model is $2p-1 = 1,999$ 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. (2018) which requires knowing the exact distribution of X.



(a) (b)

Figure 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 4b, the $\boldsymbol{x}_i \in \mathbb{R}^{32\times 32}$ are i.i.d. draws from an Ising model[1] given by:

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

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.[2] The conditional knockoffs are generated by Algorithm 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.

[1] We use the *coupling from the past algorithm* (Propp and Wilson, 1996) to sample exactly from this distribution.
[2] We use the default subgraph width $w = 5$ in Bates et al. (2019) for generating unconditional knockoffs.

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. (2018), 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 1, or the blocks B_i in Algorithms 4 and 6)?

Robustness: Can techniques like those in Barber et al. (2018) be used to quantify the robustness of conditional knockoffs to model misspecification? Empirical evidence for such robustness is provided in Appendix D.2. Also, it is worth pointing out that there are models for which no 'small' sufficient statistic exists, i.e., every sufficient statistic $T(\boldsymbol{X})$ has the property that $\boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})$ 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 Berrett et al. (2018) 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 8) 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 (1984)?

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.

References

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

Barber, R. F., Candès, E. J., and Samworth, R. J. (2018). Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*.

Bates, S., Candès, E., Janson, L., and Wang, W. (2019). Metropolized knockoff sampling. *arXiv preprint arXiv:1903.00434*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

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 (Methodological)*, 57(1):pp. 289–300.

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

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2018). The conditional permutation test. *arXiv preprint arXiv:1807.05405*.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Candès, E., 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(3):551–577.

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(3):1232–1259.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688.

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*, pages 102–125. Institute of Mathematical Statistics.

Diaconis, P., Sturmfels, B., et al. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of statistics*, 26(1):363–397.

Eaton, M. L. (1983). *Multivariate statistics: a vector space approach*. Wiley New York.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer Science & Business Media.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Fremlin, D. H. (2003). *Measure theory*, volume 4. Torres Fremlin.

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

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to NP-completeness*. W.H. Freeman.

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

Janson, L. (2017). *A Model-Free Approach to High-Dimensional Inference*. PhD thesis, Stanford University.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Statist.*, 10(1):960–975.

Javanmard, A. and Javadi, H. (2018). False Discovery Rate Control via Debiased Lasso. *arXiv.org*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

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*.

Kollo, T. and von Rosen, D. (2006). *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lewis, R. (2016). *A Guide to Graph Colouring: Algorithms and Applications*. Springer.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440.

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

Malaguti, E. and Toth, P. (2010). A survey on vertex coloring problems. *International transactions in operational research*, 17(1):1–34.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, 45(1):158–195.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252.

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

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Schneider, R. and Weil, W. (2008). *Stochastic and integral geometry*. Springer Science & Business Media.

Sesia, M., Sabatti, C., and Candès, E. J. (2018). Gene hunting with knockoffs for hidden markov models. *Biometrika*, 106(1):1–18.

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(3):1166–1202.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhu, Y. and Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 0(0):1–18.

Zhu, Y., Bradic, J., et al. (2018). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364.

A Proofs for Main Text

A.1 Integration of Unlabled Data

Proof of Proposition 2.3. Denote by $\boldsymbol{X}^{(u)}$ the last $n^{(u)} = n^* - n$ rows of $\boldsymbol{X}^*$. Since the rows of $\boldsymbol{X}^*$ are independent, $\boldsymbol{X}^{(u)} \perp\!\!\!\perp (\boldsymbol{y}, \boldsymbol{X})$. Then by the weak union property, $\boldsymbol{X}^{(u)} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{X}$. In addition, the condition that $\tilde{\boldsymbol{X}}^* \perp\!\!\!\perp \boldsymbol{y} \,|\, (\boldsymbol{X}, \boldsymbol{X}^{(u)})$ and the fact that $\tilde{\boldsymbol{X}}$ is a function of $\tilde{\boldsymbol{X}}^*$ imply $\tilde{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{y} \,|\, (\boldsymbol{X}, \boldsymbol{X}^{(u)})$. By the contraction property, these two together show $\tilde{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{X}$.

Let $\phi : \mathbb{R}^{n^* \times 2p} \mapsto \mathbb{R}^{n \times 2p}$ be the mapping that keeps the first n rows of a matrix. We have $[\boldsymbol{X},\, \tilde{\boldsymbol{X}}] = \phi([\boldsymbol{X}^*,\, \tilde{\boldsymbol{X}}^*])$ and $[\boldsymbol{X},\, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(A)} = \phi([\boldsymbol{X}^*,\, \tilde{\boldsymbol{X}}^*]_{\mathrm{swap}(A)})$ for any subset $A \subseteq [p]$. The given exchangeability condition implies that

$$\phi([\boldsymbol{X}^*,\, \tilde{\boldsymbol{X}}^*]_{\mathrm{swap}(A)}) \overset{\mathcal{D}}{=} \phi([\boldsymbol{X}^*,\, \tilde{\boldsymbol{X}}^*]) \,\Big|\, T(\boldsymbol{X}^*),$$

which is simply $[\boldsymbol{X},\, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X},\, \tilde{\boldsymbol{X}}] \,\Big|\, T(\boldsymbol{X}^*)$. It then follows that

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

and we conclude that $\tilde{\boldsymbol{X}}$ is a model-X knockoff matrix for $\boldsymbol{X}$. □

A.2 Low-Dimensional Gaussian Models

Throughout the appendix, bold-faced capital letters such as $\boldsymbol{A}$ are used for any matrix (random or not) *except* when we need to distinguish between a random matrix and the values it may take, in which case we use bold sans serif letters for the values. For example, we will write $\mathbb{P}(\boldsymbol{A} = \mathbf{A})$ to denote the probability that the random matrix $\boldsymbol{A}$ takes the (nonrandom) value $\mathbf{A}$.

This section is planned as follows. Section A.2.1 clarifies a difficulty in the joint uniform distribution on a manifold. Section A.2.2 contains the proof of Theorem 3.1, leaving the proofs of the lemmas in Section A.2.3. Section A.2.4 discusses why a seemingly simpler proof for the theorem fails and thus justifies our technical contribution.

A.2.1 Counterexample for Conditional Uniformity

The following statement is false: 'If a random variable A is uniform on its support and another random variable B is such that $B \mid A$ is conditionally uniform on its support for every A, with normalizing constant that does not depend on A, then (A, B) is uniform on its support.' Although this statement seems intuitively true and holds for many simple examples (especially when A and B are both univariate), Figure 5 shows a counterexample. In it, although X is uniform on $(0, 1)$ and $(Y, Z) \mid X$ is uniform for every X on a line whose length does not depend on X, the joint distribution of (X, Y, Z) is not uniform on its 2-dimensional support.

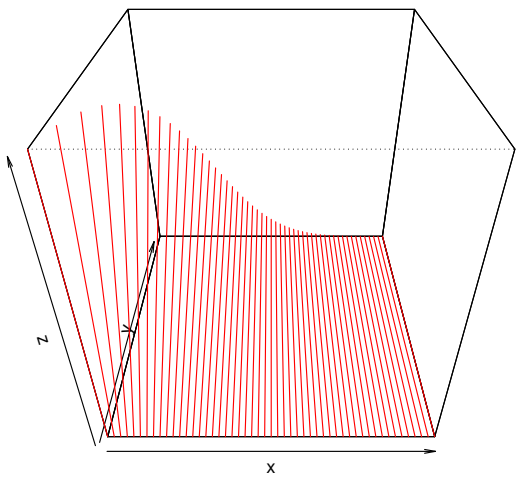


Figure 5: A non-uniform distribution on a surface. $X \sim \text{Unif}(0, 1)$ and $(Y, Z) \mid X \sim \text{Unif}(L_X)$, where the line segment L_X of length 1 is orthogonal to the X-axis and has an angle with the Y-axis of $(1 - X)^{10}\pi/2$.

A.2.2 Proof of Theorem 3.1

The proof of Theorem 3.1 follows three steps: Lemma A.1 states 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, Lemma A.2 states that the conditional

distribution remains invariant (on the same support) after swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$, and Lemma A.3 states 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 $\boldsymbol{X}$.

To streamline notation, we redefine $\hat{\boldsymbol{\Sigma}} := (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)^\top(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)$ as n times the sample covariance matrix (it was defined as just the sample covariance matrix in the main text), and redefine $\boldsymbol{s}$ such that $\mathbf{0}_{p\times p} \prec \mathrm{diag}\,\{\boldsymbol{s}\} \prec 2\hat{\boldsymbol{\Sigma}}$ accordingly. With this new notation, $\boldsymbol{L}$ is the Cholesky decomposition such that $\boldsymbol{L}^\top\boldsymbol{L} = 2\mathrm{diag}\,\{\boldsymbol{s}\} - \mathrm{diag}\,\{\boldsymbol{s}\}\hat{\boldsymbol{\Sigma}}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}$. Let $\boldsymbol{C} \in \mathbb{R}^{(n-1)\times n}$ be a matrix with orthonormal rows that are also orthogonal to $\mathbf{1}_n$. Then $\boldsymbol{C}^\top\boldsymbol{C} = \boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n$ is the centering matrix, $\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{X} = \boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top$ and $(\boldsymbol{C}\boldsymbol{X})^\top\boldsymbol{C}\boldsymbol{X} = \hat{\boldsymbol{\Sigma}}$; note $\boldsymbol{C}$ is just a constant, nonrandom matrix. The statistic being conditioned on this this proof is $T(\boldsymbol{X}) = (\boldsymbol{X}^\top\mathbf{1}_n/n, (\boldsymbol{C}\boldsymbol{X})^\top\boldsymbol{C}\boldsymbol{X}) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. For any positive integers s and t such that $s \geq t$, denote by $\mathcal{O}_s$ the group of $s \times s$ orthogonal matrices (also known as the orthogonal group) and denote by $\mathcal{F}_{s,t}$ the set of $s \times t$ real matrices whose columns form an orthonormal set in $\mathbb{R}^s$ (also known as the Stiefel manifold).

We will use techniques from topological measure theory to prove Theorem 3.1, specifically on invariant measures (see e.g. Schneider and Weil (2008, Chapter 13) and (Fremlin, 2003, Chapter 44)). For readers unfamiliar with the field, the following is a short list of definitions we will use:

- A group $\mathcal{G}$ is a *topological group* if it has a topology such that the functions of multiplication and inversion, i.e., $(x, y) \mapsto xy$ and $x \mapsto x^{-1}$, are continuous.[1]
- An *operation* of a group $\mathcal{G}$ on a nonempty set $\mathcal{M}$ is a function $\psi : \mathcal{G} \times \mathcal{M} \mapsto \mathcal{M}$ satisfying $\psi(g, \psi(g', x)) = \psi(gg', x)$ and $\psi(e, x) = x$. The operation $\psi(g, x)$ is also written as gx when there is no risk of confusion. For any subset $\mathcal{B} \subseteq \mathcal{M}$ and $g \in \mathcal{G}$, denote by $g\mathcal{B}$ the image under the operation with g, i.e., $g\mathcal{B} = \{\psi(g, x) : x \in \mathcal{B}\}$.
- An operation ψ is *transitive* if for any $x, y \in \mathcal{M}$ there exists $g \in \mathcal{G}$ such that $\psi(g, x) = y$.
- Suppose $\mathcal{M}$ is a topological space and $\mathcal{G}$ is a topological group, the operation ψ is *continuous* if ψ, as a function of two arguments, is continuous.
- Suppose $\mathcal{M}$ is a locally compact metric space. A Borel measure ρ on $\mathcal{M}$ is called $\mathcal{G}$-*invariant* if for any $g \in \mathcal{G}$ and Borel subset $\mathcal{B} \subseteq \mathcal{M}$, it holds that $\rho(\mathcal{B}) = \rho(g\mathcal{B})$.

We can now define the elements of the proof. Suppose $\boldsymbol{S} \in \mathbb{R}^{p\times p}$ is a positive definite matrix and $\boldsymbol{m} \in \mathbb{R}^p$. Define a metric space

$$\begin{aligned}\mathcal{M} = \Big\{[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n\times(2p)} : \mathbf{X}^\top\mathbf{1}_n/n = \boldsymbol{m},\ (\boldsymbol{C}\mathbf{X})^\top\boldsymbol{C}\mathbf{X} = \boldsymbol{S},\\ \tilde{\mathbf{X}}^\top\mathbf{1}_n/n = \boldsymbol{m},\ (\boldsymbol{C}\tilde{\mathbf{X}})^\top\boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S},\ (\boldsymbol{C}\tilde{\mathbf{X}})^\top\boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\Big\},\end{aligned} \tag{A.1}$$

equipped with the Euclidean metric in the vectorized space, stacked column-wise. By Equation (3.2), it is straightforward to check that if $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{m}, \boldsymbol{S})$ then $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \in \mathcal{M}$.

Define $\mathcal{G} = \{\boldsymbol{G} \in \mathcal{O}_n : \boldsymbol{G}\mathbf{1}_n = \mathbf{1}_n\}$. It is easy to check that $\mathcal{G}$ is a group whose identity element is $\boldsymbol{I}_n$. $\mathcal{G}$ is also a topological group with the induced metric of $\mathbb{R}^{n\times n}$ because matrix inversion and multiplication are continuous.

[1] A function between two topological spaces is *continuous* if the inverse image of any open set is an open set.

Define a mapping $\psi : \mathcal{G} \times \mathcal{M} \mapsto \mathcal{M}$ by $\psi(\boldsymbol{G}, [\mathbf{X}, \tilde{\mathbf{X}}]) = [\boldsymbol{G}\mathbf{X}, \boldsymbol{G}\tilde{\mathbf{X}}]$. Note that

$$\boldsymbol{G}^\top \boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{G} = \boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{G}^\top (\mathbf{1}_n \mathbf{1}_n^\top / n)\boldsymbol{G} = \boldsymbol{I}_n - (\mathbf{1}_n \mathbf{1}_n^\top / n) = \boldsymbol{C}^\top \boldsymbol{C}, \tag{A.2}$$

thus for any $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M}$, we have $\psi(\boldsymbol{G}, [\mathbf{X}, \tilde{\mathbf{X}}]) \in \mathcal{M}$. It is also seen that $\psi(\boldsymbol{G}_1\boldsymbol{G}_2, [\mathbf{X}, \tilde{\mathbf{X}}]) = \psi(\boldsymbol{G}_1, \psi(\boldsymbol{G}_2, [\mathbf{X}, \tilde{\mathbf{X}}]))$ and $\psi(\boldsymbol{I}_n, [\mathbf{X}, \tilde{\mathbf{X}}]) = [\mathbf{X}, \tilde{\mathbf{X}}]$, so ψ is an operation of $\mathcal{G}$ on $\mathcal{M}$. By the continuity of matrix multiplication, ψ is a continuous operation.

We can now state the three lemmas which comprise the proof.

Lemma A.1 (Invariance)**.** *The probability measure of* $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ *conditional on* $\hat{\boldsymbol{\mu}} = \boldsymbol{m}$ *and* $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ *is* $\mathcal{G}$*-invariant on* $\mathcal{M}$*.*

Lemma A.2 (Invariance after swapping)**.** *The probability measure of* $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}(j)}$ *conditional on* $\hat{\boldsymbol{\mu}} = \boldsymbol{m}$ *and* $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ *is* $\mathcal{G}$*-invariant on* $\mathcal{M}$*.*

Lemma A.3 (Uniqueness)**.** *The* $\mathcal{G}$*-invariant probability measure on* $\mathcal{M}$ *is unique.*

Combining Lemmas A.1, A.2 and A.3 together, we conclude that given $\hat{\boldsymbol{\mu}} = \boldsymbol{m}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, swapping $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ leaves the distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ unchanged. Since if swapping one column does not change the distribution, then by induction swapping any set of columns will not change the distribution and this completes the proof.

Remark 4. Although not shown here, one can define the uniform distribution on $\mathcal{M}$ via the Hausdorff measure and show that it is also $\mathcal{G}$-invariant. Therefore, by the uniqueness of the invariant measure, $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ is distributed uniformly on $\mathcal{M}$.

A.2.3 Proofs of Lemmas

Before proving the lemmas, we introduce some notation and properties for Gaussian matrices. Let r, s, and t be any positive integers. For any matrix $\boldsymbol{A} \in \mathbb{R}^{s \times t}$, denote by $\text{vec}(\boldsymbol{A})$ the vector that concatenates its columns, i.e., $(\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_t^\top)^\top$. Denote by $\otimes$ the Kronecker product. A $s \times t$ random matrix $\boldsymbol{A}$ is a Gaussian random matrix $\boldsymbol{A} \sim \mathcal{N}_{s,t}(\boldsymbol{M}, \boldsymbol{\Upsilon} \otimes \boldsymbol{\Sigma})$ if $\text{vec}(\boldsymbol{A}^\top) \sim \mathcal{N}(\text{vec}(\boldsymbol{M}^\top), \boldsymbol{\Upsilon} \otimes \boldsymbol{\Sigma})$ for some $\boldsymbol{M} \in \mathbb{R}^{s \times t}$ and matrices $\boldsymbol{\Upsilon} \succeq \mathbf{0}_{s \times s}$ and $\boldsymbol{\Sigma} \succeq \mathbf{0}_{t \times t}$.

If $\boldsymbol{A} \sim \mathcal{N}_{s,t}(\boldsymbol{M}, \boldsymbol{\Upsilon} \otimes \boldsymbol{\Sigma})$, then for any matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times s}$, $\text{vec}((\boldsymbol{\Gamma}\boldsymbol{A})^\top) = (\boldsymbol{\Gamma} \otimes \boldsymbol{I}_t)\,\text{vec}(\boldsymbol{A}^\top)$ is still multivariate Gaussian and

$$\boldsymbol{\Gamma}\boldsymbol{A} \sim \mathcal{N}_{r,t}(\boldsymbol{\Gamma}\boldsymbol{M}, (\boldsymbol{\Gamma}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}^\top) \otimes \boldsymbol{\Sigma}),$$

because $(\boldsymbol{\Gamma} \otimes \boldsymbol{I}_t)(\boldsymbol{\Upsilon} \otimes \boldsymbol{\Sigma})(\boldsymbol{\Gamma} \otimes \boldsymbol{I}_t)^\top = (\boldsymbol{\Gamma}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}^\top) \otimes (\boldsymbol{I}_t\boldsymbol{\Sigma}\boldsymbol{I}_t)$ by the mixed-product property and transpose of Kronecker product. When the rows of $\boldsymbol{A}$ are i.i.d. samples from a multivariate Gaussian, $\boldsymbol{\Upsilon} = \boldsymbol{I}_s$ and $\boldsymbol{M} = \mathbf{1}_s \boldsymbol{\mu}^\top$ for some $\boldsymbol{\mu} \in \mathbb{R}^t$. If further, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{I}_r$, then

$$\boldsymbol{\Gamma}\boldsymbol{A} \sim \mathcal{N}_{r,t}(\boldsymbol{\Gamma}\mathbf{1}_s\boldsymbol{\mu}^\top, \boldsymbol{I}_r \otimes \boldsymbol{\Sigma}). \tag{A.3}$$

We write the Gram–Schmidt orthonormalization as a function $\Psi(\cdot)$. We will make use of the property that for any $\boldsymbol{\Gamma}_0 \in \mathcal{O}_s$ and any matrix $\boldsymbol{U}_0 \in \mathbb{R}^{s \times t}$ (for $s \geq t$), it holds that

$$\Psi(\boldsymbol{\Gamma}_0\boldsymbol{U}_0) = \boldsymbol{\Gamma}_0\Psi(\boldsymbol{U}_0). \tag{A.4}$$

See, e.g., Eaton (1983, Proposition 7.2).

Proof of Lemma A.1. Define $\nu(\mathcal{B}) := \mathbb{P}\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}] \in \mathcal{B} \,\middle|\, \hat{\boldsymbol{\mu}} = \boldsymbol{m}, \hat{\boldsymbol{\Sigma}} = \boldsymbol{S}\right)$ for any Borel subset $\mathcal{B} \subseteq \mathcal{M}$. For fixed $\boldsymbol{G} \in \mathcal{G}$, we need to show the group operation given $\boldsymbol{G}$, i.e., $g_{\boldsymbol{G}} = \psi(\boldsymbol{G}, \cdot)$, leaves ν unchanged. Define $\boldsymbol{X}' = \boldsymbol{G}\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}' = \boldsymbol{G}\tilde{\boldsymbol{X}}$. We will show

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \overset{\mathcal{D}}{=} [\boldsymbol{X}', \tilde{\boldsymbol{X}}'] \mid T(\boldsymbol{X}).$$

By Equation (A.2) and $\boldsymbol{G}\mathbf{1}_n = \mathbf{1}_n$, we have $T(\boldsymbol{G}\mathbf{X}) = T(\mathbf{X})$ for any $\mathbf{X} \in \mathbb{R}^{n\times p}$. Applying the property in Equation (A.3), we have

$$\boldsymbol{G}\boldsymbol{X} \sim \mathcal{N}_{n,p}(\mathbf{1}_n\boldsymbol{\mu}^\top, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}),$$

where we have used $\boldsymbol{G}\mathbf{1}_n = \mathbf{1}_n$ and $\boldsymbol{G} \in \mathcal{O}_n$. Thus $\boldsymbol{X}' \overset{\mathcal{D}}{=} \boldsymbol{X}$. By Equation (A.4) and the definition of $[\boldsymbol{Q}, \boldsymbol{U}]$ in Algorithm 1,

$$\Psi([\mathbf{1}_n, \boldsymbol{X}', \boldsymbol{G}\boldsymbol{W}]) = \boldsymbol{G}\Psi(\boldsymbol{G}^\top[\mathbf{1}_n, \boldsymbol{X}', \boldsymbol{G}\boldsymbol{W}]) = \boldsymbol{G}\Psi([\mathbf{1}_n, \boldsymbol{X}, \boldsymbol{W}]) = [\boldsymbol{G}\boldsymbol{Q}, \boldsymbol{G}\boldsymbol{U}]. \tag{A.5}$$

Let $\boldsymbol{U}' = \boldsymbol{G}\boldsymbol{U}$. Since $\boldsymbol{W}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{G}\boldsymbol{W}$ has the same distribution as $\boldsymbol{W}$, we have $(\boldsymbol{X}, \boldsymbol{W}) \overset{\mathcal{D}}{=} (\boldsymbol{X}', \boldsymbol{G}\boldsymbol{W})$. This together with Equation (A.5) implies $(\boldsymbol{X}, \boldsymbol{U}) \overset{\mathcal{D}}{=} (\boldsymbol{X}', \boldsymbol{U}')$. Hence

$$\mathbb{P}\left(\boldsymbol{X}, \boldsymbol{U} \mid T(\boldsymbol{X})\right) = \mathbb{P}\left(\boldsymbol{X}', \boldsymbol{U}' \mid T(\boldsymbol{X}')\right)$$

and since $T(\boldsymbol{X}') = T(\boldsymbol{X})$, we conclude

$$(\boldsymbol{X}, \boldsymbol{U}) \overset{\mathcal{D}}{=} (\boldsymbol{X}', \boldsymbol{U}') \mid T(\boldsymbol{X}).$$

Now recall we are conditioning on $T(\boldsymbol{X}) = (\boldsymbol{m}, \boldsymbol{S})$, and thus also $\boldsymbol{s}$ and $\boldsymbol{L}$. By Equation (3.2) and the definition of $\tilde{\boldsymbol{X}}'$,

$$\begin{aligned}
\tilde{\boldsymbol{X}}' =&\ \boldsymbol{G}\left(\mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{U}\boldsymbol{L}\right) && \text{(A.6)}\\
=&\ \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top + (\boldsymbol{X}' - \mathbf{1}_n\hat{\boldsymbol{\mu}}^\top)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{U}'\boldsymbol{L}, && \text{(A.7)}
\end{aligned}$$

which would be the knockoff generated by Algorithm 1 if $\boldsymbol{X}'$ was observed. As a consequence,

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \overset{\mathcal{D}}{=} (\boldsymbol{X}', \tilde{\boldsymbol{X}}') \mid T(\boldsymbol{X}).$$

This shows that for any Borel subset $\mathcal{B} \subseteq \mathcal{M}$, $\nu(\mathcal{B}) = \nu(g_{\boldsymbol{G}^{-1}}(\mathcal{B}))$. We conclude that for any $\boldsymbol{G} \in \mathcal{G}$ and any Borel subset $\mathcal{B} \subseteq \mathcal{M}$

$$\nu(g_{\boldsymbol{G}}(\mathcal{B})) = \nu(\mathcal{B}),$$

that is, the conditional probability measure of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ given $T(\boldsymbol{X}) = (\boldsymbol{m}, \boldsymbol{S})$ is $\mathcal{G}$-invariant. □

Proof of Lemma A.2. Without loss of generality, we take $j = 1$. Define a mapping $\phi : \mathbb{R}^{n\times(2p)} \mapsto \mathbb{R}^{n\times(2p)}$ by $\phi([\mathbf{X}, \tilde{\mathbf{X}}]) = [[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}], [\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]]$, i.e., replacing $\mathbf{X}$ and $\tilde{\mathbf{X}}$ with $[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}]$ and $[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]$, respectively. It is easy to see that ϕ is isometric and $\phi^{-1} = \phi$. Furthermore, we will prove that ϕ is a bijective mapping of $\mathcal{M}$ to itself (Lemma A.4). The conditional distribution of $\phi([\boldsymbol{X}, \tilde{\boldsymbol{X}}])$ is the measure ν_ϕ on $\mathcal{M}$ such that $\nu_\phi(\mathcal{B}) = \nu(\phi^{-1}(\mathcal{B}))$, for any Borel subset $\mathcal{B} \subseteq \mathcal{M}$. We will show that ν_ϕ is $\mathcal{G}$-invariant on $\mathcal{M}$ (Lemma A.5).

Lemma A.4. *ϕ is a bijective mapping of $\mathcal{M}$ to itself.*

Proof. ϕ is easily seen to be injective, and to show surjectivity, we will first show $\phi(\mathcal{M}) \subseteq \mathcal{M}$. Combining this with $\phi^{-1} = \phi$ gives $\mathcal{M} \subseteq \phi^{-1}(\mathcal{M}) = \phi(\mathcal{M})$, and thus $\phi(\mathcal{M}) = \mathcal{M}$ so ϕ is surjective from $\mathcal{M}$ to $\mathcal{M}$. We now complete the proof by showing something even stronger than $\phi(\mathcal{M}) \subseteq \mathcal{M}$, namely the equivalence $\phi([\mathbf{X}, \tilde{\mathbf{X}}]) \in \mathcal{M} \iff [\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M}$.

Translating this equivalence to an equality of indicator functions, we need to show that

$$\begin{aligned}
&\mathbf{1}_{\{\mathbf{X}^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{(\boldsymbol{C}\mathbf{X})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S}\}} \mathbf{1}_{\{\tilde{\mathbf{X}}^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}} \\
= \; &\mathbf{1}_{\{[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}]^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{(\boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S}\}} \\
&\cdot \mathbf{1}_{\{[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}] = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}},
\end{aligned}$$

where the righthand side is the same as the lefthand side but with $\mathbf{X}$ and $\tilde{\mathbf{X}}$ replaced with $[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}]$ and $[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]$, respectively. First note that for the first and third indicator functions on the lefthand side,

$$\mathbf{1}_{\{\mathbf{X}^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{\tilde{\mathbf{X}}^\top \mathbf{1}_n/n = \boldsymbol{m}\}} = \left(\prod_{j=1}^{p} \mathbf{1}_{\{\mathbf{X}_j^\top \mathbf{1}_n/n = m_j\}} \right) \left(\prod_{j=1}^{p} \mathbf{1}_{\{\tilde{\mathbf{X}}_j^\top \mathbf{1}_n/n = m_j\}} \right)$$

and exchanging the first term in each product and compressing the products each back into single indicator functions gives $\mathbf{1}_{\{[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}]^\top \mathbf{1}_n/n = \boldsymbol{m}\}} \mathbf{1}_{\{[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]^\top \mathbf{1}_n/n = \boldsymbol{m}\}}$, so it just remains to show that

$$\begin{aligned}
&\mathbf{1}_{\{(\boldsymbol{C}\mathbf{X})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}} \\
= \; &\mathbf{1}_{\{(\boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}] = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}}.
\end{aligned}$$

Again it is useful to rewrite the three indicator functions as products:

$$\begin{aligned}
&\mathbf{1}_{\{(\boldsymbol{C}\mathbf{X})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}} \\
= \; &\left(\prod_{1 \le j \le k \le p} \mathbf{1}_{\{(\boldsymbol{C}\mathbf{X}_j)^\top \boldsymbol{C}\mathbf{X}_k = \boldsymbol{S}_{j,k}\}} \right) \left(\prod_{1 \le j \le k \le p} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}}_j)^\top \boldsymbol{C}\tilde{\mathbf{X}}_k = \boldsymbol{S}_{j,k}\}} \right) \left(\prod_{j,k=1}^{p} \mathbf{1}_{\{(\boldsymbol{C}\tilde{\mathbf{X}}_j)^\top \boldsymbol{C}\mathbf{X}_k = \boldsymbol{S}_{j,k} - 1_{\{j=k\}} s_j\}} \right).
\end{aligned}$$

Now if we exchange the terms in the first product with $k > j = 1$ with the same terms in the third product, and exchange the terms in the second product with $k > j = 1$ with the terms in the third product with $j > k = 1$, we can compress the products each back into single indicator functions again to get $\mathbf{1}_{\{(\boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}] = \boldsymbol{S}\}} \mathbf{1}_{\{(\boldsymbol{C}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}])^\top \boldsymbol{C}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}] = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\}}$. We conclude that $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M} \iff \phi([\mathbf{X}, \tilde{\mathbf{X}}]) \in \mathcal{M}$. □

Lemma A.5. *ν_ϕ is $\mathcal{G}$-invariant on $\mathcal{M}$.*

Proof. For any $\boldsymbol{G} \in \mathcal{G}$, the group operation $g_{\boldsymbol{G}} = \psi(\boldsymbol{G}, \cdot)$ is exchangeable with ϕ because

$$\begin{aligned}
\psi(\boldsymbol{G}, \phi([\mathbf{X}, \tilde{\mathbf{X}}])) &= [\boldsymbol{G}[\tilde{\mathbf{X}}_1, \mathbf{X}_{\text{-}1}], \boldsymbol{G}[\mathbf{X}_1, \tilde{\mathbf{X}}_{\text{-}1}]] \\
&= [[\boldsymbol{G}\tilde{\mathbf{X}}_1, \boldsymbol{G}\mathbf{X}_{\text{-}1}], [\boldsymbol{G}\mathbf{X}_1, \boldsymbol{G}\tilde{\mathbf{X}}_{\text{-}1}]] \\
&= \phi([\boldsymbol{G}\mathbf{X}, \boldsymbol{G}\tilde{\mathbf{X}}])). \\
&= \phi(\psi(\boldsymbol{G}, [\mathbf{X}, \tilde{\mathbf{X}}])).
\end{aligned}$$

Thus for any Borel subset $\mathcal{B} \subseteq \mathcal{M}$,

$$\begin{aligned}\nu_\phi(g_{\boldsymbol{G}}\mathcal{B}) &= \nu(\phi(g_{\boldsymbol{G}}\mathcal{B})) \\ &= \nu(g_{\boldsymbol{G}}(\phi(\mathcal{B}))) \\ &= \nu(\phi(\mathcal{B})) \\ &= \nu_\phi(\mathcal{B}),\end{aligned}$$

where the third equality follows from Lemma A.1. Thus we conclude that ν_ϕ is $\mathcal{G}$-invariant. □

□

Proof of Lemma A.3. Before the proof, we list a few results that will be used.

Fact 1. For an operation ψ of a group $\mathcal{G}$ on a space $\mathcal{M}$, if there is some $z \in \mathcal{M}$ such that for any $y \in \mathcal{M}$ there exists $g_y \in \mathcal{G}$ such that $\psi(g_y, z) = y$, then ψ is transitive. This is because for any $x, y \in \mathcal{M}$, $\psi(g_x^{-1}, x) = \psi(g_x^{-1}, \psi(g_x, z)) = \psi(g_x^{-1} g_x, z) = z$ and $\psi(g_y g_x^{-1}, x) = \psi(g_y, \psi(g_x^{-1}, x)) = \psi(g_y, z) = y$.

Fact 2. For any compact Hausdorff[1] topological group $\mathcal{G}$, there exists a finite Borel measure ν, called a *Haar measure*, such that for any $g \in \mathcal{G}$ and Borel subset $\mathcal{B} \subseteq \mathcal{G}$, $\nu(\mathcal{B}) = \nu(g\mathcal{B}) = \nu(\mathcal{B}g)$ (Fremlin, 2003, 441E, 442I(c)). As an example, the orthogonal group $\mathcal{O}_n$ has a Haar measure (Eaton, 1983, Chapter 6.2).

The key theorem we use is the following.

Lemma A.6 (Theorem 13.1.5 in Schneider and Weil (2008))**.** *Suppose that the compact group $\mathcal{G}$ operates continuously and transitively on the Hausdorff space $\mathcal{M}$ and that $\mathcal{G}$ and $\mathcal{M}$ have countable bases. Let ν be a Haar measure on $\mathcal{G}$ with $\nu(\mathcal{G}) = 1$. Then there exists a unique $\mathcal{G}$-invariant Borel measure ρ on $\mathcal{M}$ with $\rho(\mathcal{M}) = 1$.*

Now we are ready to prove Lemma A.3. Note $\mathcal{G}$ and $\mathcal{M}$ are compact subspaces of the vectorized spaces, and Fact 2 ensures the existence of a Haar measure on $\mathcal{G}$. Since ψ is continuous, as long as ψ is transitive we can apply Lemma A.6 and conclude that the $\mathcal{G}$-invariant probability measure on $\mathcal{M}$ is unique.

To show ψ is transitive by Fact 1, we first fix a point $[\mathbf{X}_0, \tilde{\mathbf{X}}_0]$ and then show for any $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M}$, we can find $\boldsymbol{G} \in \mathcal{G}$ such that $\psi(\boldsymbol{G}, [\mathbf{X}_0, \tilde{\mathbf{X}}_0]) = [\mathbf{X}, \tilde{\mathbf{X}}]$.

Part 1. We begin with representing $\tilde{\mathbf{X}}$ using the Stiefel Manifold. Define $\mathcal{M}_1 = \{\mathbf{X} \in \mathbb{R}^{n\times p} : \mathbf{X}^\top \mathbf{1}_n/n = \boldsymbol{m},\ (\boldsymbol{C}\mathbf{X})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S}\} = \{\mathbf{X} \in \mathbb{R}^{n\times p} : T(\mathbf{X}) = (\boldsymbol{m}, \boldsymbol{S})\}$. For any $\mathbf{X} \in \mathcal{M}_1$, define

$$\mathcal{M}_{\mathbf{X}} = \left\{\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p} : \tilde{\mathbf{X}}^\top \mathbf{1}_n/n = \boldsymbol{m},\ (\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S},\ (\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}\right\}.$$

Let $\boldsymbol{Z}_{\mathbf{X}}$ be a $n \times (n-1-p)$ matrix whose columns form an orthonormal basis for the orthogonal complement of span($[\mathbf{1}_n, \mathbf{X}]$). Recall that $\mathcal{F}_{n-1-p,p}$ is the set of $(n-1-p)\times p$ real matrices whose columns form an orthonormal set in $\mathbb{R}^{n-1-p}$. Define $\varphi_{\mathbf{X}} : \mathcal{F}_{n-1-p,p} \mapsto \mathbb{R}^{n\times p}$ by

$$\varphi_{\mathbf{X}}(\mathbf{V}) = \mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}.$$

[1] A topological space is *Hausdorff* if every two different points can be separated by two disjoint open sets.

The following result tells us that for any $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M}$, there exists a $\mathbf{V} \in \mathcal{F}_{n-1-p,p}$ such that $\tilde{\mathbf{X}} = \varphi_{\mathbf{X}}(\mathbf{V})$, and thus we are implicitly decomposing $\boldsymbol{U}$ from Algorithm 1 into $\boldsymbol{Z_X V}$ for some random $\boldsymbol{V}$, and we think of $\mathbf{V}$ as a realization of this $\boldsymbol{V}$.

Lemma A.7. *$\varphi_{\mathbf{X}}$ is a bijective mapping from $\mathcal{F}_{n-1-p,p}$ to $\mathcal{M}_{\mathbf{X}}$.*

The proof of Lemma A.7 involves mainly linear algebra and is deferred to the end of this section.

Part 2. We now define $[\mathbf{X}_0, \tilde{\mathbf{X}}_0]$. Let the eigenvalue decomposition of $\boldsymbol{S}$ be $\boldsymbol{G}_0\boldsymbol{D}^2\boldsymbol{G}_0^\top$, where $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with positive non-increasing diagonal entries and $\boldsymbol{G}_0 \in \mathcal{O}_p$. Define a $(n-1) \times p$ matrix $\mathbf{X}_*$ and a $(n-1) \times (n-1-p)$ matrix $\boldsymbol{Z}_*$ as

$$\mathbf{X}_* = \begin{bmatrix} \boldsymbol{D}\boldsymbol{G}_0^\top \\ \mathbf{0}_{(n-1-p)\times p} \end{bmatrix}, \ \boldsymbol{Z}_* = \begin{bmatrix} \mathbf{0}_{p\times(n-1-p)} \\ \boldsymbol{I}_{n-1-p} \end{bmatrix}.$$

Then $\boldsymbol{Z}_*^\top \boldsymbol{Z}_* = \boldsymbol{I}_{n-1-p}$, $\mathbf{X}_*^\top \mathbf{X}_* = \boldsymbol{S}$ and $\mathbf{X}_*^\top \boldsymbol{Z}_* = \mathbf{0}$. Next define

$$\begin{aligned} \mathbf{X}_0 &= \mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \mathbf{X}_*, \\ \boldsymbol{Z}_0 &= \boldsymbol{C}^\top \boldsymbol{Z}_*, \qquad\qquad \mathbf{V}_0 = \begin{bmatrix} \boldsymbol{I}_p \\ \mathbf{0}_{(n-1-2p)\times p} \end{bmatrix}, \\ \tilde{\mathbf{X}}_0 &= \mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X}_0 - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_0 \mathbf{V}_0 \boldsymbol{L}. \end{aligned}$$

One can check that $[\mathbf{X}_0, \tilde{\mathbf{X}}_0] \in \mathcal{M}$.

Part 3. Now for any $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{M}$, we will find a $\boldsymbol{G} \in \mathcal{G}$ such that $\psi(\boldsymbol{G}, [\mathbf{X}_0, \tilde{\mathbf{X}}_0]) = [\mathbf{X}, \tilde{\mathbf{X}}]$.

Let $\boldsymbol{Q}_{\mathbf{X}} = \boldsymbol{C}\mathbf{X}\boldsymbol{G}_0\boldsymbol{D}^{-1}$, which is a $(n-1)\times p$ matrix. Since $(\boldsymbol{C}\mathbf{X})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S}$, we have $\boldsymbol{Q}_{\mathbf{X}}^\top \boldsymbol{Q}_{\mathbf{X}} = \boldsymbol{I}_p$. Thus $\boldsymbol{Q}_{\mathbf{X}} \in \mathcal{F}_{n-1,p}$. By Lemma A.7, there is some $\mathbf{V} \in \mathcal{F}_{n-1-p,p}$ such that $\tilde{\mathbf{X}} = \mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}$. Let $\boldsymbol{Q}_{\tilde{\mathbf{X}}}$ be $\boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}}\mathbf{V}$. We will show $\boldsymbol{Q}_{\tilde{\mathbf{X}}} \in \mathcal{F}_{n-1,p}$ and $\boldsymbol{Q}_{\tilde{\mathbf{X}}}^\top \boldsymbol{Q}_{\mathbf{X}} = \mathbf{0}$: Because $\boldsymbol{Z}_{\mathbf{X}}^\top \mathbf{1}_n = \mathbf{0}$, it holds $\boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{Z}_{\mathbf{X}} = \boldsymbol{Z}_{\mathbf{X}}$. Thus

$$\begin{aligned} \boldsymbol{Q}_{\tilde{\mathbf{X}}}^\top \boldsymbol{Q}_{\tilde{\mathbf{X}}} &= (\boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}}\mathbf{V})^\top \boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}}\mathbf{V} \\ &= \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{Z}_{\mathbf{X}} \mathbf{V} \\ &= \boldsymbol{I}_p. \end{aligned}$$

In addition, because $\boldsymbol{Z}_{\mathbf{X}}^\top \mathbf{X} = \mathbf{0}$, it holds that

$$\begin{aligned} \boldsymbol{Q}_{\tilde{\mathbf{X}}}^\top \boldsymbol{Q}_{\mathbf{X}} &= \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{C}^\top \boldsymbol{Q}_{\mathbf{X}} \\ &= \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{C}^\top \boldsymbol{C}\mathbf{X}\boldsymbol{G}_0\boldsymbol{D}^{-1} \\ &= \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \mathbf{X}\boldsymbol{G}_0\boldsymbol{D}^{-1} \\ &= \mathbf{0}. \end{aligned}$$

Then we can find some $\boldsymbol{G}_* \in \mathcal{O}_{n-1}$ such that

$$(\boldsymbol{G}_*)_{1:(2p)} = [\boldsymbol{Q}_{\mathbf{X}}, \boldsymbol{Q}_{\tilde{\mathbf{X}}}] = [\boldsymbol{C}\mathbf{X}\boldsymbol{G}_0\boldsymbol{D}^{-1}, \boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}}\mathbf{V}]$$

Define $\boldsymbol{G} = \boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{C} + \mathbf{1}_n \mathbf{1}_n^\top / n$. One can check that $\boldsymbol{G}^\top \boldsymbol{G} = \boldsymbol{I}_n$ and $\boldsymbol{G}\mathbf{1}_n = \mathbf{1}_n$, and conclude that $\boldsymbol{G} \in \mathcal{G}$. We next show $[\boldsymbol{G}\mathbf{X}_0, \boldsymbol{G}\tilde{\mathbf{X}}_0] = [\mathbf{X}, \tilde{\mathbf{X}}]$.

We first check

$$\begin{aligned}
\boldsymbol{G}\mathbf{X}_0 &= (\boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{C} + \mathbf{1}_n \mathbf{1}_n^\top/n)(\mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \mathbf{X}_*) \\
&= \mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{C}\boldsymbol{C}^\top \mathbf{X}_* && (\because \boldsymbol{C}\mathbf{1}_n = \mathbf{0}) \\
&= \mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \boldsymbol{G}_* \mathbf{X}_* && (\because \boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{I}_{n-1}) \\
&= \mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \boldsymbol{Q}_{\mathbf{X}} \boldsymbol{D} \boldsymbol{G}_0^\top && (\because \text{definitions of } \boldsymbol{G}_*, \mathbf{X}_*) \\
&= \mathbf{1}_n \boldsymbol{m}^\top + \boldsymbol{C}^\top \boldsymbol{C}\mathbf{X} && (\because \text{definition of } \boldsymbol{Q}_{\mathbf{X}}) \\
&= \mathbf{X}.
\end{aligned}$$

Next, note that

$$\begin{aligned}
\boldsymbol{G}\boldsymbol{Z}_0\mathbf{V}_0 &= (\boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{C} + \mathbf{1}_n \mathbf{1}_n^\top/n)\boldsymbol{C}^\top \boldsymbol{Z}_* \mathbf{V}_0 \\
&= \boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{Z}_* \mathbf{V}_0 && (\because \boldsymbol{C}\mathbf{1}_n = \mathbf{0}) \\
&= \boldsymbol{C}^\top \boldsymbol{G}_* \boldsymbol{Z}_* \mathbf{V}_0 && (\because \boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{I}_{n-1}) \\
&= \boldsymbol{C}^\top \boldsymbol{G}_* \begin{bmatrix} \mathbf{0}_{p\times p} \\ \boldsymbol{I}_p \\ \mathbf{0}_{(n-1-2p)\times p} \end{bmatrix} && (\because \text{definitions of } \boldsymbol{Z}_*, \mathbf{V}_0) \\
&= \boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}}\mathbf{V} && (\because \text{definition of } \boldsymbol{G}_*) \\
&= \boldsymbol{Z}_{\mathbf{X}}\mathbf{V},
\end{aligned}$$

and hence it holds that

$$\begin{aligned}
\boldsymbol{G}\tilde{\mathbf{X}}_0 &= \boldsymbol{G}\left(\mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X}_0 - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_0 \mathbf{V}_0 \boldsymbol{L}\right) \\
&= \mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L} \\
&= \tilde{\mathbf{X}}.
\end{aligned}$$

Hence the operation ψ is transitive, and the proof is complete. □

Proof of Lemma A.7. The proof takes four steps.

Step 1: $\boldsymbol{L}$ is invertible.

Let

$$\boldsymbol{S}_* = \begin{bmatrix} 2\mathrm{diag}\,\{\boldsymbol{s}\} & \mathrm{diag}\,\{\boldsymbol{s}\} \\ \mathrm{diag}\,\{\boldsymbol{s}\} & \boldsymbol{S} \end{bmatrix}.$$

By construction of $\boldsymbol{s}$, $2\mathrm{diag}\,\{\boldsymbol{s}\} \succ \mathbf{0}_{p\times p}$ and

$$\begin{aligned}
& & 2\boldsymbol{S} &\succ \mathrm{diag}\,\{\boldsymbol{s}\} \\
\Rightarrow & & \boldsymbol{S} - \frac{1}{2}\mathrm{diag}\,\{\boldsymbol{s}\} &\succ \mathbf{0}_{p\times p} \\
\Rightarrow & & \boldsymbol{S} - \mathrm{diag}\,\{\boldsymbol{s}\}\,(2\mathrm{diag}\,\{\boldsymbol{s}\})^{-1}\,\mathrm{diag}\,\{\boldsymbol{s}\} &\succ \mathbf{0}_{p\times p},
\end{aligned}$$

where the lefthand side of the last line is exactly the Schur complement of $2\mathrm{diag}\,\{\boldsymbol{s}\}$ in $\boldsymbol{S}_*$, and therefore $\boldsymbol{S}_* \succ \mathbf{0}_{p\times p}$. But since $\boldsymbol{S} \succ \mathbf{0}_{p\times p}$, the fact that $\boldsymbol{S}_* \succ \mathbf{0}_{p\times p}$ implies that the Schur complement of $\boldsymbol{S}$ in $\boldsymbol{S}_*$ is also positive definite:

$$2\mathrm{diag}\,\{\boldsymbol{s}\} - \mathrm{diag}\,\{\boldsymbol{s}\}\boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\} = \boldsymbol{L}^\top \boldsymbol{L} \succ \mathbf{0}_{p\times p},$$

and therefore $\boldsymbol{L}$ is invertible.

Step 2: $\varphi_{\mathbf{X}}(\mathcal{F}_{n-1-p,p}) \subseteq \mathcal{M}_{\mathbf{X}}$.

Let $\tilde{\mathbf{X}} = \varphi_{\mathbf{X}}(\mathbf{V})$ for some $\mathbf{V} \in \mathcal{F}_{n-1-p,p}$. First we show $\tilde{\mathbf{X}}^\top \mathbf{1}_n/n = \boldsymbol{m}$:

$$\begin{aligned}
\tilde{\mathbf{X}}^\top \mathbf{1}_n/n &= \left(\mathbf{1}_n \boldsymbol{m}^\top + (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}\right)^\top \mathbf{1}_n/n \\
&= \boldsymbol{m} + \boldsymbol{L}^\top \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \mathbf{1}_n/n && \because (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)^\top \mathbf{1}_n = \mathbf{0}_p \\
&= \boldsymbol{m} && \because \boldsymbol{Z}_{\mathbf{X}}^\top \mathbf{1}_n = \mathbf{0}_{n-1-p}.
\end{aligned}$$

Next we show $(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S}$:

$$\begin{aligned}
(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} &= \left((\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}\right)^\top \left((\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}\right) \\
&\qquad \because \boldsymbol{C}^\top \boldsymbol{C}[\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top, \boldsymbol{Z}_{\mathbf{X}}] = [\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top, \boldsymbol{Z}_{\mathbf{X}}] \\
&= (\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\})^\top \boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{L}^\top \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L} \\
&\qquad \because (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)^\top \boldsymbol{Z}_{\mathbf{X}} = \mathbf{0}_{p\times p} \\
&= \boldsymbol{S} - 2\mathrm{diag}\,\{\boldsymbol{s}\} + \mathrm{diag}\,\{\boldsymbol{s}\}\boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\} + \boldsymbol{L}^\top \boldsymbol{L} \\
&\qquad \because \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{Z}_{\mathbf{X}} = \boldsymbol{I}_{n-1-p}, \mathbf{V} \in \mathcal{F}_{n-1-p,p} \\
&= \boldsymbol{S} && \because \boldsymbol{L}^\top \boldsymbol{L} = 2\mathrm{diag}\,\{\boldsymbol{s}\} - \mathrm{diag}\,\{\boldsymbol{s}\}\boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}.
\end{aligned}$$

And finally we show $(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} = \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\}$:

$$\begin{aligned}
(\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} &= \left((\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{Z}_{\mathbf{X}}\mathbf{V}\boldsymbol{L}\right)^\top \left(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top\right) \\
&= (\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\})^\top \boldsymbol{S} + \boldsymbol{L}^\top \mathbf{V}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \left(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top\right) \\
&= \boldsymbol{S} - \mathrm{diag}\,\{\boldsymbol{s}\} && \because (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)^\top \boldsymbol{Z}_{\mathbf{X}} = \mathbf{0}_{p\times(n-1-p)}.
\end{aligned}$$

We conclude that $\tilde{\mathbf{X}} \in \mathcal{M}_{\mathbf{X}}$ and therefore $\varphi_{\mathbf{X}}(\mathcal{F}_{n-1-p,p}) \subseteq \mathcal{M}_{\mathbf{X}}$.

Step 3: $\varphi_{\mathbf{X}}$ is injective.

Since $\boldsymbol{Z}_{\mathbf{X}}^\top [\mathbf{1}_n, \boldsymbol{X}] = \mathbf{0}$ and $\boldsymbol{L}$ is invertible, $\boldsymbol{Z}_{\mathbf{X}}^\top \varphi_{\mathbf{X}}(\mathbf{V})\boldsymbol{L}^{-1} = \mathbf{V}$. Thus $\varphi_{\mathbf{X}}$ is injective.

Step 4: $\varphi_{\mathbf{X}}$ is surjective.

Let $\tilde{\mathbf{X}} \in \mathcal{M}_{\mathbf{X}}$. By the definition of $\boldsymbol{Z}_{\mathbf{X}}$, the columns of $\left[\mathbf{1}_n, (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top), \boldsymbol{Z}_{\mathbf{X}}\right]$ form a basis of $\mathbb{R}^n$. Hence we can uniquely define $\boldsymbol{\alpha}^\top \in \mathbb{R}^{1\times p}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Theta} \in \mathbb{R}^{(n-1-p)\times p}$ such that

$$\tilde{\mathbf{X}} = \mathbf{1}_n \boldsymbol{\alpha}^\top + (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)\boldsymbol{\Lambda} + \boldsymbol{Z}_{\mathbf{X}}\boldsymbol{\Theta} \tag{A.8}$$

First, $\boldsymbol{m} = \tilde{\mathbf{X}}^\top \mathbf{1}_n/n = \boldsymbol{\alpha}$ because $\left[(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top), \boldsymbol{Z}_{\mathbf{X}}\right]^\top \mathbf{1}_n = \mathbf{0}_{(n-1)\times 1}$.

Next we show $\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}$:

$$\begin{aligned}
& \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\} = (\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\mathbf{X} \\
\Rightarrow \quad & \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\} = \boldsymbol{\Lambda}^\top (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)^\top \left(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top\right) + \boldsymbol{\Theta}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \left(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top\right) \\
\Rightarrow \quad & \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\} = \boldsymbol{\Lambda}^\top (\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top)^\top \left(\mathbf{X} - \mathbf{1}_n \boldsymbol{m}^\top\right) \\
\Rightarrow \quad & \boldsymbol{S} - \mathrm{diag}\{\boldsymbol{s}\} = \boldsymbol{\Lambda}^\top \boldsymbol{S} \\
\Rightarrow \quad & \boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\} = \boldsymbol{\Lambda}.
\end{aligned}$$

And finally, we show $\boldsymbol{\Theta} = \mathbf{V}\boldsymbol{L}$ for some $\mathbf{V} \in \mathcal{F}_{n-1-p,p}$. Using Equation (A.8),

$$\begin{aligned}
& (\boldsymbol{C}\tilde{\mathbf{X}})^\top \boldsymbol{C}\tilde{\mathbf{X}} = \boldsymbol{S} \\
\Rightarrow \quad & \boldsymbol{\Lambda}^\top \mathbf{X}^\top \boldsymbol{C}^\top \boldsymbol{C}\mathbf{X}\boldsymbol{\Lambda} + \boldsymbol{\Theta}^\top \boldsymbol{Z}_{\mathbf{X}}^\top \boldsymbol{Z}_{\mathbf{X}}\boldsymbol{\Theta} = \boldsymbol{S} \\
\Rightarrow \quad & (\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\})^\top \boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{S} \\
\Rightarrow \quad & \boldsymbol{S} - 2\mathrm{diag}\,\{\boldsymbol{s}\} + \mathrm{diag}\,\{\boldsymbol{s}\}\boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\} + \boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{S} \\
\Rightarrow \quad & \boldsymbol{\Theta}^\top\boldsymbol{\Theta} = 2\mathrm{diag}\,\{\boldsymbol{s}\} - \mathrm{diag}\,\{\boldsymbol{s}\}\boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\} \\
\Rightarrow \quad & \left(\boldsymbol{L}^{-1}\right)^\top \boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{L}^{-1} = \boldsymbol{I}_p,
\end{aligned}$$

where again the second equality uses $\boldsymbol{C}\mathbf{1}_n = \mathbf{0}$ and $\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{Z}_{\mathbf{X}} = \boldsymbol{Z}_{\mathbf{X}}$, the third equality uses $\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{S}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}$ and $\boldsymbol{Z}_{\mathbf{X}}^\top\boldsymbol{Z}_{\mathbf{X}} = \boldsymbol{I}_{n-1-p}$, and the last equality follows from the invertibility of $\boldsymbol{L}$. Define $\mathbf{V} := \boldsymbol{\Theta}\boldsymbol{L}^{-1}$, then the last equality implies $\mathbf{V} \in \mathcal{F}_{n-1-p,p}$. We conclude that $\tilde{\mathbf{X}} = \varphi_{\mathbf{X}}(\mathbf{V})$. □

A.2.4 An Intuitive Proof That Does Not Quite Work

The astute reader may think there is a more straightforward way than the previous subsection to prove Theorem 3.1 using the fact that all the randomness in the conditional knockoffs construction of Algorithm 1 comes from $[\boldsymbol{U}, \tilde{\boldsymbol{U}}]$ which follows the Haar measure on $\mathcal{F}_{n,2p}$, and this Haar measure has many known properties including swap-invariance. We show here why we were not able to follow this route, and resorted instead to a more technical proof using topological measure theory.

For simplicity, consider the special case where the mean vector is known to be zero, i.e. $\boldsymbol{x}_i \sim N(0, \boldsymbol{I} \otimes \boldsymbol{\Sigma})$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{X}$ where $\boldsymbol{U} \in \mathbb{R}^{n\times p}, \boldsymbol{D} \in \mathbb{R}^{p\times p}, \boldsymbol{V} \in \mathbb{R}^{p\times p}$. It is not hard to see that $\boldsymbol{U}$ is uniformly distributed on $\mathcal{F}_{n,p}$ and is independent of $\boldsymbol{D}\boldsymbol{V}^\top$. This claim implicitly uses the existence of a Haar measure on $\mathcal{F}_{n,p}$, but this is well-known (we denote this measure by $\mathrm{Unif}\,(\mathcal{F}_{n,p})$). Conditioning on $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}^T\boldsymbol{D}^2\boldsymbol{V} = \hat{\boldsymbol{\Sigma}}$,

$$\boldsymbol{X} \overset{d}{=} \mathrm{Unif}\,(\mathcal{F}_{n,p})\,\boldsymbol{D}\boldsymbol{V} \overset{d}{=} \mathrm{Unif}\,(\mathcal{F}_{n,p})\,\hat{\boldsymbol{\Sigma}}^{1/2}$$

Thus in principle, it would be sufficient to construct $\tilde{\boldsymbol{X}}$ such that

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \sim \mathrm{Unif}\,(\mathcal{F}_{n,2p}) \left[\begin{array}{cc} \hat{\boldsymbol{\Sigma}} & \hat{\boldsymbol{\Sigma}} - \mathrm{diag}\,\{\boldsymbol{s}\} \\ \hat{\boldsymbol{\Sigma}} - \mathrm{diag}\,\{\boldsymbol{s}\} & \hat{\boldsymbol{\Sigma}} \end{array}\right]^{1/2} \tag{A.9}$$

which simply requires generating the left singular vectors of $\mathrm{Unif}\,(\mathcal{F}_{n,2p})$ conditioned on $\boldsymbol{U}$ being the first p columns. This can be easily achieved by stacking $\boldsymbol{W}$ on the right of $\boldsymbol{X}$ and calculating the left singular values of $[\boldsymbol{X}, \boldsymbol{W}]$, which is exactly what is done in Algorithm 3.1.

To prove the validity of this construction, we just need to check that the right hand side of Equation (A.9) is swap-invariant. Indeed, $\mathrm{Unif}\,(\mathcal{F}_{n,2p})$ is easily shown to be swap-invariant, and the matrix multiplying it appears to be swap-invariant as well. However, the matrix square root complicates things. Denote

$$\boldsymbol{G} = \left[\begin{array}{cc} \hat{\boldsymbol{\Sigma}} & \hat{\boldsymbol{\Sigma}} - \mathrm{diag}\,\{\boldsymbol{s}\} \\ \hat{\boldsymbol{\Sigma}} - \mathrm{diag}\,\{\boldsymbol{s}\} & \hat{\boldsymbol{\Sigma}} \end{array}\right]^{1/2}.$$

To make the argument more precise, suppose that we want to show that swapping $\boldsymbol{X}_1$ with $\tilde{\boldsymbol{X}}_1$ does not change the joint distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$. Let $\boldsymbol{P} \in \mathbb{R}^{2p \times 2p}$ be the permutation matrix that swaps columns 1 and $1+p$ of a matrix when multiplied on the right. By Equation (A.9), what we need to show is

$$\text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{G}\boldsymbol{P} \overset{\mathcal{D}}{=} \text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{G} \tag{A.10}$$

The left hand side equals to $\left(\text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{P}\right) \boldsymbol{P}\boldsymbol{G}\boldsymbol{P}$. By known properties of the Haar measure, we have that $\text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{P} \overset{\mathcal{D}}{=} \text{Unif}\left(\mathcal{F}_{n,2p}\right)$, and hence Equation (A.10) is equivalent to

$$\text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{P}\boldsymbol{G}\boldsymbol{P} \overset{\mathcal{D}}{=} \text{Unif}\left(\mathcal{F}_{n,2p}\right) \boldsymbol{G} \tag{A.11}$$

The only way we can see how one might prove this more simply than the proof in our paper is to show that $\boldsymbol{P}\boldsymbol{G}\boldsymbol{P} = \boldsymbol{G}$, i.e., that $\boldsymbol{G}$ is swap-invariant.

$\boldsymbol{G}$ visually appears to be swap-invariant, and indeed is the square root of a swap-invariant matrix, but the fact that a matrix is swap-invariant does not directly imply that its square root is swap-invariant. The square root of a matrix in general is not unique, so we may hope that there exists (and we can identify) a swap-invariant square root in this case, but in the representation of Equation (A.9), we can actually only use the square root that has $\boldsymbol{D}\boldsymbol{V}^\top$ on its upper left block and has $\boldsymbol{0}$ on its bottom left block, in order to match $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ on the left hand side. Therefore, we can actually say with certainty that

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{D}\boldsymbol{V}^\top & \boldsymbol{D}^{-1}\boldsymbol{V}^\top \text{diag}\left\{\boldsymbol{s}\right\} - \hat{\boldsymbol{\Sigma}}^{-1}\text{diag}\left\{\boldsymbol{s}\right\} \\ 0 & \boldsymbol{L} \end{bmatrix},$$

where $\boldsymbol{L}^\top\boldsymbol{L} = 2\text{diag}\left\{\boldsymbol{s}\right\} - \text{diag}\left\{\boldsymbol{s}\right\}\hat{\boldsymbol{\Sigma}}^{-1}\text{diag}\left\{\boldsymbol{s}\right\}$ is a Cholesky decomposition. However, this matrix *cannot* be swap-invariant. This is why we were unable to prove swap-exchangeability of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ directly from swap-invariance of $\text{Unif}\left(\mathcal{F}_{n,2p}\right)$, and were instead forced to prove it directly using topological measure theory. Note that our proof uses similar machinery to the first-principles proof of the known result that $\text{Unif}\left(\mathcal{F}_{n,2p}\right)$ is swap-invariant.

A.3 Gaussian Graphical Models

Proof of Theorem 3.2. By classical results for the multivariate Gaussian distribution, we have

$$X_{B^c} \mid X_B \sim \mathcal{N}(\boldsymbol{\mu}^* + \boldsymbol{\Xi}X_B,\ \boldsymbol{\Sigma}^*), \tag{A.12}$$

where $\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{B^c,B}(\boldsymbol{\Sigma}_{B,B})^{-1}$, $\boldsymbol{\mu}^* = \boldsymbol{\mu}_{B^c} - \boldsymbol{\Xi}\boldsymbol{\mu}_B$ and $(\boldsymbol{\Sigma}^*)^{-1} = (\boldsymbol{\Sigma}^{-1})_{B^c,B^c}$. By the condition that G is n-separated by B, $(\boldsymbol{\Sigma}^*)^{-1}$ is block diagonal with blocks defined by the V_k's. Thus $X_{V_1}, \ldots, X_{V_\ell}$ are conditionally independent given X_B.

To show $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ is invariant to swapping A for any $A \subseteq [p]$, by conditional independence of the X_{V_k}'s, it suffices to show that for any $k \in [\ell]$ and $A_k := A \cap V_k$,

$$[\boldsymbol{X}_{V_k}, \tilde{\boldsymbol{X}}_{V_k}]_{\text{swap}(A_k)} \overset{\mathcal{D}}{=} [\boldsymbol{X}_{V_k}, \tilde{\boldsymbol{X}}_{V_k}] \mid \boldsymbol{X}_B. \tag{A.13}$$

Before proving Equation (A.13), we set up some notation. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, then by block matrix inversion (see, e.g., Kollo and von Rosen (2006, Proposition 1.3.3)),

$$\boldsymbol{\Sigma}_{B^c,B^c} - \boldsymbol{\Sigma}_{B^c,B}(\boldsymbol{\Sigma}_{B,B})^{-1}\boldsymbol{\Sigma}_{B,B^c} = \boldsymbol{\Omega}^{-1}_{B^c,B^c},$$

and

$$-\left(\boldsymbol{\Sigma}_{B^c,B^c}-\boldsymbol{\Sigma}_{B^c,B}(\boldsymbol{\Sigma}_{B,B})^{-1}\boldsymbol{\Sigma}_{B,B^c}\right)^{-1}\boldsymbol{\Sigma}_{B^c,B}(\boldsymbol{\Sigma}_{B,B})^{-1}=\boldsymbol{\Omega}_{B^c,B}.$$

Thus $\boldsymbol{\Xi}$ can be written as $-\boldsymbol{\Omega}_{B^c,B^c}^{-1}\boldsymbol{\Omega}_{B^c,B}$.

Now fix $k\in[\ell]$. Let $B_k=I_{V_k}\cap B$. Since V_k and $B\setminus B_k$ are not adjacent, $\boldsymbol{\Omega}_{V_k,B\setminus B_k}$, and thus $\boldsymbol{\Xi}_{V_k,B\setminus B_k}$, equals $\mathbf{0}$. Equation (A.12) implies

$$X_{V_k}\mid X_B\sim\mathcal{N}(\boldsymbol{\mu}^*_{V_k}+\boldsymbol{\Xi}_{V_k,B_k}X_{B_k},\ \boldsymbol{\Sigma}^*_{V_k,V_k}).$$

This also implies that

$$X_{V_k}\perp\!\!\!\perp X_{B\setminus B_k}\mid X_{B_k}. \tag{A.14}$$

Since the rows of $\boldsymbol{X}_{V_k\uplus B_k}$ are i.i.d. Gaussian, the validity of Algorithm 8 (see Theorem B.2 in Appendix B.1.3) says that $\tilde{\boldsymbol{X}}_{V_k}$ generated in Line 2 of Algorithm 2 satisfies

$$[\boldsymbol{X}_{V_k},\tilde{\boldsymbol{X}}_{V_k}]_{\mathrm{swap}(A_k)}\overset{\mathcal{D}}{=}[\boldsymbol{X}_{V_k},\tilde{\boldsymbol{X}}_{V_k}]\mid X_{B_k},$$

This together with Equation (A.14) shows Equation (A.13). This completes the proof. □

Proof of Proposition 3.3. This proof will be about Algorithm 10 in Appendix B.2, which is shown there to be equivalent to Algorithm 3. Without loss of generality, we assume $\pi=(1,\dots,p)$. Denote by $N_j^{(h)}$ the set N_j in the Algorithm 10 after the hth step. The updating steps of the algorithm ensure $j\notin N_j^{(h)}$ for any j and h. Note that N_j does not change after the $(j-1)$th step, i.e., $N_j^{(j-1)}=N_j^{(j)}=\cdots=N_j^{(p)}$.

It suffices to show the following inequality for each connected component W, whose vertex set is denoted by V, of the subgraph induced by deleting B:

$$1+2|V|+|I_V\cap B|\le n',\ \text{where } I_V:=\bigcup_{j\in V}I_j.$$

Part 1. First note that by definition of V, every element of I_V is either in V or B. Now define $F=[p]\setminus(V\uplus(I_V\cap B))$. We will show that $k\in F$ will never appear in N_j for any $j\in V$.

Initially, for any $j\in V$, $N_j^{(0)}=I_j$ does not intersect F. Suppose h is the smallest integer such that there exists some $j\in V$ such that $N_j^{(h)}$ contains some $k\in F$. By the construction of the algorithm, $h\notin B$, $j>h$ and $j\in N_h^{(h-1)}$ (otherwise $N_j^{(h)}$ would not have been altered in the hth step), $k\in N_h^{(h-1)}$ (otherwise k could not have entered $N_j^{(h)}$ at the hth step), and $h\in N_j^{(h-1)}$ (by symmetry of $N_j^{(i)}$ and $N_h^{(i)}$ for $i<\min(h,j)$).

Since $h\in N_j^{(h-1)}$, the definition of h guarantees $h\notin F$ (otherwise $h-1$ would be smaller and satisfy the condition defining h), and thus h is in either V or $I_V\cap B$. But since $h\notin B$, we must have $h\in V$. Now we have shown $k\in N_h^{(h-1)}$, i.e., F intersects N_h before the hth step, and $h\in V$, but this contradicts the definition of h. We conclude that for any $j\in V$ and any $h\in[p]$, $F\cap N_j^{(h)}=\emptyset$ and thus $N_j^{(p)}\subseteq(I_V\cap B)\uplus(V\setminus\{j\})$.

Part 2: We now characterize $N_j^{(p)}$. For any $j\in V$, define

$$L_j:=\{v\in(I_V\cap B)\uplus(V\setminus\{j\}):\ \exists\text{ a path }(j,j_1,\dots,j_m,v)\text{ in }G,\\ \text{where } j_i<j\text{ and }j_i\in V,\ \forall i\in[m]\}.$$

We will show $L_j \subseteq N_j^{(p)}$ by induction. This is true for the smallest $j \in V$ because $L_j = I_j \subseteq N_j^{(p)}$. Now assume $L_j \subseteq N_j^{(p)}$ for any $j < j_0$ (both in V), we will show $L_{j_0} \subseteq N_{j_0}^{(p)}$. For any $v \in L_{j_0}$, if $v \in I_{j_0}$ it is trivial that $v \in N_{j_0}^{(p)}$. If $v \in L_{j_0} \setminus I_j$, there is a path $(j_0, j_1, \ldots, j_m, v)$ in G where $\{j_i\}_{i=1}^m \subseteq V$ are all smaller than j_0. Let j_{i^*} be the largest among $\{j_i\}_{i=1}^m$. With the two paths $(j_0, j_1, \ldots, j_{i^*})$ and $(j_{i^*}, \ldots, j_m, v)$, we have $j_0, v \in L_{j_{i^*}} \subseteq N_{j_{i^*}}^{(p)}$ by the inductive hypothesis. Since $j_0 \in N_{j_{i^*}}^{(p)}$ and $j_0 > j_{i^*}$, in the j_{i^*}th step on Line 5, N_{j_0} absorbs $N_{j_{i^*}} \setminus \{j_0\}$, and it follows that $v \in N_{j_0}^{(j_{i^*})}$ and thus $v \in N_{j_0}^{(p)}$. We finally conclude that $L_{j_0} \subseteq N_{j_0}^{(p)}$, and by induction, $L_j \subseteq N_j^{(p)}$ for all $j \in V$.

Part 3. Let j^* be the largest number in V. Since W is connected and j^* is the largest, the definition of L_{j^*} implies $(I_V \cap B) \uplus (V \setminus \{j^*\}) = L_{j^*}$. Part 1 showed that $N_{j^*}^{(p)} \subseteq (I_V \cap B) \uplus (V \setminus \{j^*\})$ and Part 2 showed that $L_{j^*} \subseteq N_{j^*}^{(p)}$. Thus $N_{j^*}^{(p)} = (I_V \cap B) \uplus (V \setminus \{j^*\})$.

Since B keeps growing, at the jth step of Algorithm 10, the set $\{1, \ldots, j-1\} \setminus B$ with the current B is the same as that with the final B. At the j^*th step of the algorithm, $N_{j^*} \cap (\{1, \ldots, j^* - 1\} \setminus B)$ equals $V \setminus \{j^*\}$ (since j^* is the largest in V). Hence

$$|N_{j^*}| + |N_{j^*} \cap (\{1, \ldots, j^* - 1\} \setminus B)| = (|(I_V \cap B) \uplus V| - 1) + (|V| - 1) = 2|V| + |I_V \cap B| - 2.$$

Since $j^* \notin B$, the requirement in Line 4 and the equality above implies

$$1 + 2|V| + |I_V \cap B| \leq n',$$

and this completes the proof. □

A.4 Discrete Graphical Models

Proof of Theorem 3.4. We first show

$$\mathbb{P}(X_{B^c} \mid X_B) = \prod_{j \in B^c} \mathbb{P}(X_j \mid X_B) = \prod_{j \in B^c} \mathbb{P}\big(X_j \,\big|\, X_{I_j}\big). \tag{A.15}$$

Suppose $j \in B^c$, then $I_j \subseteq B$. By the local Markov property,

$$X_j \perp\!\!\!\perp (X_{B^c \setminus \{j\}}, X_{B \setminus I_j}) \mid X_{I_j}.$$

By the weak union property, we have

$$X_j \perp\!\!\!\perp X_{B^c \setminus \{j\}} \mid (X_{I_j}, X_{B \setminus I_j}),$$

which implies $\mathbb{P}(X_{B^c} \mid X_B) = \mathbb{P}(X_j \mid X_B)\mathbb{P}\big(X_{B^c \setminus \{j\}} \mid X_B\big)$. Following this logic for the remaining elements of $B^c \setminus \{j\}$, we have $\mathbb{P}(X_{B^c} \mid X_B) = \prod_{j \in B^c} \mathbb{P}(X_j \mid X_B)$, which is then equal to $\prod_{j \in B^c} \mathbb{P}\big(X_j \,\big|\, X_{I_j}\big)$ because $X_j \perp\!\!\!\perp X_{B \setminus I_j} \mid X_{I_j}$.

Secondly, as justified in Section 3.3.1, the construction of $\tilde{\boldsymbol{X}}_j$ in Algorithm 5 implies that conditional on $T_B(\boldsymbol{X})$, $\tilde{\boldsymbol{X}}_j$ and $\boldsymbol{X}_j$ are independent and identically distributed, and thus

$$(\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{X}}_j, \boldsymbol{X}_j) \mid T_B(\boldsymbol{X}).$$

By the law of total probability, it follows that

$$(\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{X}}_j, \boldsymbol{X}_j) \mid \boldsymbol{X}_B \tag{A.16}$$

Since $\tilde{\boldsymbol{X}}_j$ is generated without looking at $\boldsymbol{X}_{B^c \setminus \{j\}}$, it holds that

$$\tilde{\boldsymbol{X}}_j \perp\!\!\!\perp (\boldsymbol{X}_{B^c \setminus \{j\}}, \tilde{\boldsymbol{X}}_{B^c \setminus \{j\}}) \mid (\boldsymbol{X}_B, \boldsymbol{X}_j). \tag{A.17}$$

Next we show $(\boldsymbol{X}_{B^c}, \tilde{\boldsymbol{X}}_{B^c})_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} (\boldsymbol{X}_{B^c}, \tilde{\boldsymbol{X}}_{B^c})$ for any $A \subseteq B^c$. For any pair of column vectors $(\mathbf{X}_j, \tilde{\mathbf{X}}_j) \in [K_j]^n \times [K_j]^n$, define

$$(\mathbf{X}_j^A, \tilde{\mathbf{X}}_j^A) = \begin{cases} (\mathbf{X}_j, \tilde{\mathbf{X}}_j) & j \notin A \\ (\tilde{\mathbf{X}}_j, \mathbf{X}_j) & j \in A \end{cases}$$

By Equations (A.15) and (A.17),

$$\begin{aligned}
&\mathbb{P}\left((\boldsymbol{X}_{B^c}, \tilde{\boldsymbol{X}}_{B^c}) = (\mathbf{X}_{B^c}, \tilde{\mathbf{X}}_{B^c}) \,\middle|\, \boldsymbol{X}_B \right) \\
&= \prod_{j \in B^c} \mathbb{P}\left((\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) = (\mathbf{X}_j, \tilde{\mathbf{X}}_j) \,\middle|\, \boldsymbol{X}_B \right) \\
&= \prod_{j \in B^c \setminus A} \mathbb{P}\left((\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) = (\mathbf{X}_j, \tilde{\mathbf{X}}_j) \,\middle|\, \boldsymbol{X}_B \right) \times \prod_{j \in A} \mathbb{P}\left((\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) = (\tilde{\mathbf{X}}_j, \mathbf{X}_j) \,\middle|\, \boldsymbol{X}_B \right) \\
&= \prod_{j \in B^c \setminus A} \mathbb{P}\left((\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) = (\mathbf{X}_j^A, \tilde{\mathbf{X}}_j^A) \,\middle|\, \boldsymbol{X}_B \right) \times \prod_{j \in A} \mathbb{P}\left((\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) = (\mathbf{X}_j^A, \tilde{\mathbf{X}}_j^A) \,\middle|\, \boldsymbol{X}_B \right) \\
&= \mathbb{P}\left((\boldsymbol{X}_{B^c}, \tilde{\boldsymbol{X}}_{B^c}) = (\mathbf{X}_{B^c}^A, \tilde{\mathbf{X}}_{B^c}^A) \,\middle|\, \boldsymbol{X}_B \right), \\
&= \mathbb{P}\left((\boldsymbol{X}_{B^c}, \tilde{\boldsymbol{X}}_{B^c})_{\mathrm{swap}(A)} = (\mathbf{X}_{B^c}, \tilde{\mathbf{X}}_{B^c}) \,\middle|\, \boldsymbol{X}_B \right),
\end{aligned}$$

where the third equality (which swaps the order of $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$ and adds superscript A's in the second product) follows from Equation (A.16).

Together with $\tilde{\boldsymbol{X}}_B = \boldsymbol{X}_B$, we conclude $\tilde{\boldsymbol{X}}$ is a valid knockoff for $\boldsymbol{X}$. □

B Algorithmic Details

B.1 Low Dimensional Gaussian

B.1.1 Additional Details on Algorithm 1

We begin with the construction of a suitable $\boldsymbol{s}$ by extending existing algorithms for computing $\boldsymbol{s}$ to our situation. Without loss of generality we assume $\hat{\boldsymbol{\Sigma}}_{j,j} = 1$ for $j = 1, \ldots, p$ here; otherwise denote by $\hat{\boldsymbol{D}}$ the diagonal matrix with $\hat{\boldsymbol{D}}_{j,j} = \hat{\boldsymbol{\Sigma}}_{j,j}$, set $\hat{\boldsymbol{\Sigma}}^0$ to be $\hat{\boldsymbol{D}}^{-1/2}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}^{-1/2}$, define $\boldsymbol{s}^0 = \hat{\boldsymbol{D}}^{-1}\boldsymbol{s}$, and proceed with $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{s}$ replaced by $\hat{\boldsymbol{\Sigma}}^0$ and $\boldsymbol{s}^0$ respectively. For any $\epsilon, \delta \in (0, 1)$, we can compute $\boldsymbol{s}$ in any of the following ways:

- *Equicorrelated* (Barber and Candès, 2015): Take $s_j^{\mathrm{EQ}} = (1 - \epsilon) \min\left(2\lambda_{\min}(\hat{\boldsymbol{\Sigma}}), 1\right)$ for all $j = 1, \ldots, p$.

- *Semidefinite program (SDP)* (Barber and Candès, 2015): Take $\boldsymbol{s}^{\text{SDP}}$ to be the solution to the following convex optimization:

$$\min \sum_{j=1}^{p} (1 - s_j) \quad \text{subject to:} \quad \begin{array}{l} \delta \le s_j \le 1,\ j = 1, \dots, p \\ \text{diag}\,\{\boldsymbol{s}\} \preceq (1-\epsilon) 2\hat{\boldsymbol{\Sigma}}. \end{array} \tag{B.1}$$

- *Approximate SDP* (Candès et al., 2018): Choose an approximation $\hat{\boldsymbol{\Sigma}}_{\text{approx}}$ of $\hat{\boldsymbol{\Sigma}}$ and compute $\boldsymbol{s}^{\text{approx}}$ by the SDP method as if $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{approx}}$. Then set $\boldsymbol{s} = \gamma \boldsymbol{s}_{\text{approx}}$ where γ solves

$$\max \gamma \qquad \text{subject to:} \quad \text{diag}\,\{\gamma \boldsymbol{s}_{\text{approx}}\} \preceq (1-\epsilon) 2\hat{\boldsymbol{\Sigma}}.$$

Noting that $\tilde{\boldsymbol{X}}_j^\top \boldsymbol{X}_j / n = \hat{\Sigma}_{j,j} - s_j$, it will always be preferable to take ϵ as small as possible (for all methods), so that $\boldsymbol{s}$ is as large as possible and $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ are as different as possible. For the SDP method, the lower bound δ can be set to be s_j^{EQ} multiplied by a small number, e.g., $\delta = 0.1 \cdot 2\lambda_{\min}(\hat{\boldsymbol{\Sigma}})$, to guarantee feasibility; this choice is used in the simulations in Sections 3.1 and 3.2.

We now prove the computational complexity of Algorithm 1. The Cholesky decomposition takes $O(p^3)$ operations and the Gram–Schmidt orthonormalization takes $O(np^2)$ operations. If $\boldsymbol{s}$ is computed by the *Equicorrelated* method whose complexity is no larger than $O(p^3)$, the overall complexity of Algorithm 1 is $O(np^2)$.

B.1.2 Gaussian Knockoffs with Known Mean

Algorithm 7 is a slight modification of Algorithm 1 for mulitvariate Gaussian models with mean parameter $\boldsymbol{\mu}$ known. The proof of its validity requires only minor modification of the proof of Theorem 3.1, and is thus omitted.

Algorithm 7 Conditional Knockoffs for Low-Dimensional Gaussian Models with Known Mean

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\mu} \in \mathbb{R}^p$.
Require: $n \ge 2p$.
1: Define $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{X} - \boldsymbol{1}_n \boldsymbol{\mu}^\top)^\top (\boldsymbol{X} - \boldsymbol{1}_n \boldsymbol{\mu}^\top)$.
2: Find $\boldsymbol{s} \in \mathbb{R}^p$ such that $\boldsymbol{0}_{p \times p} \prec \text{diag}\,\{\boldsymbol{s}\} \prec 2\hat{\boldsymbol{\Sigma}}$.
3: Compute the Cholesky decomposition of $2\text{diag}\,\{\boldsymbol{s}\} - \text{diag}\,\{\boldsymbol{s}\}\hat{\boldsymbol{\Sigma}}^{-1}\text{diag}\,\{\boldsymbol{s}\}$ as $\boldsymbol{L}^\top \boldsymbol{L}$.
4: 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 $[\underbrace{\boldsymbol{Q}}_{n \times p}, \underbrace{\boldsymbol{U}}_{n \times p}]$ of $[\boldsymbol{X}, \boldsymbol{W}]$.
5: Set

$$\tilde{\boldsymbol{X}} = \boldsymbol{1}_n \boldsymbol{\mu}^\top + (\boldsymbol{X} - \boldsymbol{1}_n \boldsymbol{\mu}^\top)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}^{-1}\text{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{U}\boldsymbol{L}. \tag{B.2}$$

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

B.1.3 Partial Gaussian Knockoffs with Fixed Columns

Consider the case where some of the variables are known to be relevant and thus do not need to have knockoffs generated for them. Let $B \subseteq [p]$ be the set of variables that no knockoffs are needed

for, so we only want to construct knockoffs for variables in $V = B^c$, i.e., to generate $\tilde{\boldsymbol{X}}_V$ such that for any subset $A \subseteq V$,

$$[\boldsymbol{X}_V, \tilde{\boldsymbol{X}}_V]_{\text{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}_V, \tilde{\boldsymbol{X}}_V] \mid \boldsymbol{X}_B.$$

Algorithm 8 provides a way to generate such knockoffs. We can find its computational complexity as follows. Fitting the least squares in Line 1 takes $O(n|B|^2|V|)$, computing $\hat{\boldsymbol{\Sigma}}$ takes $O(n|V|^2)$, both the most efficient construction of $\boldsymbol{s}$ and inverting $\hat{\boldsymbol{\Sigma}}$ take $O(|V|^3)$, and the Gram–Schmidt orthonormalization takes $O(n(1 + |B| + 2|V|)^2)$. Hence the overall computational complexity is $O\left(n|B|^2|V| + n|V|^2\right)$.

Algorithm 8 Partial Conditional Knockoffs for Low-Dimensional Gaussian Models

Input: $\boldsymbol{X}_V \in \mathbb{R}^{n\times|V|}$, columns to condition on: $\boldsymbol{X}_B \in \mathbb{R}^{n\times|B|}$.

Require: $n > 2|V| + |B|$.

1: Compute the least squares fitted value $\hat{\boldsymbol{X}}_j$ and residual $\boldsymbol{R}_j$ from regressing $\boldsymbol{X}_j$ on $[\boldsymbol{1}_n, \boldsymbol{X}_B]$ for each $j \in V$. Let $\boldsymbol{R} = [\ldots, \boldsymbol{R}_j, \ldots]_{j\in V}$ and compute $\hat{\boldsymbol{\Sigma}} = \boldsymbol{R}^\top \boldsymbol{R}$.

2: Find $\boldsymbol{s} \in \mathbb{R}^{|V|}$ such that $\boldsymbol{0}_{|V|\times|V|} \prec \text{diag}\{\boldsymbol{s}\} \prec 2\hat{\boldsymbol{\Sigma}}$.

3: Compute the Cholesky decomposition of $2\text{diag}\{\boldsymbol{s}\} - \text{diag}\{\boldsymbol{s}\}\hat{\boldsymbol{\Sigma}}^{-1}\text{diag}\{\boldsymbol{s}\}$ as $\boldsymbol{L}^\top \boldsymbol{L}$.

4: Generate $\boldsymbol{W}$ a $n \times |V|$ matrix whose entries are i.i.d. $\mathcal{N}(0,1)$ and independent of $\boldsymbol{X}$ and compute the Gram–Schmidt orthonormalization $[\underbrace{\boldsymbol{U}_0}_{n\times(1+|B|+|V|)}, \underbrace{\boldsymbol{U}}_{n\times|V|}]$ of $[\boldsymbol{1}_n, \boldsymbol{X}_B, \boldsymbol{R}, \boldsymbol{W}]$.

5: Set $\tilde{\boldsymbol{X}}_V = \hat{\boldsymbol{X}}_V + \boldsymbol{R}\left(\boldsymbol{I}_{|V|} - (\hat{\boldsymbol{\Sigma}})^{-1}\text{diag}\{\boldsymbol{s}\}\right) + \boldsymbol{U}\boldsymbol{L}$.

6: **return** $\tilde{\boldsymbol{X}}_V$.

The validity of Algorithm 8 relies on its equivalence to a straightforward but slow algorithm, Algorithm 9. We first show the validity of Algorithm 9 and then show the equivalence.

Algorithm 9 Alternative Form of Algorithm 8

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $B \subseteq [p]$ and $V = B^c$.

Require: $n > 2|V| + |B|$.

1: Generate a $n \times (n-1-|B|)$ orthonormal matrix $\boldsymbol{Q}$ that is orthogonal to $[\boldsymbol{1}_n, \boldsymbol{X}_B]$, and compute $\boldsymbol{Q}_\perp$ an orthonormal basis for the column space of $[\boldsymbol{1}_n, \boldsymbol{X}_B]$.

2: Construct low-dimensional knockoffs $\boldsymbol{J}$ for $\boldsymbol{Q}^\top \boldsymbol{X}_V$ via Algorithm 7 with $\boldsymbol{\mu} = \boldsymbol{0}$.

3: Set $\tilde{\boldsymbol{X}}_V = \boldsymbol{Q}\boldsymbol{J} + \boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top \boldsymbol{X}_V$.

4: **return** $\tilde{\boldsymbol{X}}_V$.

Proposition B.1. *Algorithm 9 generates valid knockoff for $\boldsymbol{X}_V$ conditional on $\boldsymbol{X}_B$.*

Proof. By classical results for the multivariate Gaussian distribution, we have

$$X_V \mid X_B \sim \mathcal{N}(\boldsymbol{\mu}^* + \boldsymbol{\Xi} X_B,\ \boldsymbol{\Sigma}^*), \tag{B.3}$$

where $\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{V,B}(\boldsymbol{\Sigma}_{B,B})^{-1}$, $\boldsymbol{\mu}^* = \boldsymbol{\mu}_V - \boldsymbol{\Xi}\boldsymbol{\mu}_B$ and $(\boldsymbol{\Sigma}^*)^{-1} = (\boldsymbol{\Sigma}^{-1})_{V,V}$.

We want to show that for any $A \subseteq V$,

$$[\boldsymbol{X}_V, \tilde{\boldsymbol{X}}_V]_{\text{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{X}_V, \tilde{\boldsymbol{X}}_V] \mid \boldsymbol{X}_B. \tag{B.4}$$

For n i.i.d. samples,

$$\boldsymbol{X}_V \mid \boldsymbol{X}_B \sim \mathcal{N}_{n,|V|}(\mathbf{1}_n(\boldsymbol{\mu}^*)^\top + \boldsymbol{X}_B\boldsymbol{\Xi}^\top,\ \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^*).$$

By the definition of $\boldsymbol{Q}$ in Algorithm 2, $\boldsymbol{Q}^\top[\mathbf{1}_n, \boldsymbol{X}_B] = \mathbf{0}_{(n-1-|B|)\times(1+|B|)}$ and $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_{n-1-|B|}$. This together with the property in Equation (A.3) implies

$$\boldsymbol{Q}^\top\boldsymbol{X}_V \mid \boldsymbol{X}_B \sim \mathcal{N}_{n-1-|B|,|V|}(\mathbf{0},\ \boldsymbol{I}_{n-1-|B|} \otimes \boldsymbol{\Sigma}^*).$$

Since $n - 1 - |B| \geq 2|V|$, Algorithm 7 can be used to generate knockoffs $\boldsymbol{J}$ for $\boldsymbol{Q}^\top\boldsymbol{X}_V$, which satisfies that

$$[\boldsymbol{Q}^\top\boldsymbol{X}_V, \boldsymbol{J}]_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{Q}^\top\boldsymbol{X}_V, \boldsymbol{J}] \mid \boldsymbol{X}_B,$$

and thus

$$[\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{X}_V, \boldsymbol{Q}\boldsymbol{J}]_{\mathrm{swap}(A)} \overset{\mathcal{D}}{=} [\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{X}_V, \boldsymbol{Q}\boldsymbol{J}] \mid \boldsymbol{X}_B.$$

Adding $[\boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top\boldsymbol{X}_V, \boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top\boldsymbol{X}_V]$, which is trivially invariant to swapping, to both sides and using $\boldsymbol{I}_n = \boldsymbol{Q}\boldsymbol{Q}^\top + \boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top$ and the definition of $\tilde{\boldsymbol{X}}_V$ in Line 3 of Algorithm 2, we have

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

Since this holds for any $A \subseteq V$, $\tilde{\boldsymbol{X}}_V$ is a valid knockoff matrix for $\boldsymbol{X}_V$ conditional on $\boldsymbol{X}_B$. □

Theorem B.2. *Algorithm 8 generates valid knockoffs for $\boldsymbol{X}_V$ conditional on $\boldsymbol{X}_B$.*

Proof. It suffices to show that if the same $\boldsymbol{s}$ and $\boldsymbol{L}$ in Algorithm 8 are used to generate $\boldsymbol{J}$ in Line 2 of Algorithm 9, then the output $\tilde{\boldsymbol{X}}_V$ in Algorithm 8 and the output in Algorithm 9, which is denoted by $\tilde{\boldsymbol{X}}'_V$ to avoid confusion, have the same conditional distribution given $\boldsymbol{X}_B$ and $\boldsymbol{X}_V$.

We write the Gram–Schmidt orthonormalization as a function $\Psi(\cdot)$. Let $b = 1+|B|$ and $d = |V|$. By assumption, $b + 2d \leq n$.

By the definition of $\boldsymbol{Q}$ and $\boldsymbol{Q}_\perp$ in Line 1 of Algorithm 9, we have

$$\hat{\boldsymbol{X}}_V = \boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top\boldsymbol{X}_V,\ \boldsymbol{R} = \boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{X}_V.$$

First, we express $\tilde{\boldsymbol{X}}'_V$ in a similar form as $\tilde{\boldsymbol{X}}_V$ in Line 5 of Algorithm 8. The conditional knockoff matrix for $\boldsymbol{Q}^\top\boldsymbol{X}_V$ generated by Algorithm 7 (with $\boldsymbol{\mu} = \mathbf{0}_{(n-b)\times 1}$) is given by

$$\boldsymbol{J} = \boldsymbol{Q}^\top\boldsymbol{X}_V(\boldsymbol{I}_d - (\hat{\boldsymbol{\Sigma}}')^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}) + \boldsymbol{U}'\boldsymbol{L},$$

where $\hat{\boldsymbol{\Sigma}}' = \boldsymbol{X}_V^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{X}_V = \boldsymbol{R}^\top\boldsymbol{R} = \hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{U}'$ is the last d columns of the Gram–Schmidt orthonormalization of $[\boldsymbol{Q}^\top\boldsymbol{X}_V, \boldsymbol{W}']$ with $\boldsymbol{W}' \sim \mathcal{N}_{n-b,d}(\mathbf{0}, \boldsymbol{I}_{n-b} \otimes \boldsymbol{I}_d)$ independent of $\boldsymbol{X}_V$ and $\boldsymbol{X}_B$. Hence we have

$$\begin{aligned}\tilde{\boldsymbol{X}}'_V &= \boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top\boldsymbol{X}_V + \boldsymbol{Q}\boldsymbol{J} \\ &= \hat{\boldsymbol{X}}_V + \boldsymbol{R}\left(\boldsymbol{I}_d - \hat{\boldsymbol{\Sigma}}^{-1}\mathrm{diag}\,\{\boldsymbol{s}\}\right) + \boldsymbol{Q}\boldsymbol{U}'\boldsymbol{L}. \end{aligned} \tag{B.5}$$

It suffices to show $\boldsymbol{U}$ in Line 4 of Algorithm 8 is distributed the same as $\boldsymbol{Q}\boldsymbol{U}'$ conditional on $\boldsymbol{X}$.

Without loss of generality (by choosing $\boldsymbol{Q}_\perp$ in Line 1 of Algorithm 9), assume the Gram–Schmidt orthonormalization of $[\mathbf{1}_n, \boldsymbol{X}_B, \boldsymbol{X}_V]$ is $[\boldsymbol{Q}_\perp, \boldsymbol{M}]$, where $\boldsymbol{M}$ is a $n \times d$ matrix. Hence $\operatorname{span}(\boldsymbol{M}) = \operatorname{span}(\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{X}_V)$. Let $\boldsymbol{Z}$ be a $(n-b) \times (n-b-d)$ matrix whose columns form an orthonormal basis for the orthogonal complement of $\operatorname{span}(\boldsymbol{Q}^\top \boldsymbol{X}_V)$.

Characterizing $\boldsymbol{U}$: Let $\boldsymbol{\Gamma} = [\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{Q}\boldsymbol{Z}]$. Since $\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{X}_V = \mathbf{0}$, we have $\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{X}_V = \mathbf{0}$ and thus $\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{M} = \mathbf{0}$. Together with $\boldsymbol{Q}_\perp^\top \boldsymbol{Q}\boldsymbol{Z} = \mathbf{0}$ and $(\boldsymbol{Q}\boldsymbol{Z})^\top \boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{I}_{n-b-d}$, we have $\boldsymbol{\Gamma} \in \boldsymbol{O}_n$.

Using Equation (A.4),

$$\begin{aligned}\Psi([\mathbf{1}_n, \boldsymbol{X}_B, \boldsymbol{X}_V, \boldsymbol{W}]) &= \Psi([\Psi([\mathbf{1}_n, \boldsymbol{X}_B, \boldsymbol{X}_V]), \boldsymbol{W}]) \\ &= \Psi([\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{W}]) \\ &= \boldsymbol{\Gamma}\Psi(\boldsymbol{\Gamma}^\top [\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{W}]),\end{aligned} \tag{B.6}$$

where the first equality is due to the fact that Gram–Schmidt orthonormalization treats the columns of its inputs sequentially. An elementary calculation shows

$$\boldsymbol{\Gamma}^\top [\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{W}] = \begin{bmatrix} \boldsymbol{I}_b & \mathbf{0}_{b\times d} & \boldsymbol{Q}_\perp^\top \boldsymbol{W} \\ \mathbf{0}_{d\times b} & \boldsymbol{I}_d & \boldsymbol{M}^\top \boldsymbol{W} \\ \mathbf{0}_{(n-b-d)\times b} & \mathbf{0}_{(n-b-d)\times d} & \boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W} \end{bmatrix},$$

thus

$$\Psi\left(\boldsymbol{\Gamma}^\top [\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{W}]\right) = \begin{bmatrix} \boldsymbol{I}_b & \mathbf{0}_{b\times d} & \mathbf{0}_{b\times(n-b-d)} \\ \mathbf{0}_{d\times b} & \boldsymbol{I}_d & \mathbf{0}_{d\times(n-b-d)} \\ \mathbf{0}_{(n-b-d)\times b} & \mathbf{0}_{(n-b-d)\times d} & \Psi(\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W}) \end{bmatrix}. \tag{B.7}$$

Using the definition of $\boldsymbol{\Gamma}$ and Equations (B.6) and (B.7), we conclude

$$\Psi([\mathbf{1}_n, \boldsymbol{X}_B, \boldsymbol{X}_V, \boldsymbol{W}]) = [\boldsymbol{Q}_\perp, \boldsymbol{M}, \boldsymbol{Q}\boldsymbol{Z}\Psi(\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W})],$$

which implies

$$\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{Z}\Psi(\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W}).$$

Noting that $\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{I}_{n-b-d}$ and $\boldsymbol{W} \sim \mathcal{N}_{n,d}(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{I}_d)$, Equation (A.3) implies $\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W} \sim \mathcal{N}_{n-b-d,d}(\mathbf{0}, \boldsymbol{I}_{n-b-d} \otimes \boldsymbol{I}_d)$. By the classic result in Eaton (1983, Proposition 7.2), the conditional distribution of $\Psi(\boldsymbol{Z}^\top \boldsymbol{Q}^\top \boldsymbol{W})$ given $(\boldsymbol{X}_B, \boldsymbol{X}_V)$ is the unique $\mathcal{O}_{n-b-d}$-invariant probability measure on $\mathcal{F}_{n-b-d,d}$.

Characterizing $\boldsymbol{U}'$: Let $\boldsymbol{Z}_\perp \in \mathbb{R}^{(n-b)\times d}$ be the Gram–Schmidt orthonormalization of $\boldsymbol{Q}^\top \boldsymbol{X}_V$, and thus $\boldsymbol{Z}^\top \boldsymbol{Z}_\perp = \mathbf{0}$. Let $\boldsymbol{\Gamma}_z = [\boldsymbol{Z}_\perp, \boldsymbol{Z}]$, then $\boldsymbol{\Gamma}_z \in \mathcal{O}_{n-b}$. Again using the properties of Gram–Schmidt orthonormalization,

$$\Psi([\boldsymbol{Q}^\top \boldsymbol{X}_V, \boldsymbol{W}']) = \Psi([\boldsymbol{Z}_\perp, \boldsymbol{W}']) = \boldsymbol{\Gamma}_z \Psi(\boldsymbol{\Gamma}_z^\top [\boldsymbol{Z}_\perp, \boldsymbol{W}']). \tag{B.8}$$

Since

$$\boldsymbol{\Gamma}_z^\top [\boldsymbol{Z}_\perp, \boldsymbol{W}'] = \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{Z}_\perp^\top \boldsymbol{W}' \\ \mathbf{0}_{(n-b-d)\times d} & \boldsymbol{Z}^\top \boldsymbol{W}' \end{bmatrix},$$

it holds that

$$\Psi\left(\boldsymbol{\Gamma}_z^\top[\boldsymbol{Z}_\perp, \boldsymbol{W}']\right) = \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{0}_{d\times d} \\ \boldsymbol{0}_{(n-b-d)\times d} & \Psi(\boldsymbol{Z}^\top \boldsymbol{W}') \end{bmatrix}. \tag{B.9}$$

Hence $\boldsymbol{U}' = \boldsymbol{Z}\Psi(\boldsymbol{Z}^\top \boldsymbol{W}')$ by combining Equations (B.8) and (B.9). As before, we can conclude that the conditional distribution of $\Psi(\boldsymbol{Z}^\top \boldsymbol{W}')$ given $(\boldsymbol{X}_B, \boldsymbol{X}_V)$ is the unique $\mathcal{O}_{n-b-d}$-invariant probability measure on $\mathcal{F}_{n-b-d,d}$.

Combining the two parts above and using the uniqueness of the invariant measure, we conclude that

$$\boldsymbol{U} \overset{\mathcal{D}}{=} \boldsymbol{Q}\boldsymbol{U}' \mid (\boldsymbol{X}_B, \boldsymbol{X}_V).$$

Using the definition of $\tilde{\boldsymbol{X}}_V$ in Line 4 and Equation (B.5), it follows that $\tilde{\boldsymbol{X}}_V \overset{\mathcal{D}}{=} \tilde{\boldsymbol{X}}'_V \mid (\boldsymbol{X}_B, \boldsymbol{X}_V)$. □

B.2 Gaussian Graphical Models

The computational complexity of Algorithm 2 can be shown by summing up the computational complexity of Algorithm 8 in Line 2 for individual connected components, which is $O\left(n\left|I_{V_k} \cap B\right|^2 |V_k| + |V_k|^2\right)$, as shown in Appendix B.1.3. Its upper bound is due to the facts that $\sum_{k=1}^{\ell} |V_k| \leq p$ and $\max_{1\leq k\leq \ell} |V_k| \leq n'$.

B.2.1 Greedy Search for a Blocking Set

Algorithm 10 is the virtual implementation of Algorithm 3. In Line 5 of Algorithm 10, we only need to keep track of N_j the neighborhood of each unvisited j in $\bar{G}$ among the vertices in $[p]$. This is because if $k \in N_j$ and $\tilde{k}$ exists in $\bar{G}$ then it is guaranteed by Algorithm 3 that $\tilde{k}$ is a neighbor of j in $\bar{G}$, and j is a neighbor of both k and $\tilde{k}$. This also implies that $|N_j \cap \{\pi_1, \ldots, \pi_{t-1}\} \setminus B|$ equals the size of the neighborhood of j in $\bar{G}$ among the knockoff vertices. Also note that the neighborhood of a visited vertex is no longer used in Line 4 of Algorithm 3, therefore the update step in Line 5 of Algorithm 10 can be restricted to the unvisited k's. In the following, we use the equivalence between Algorithm 3 and Algorithm 10 to prove the properties of Algorithm 3.

Algorithm 10 Greedy Search for a Blocking Set

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

1: Initialize $N_j = I_j$ for all $j \in [p]$, $B = \emptyset$.
2: **for** $t = 1, \ldots, p$ **do**
3: Let j be π_t.
4: **if** $n' \geq 3 + |N_j| + |N_j \cap \{\pi_1, \ldots, \pi_{t-1}\} \setminus B|$ **then**
5: Update $N_k \leftarrow N_k \cup (N_j \setminus \{k\})$ for all $k \in N_j \cap \{\pi_{t+1}, \ldots, \pi_p\}$.
6: **else**
7: $B \leftarrow B \cup \{j\}$.
8: **end if**
9: **end for**
10: **return** B.

The following proposition shows that if the tail of the input permutation to Algorithm 3 is already a blocking set of the graph, then the output from the algorithm is a subset of this blocking set. This property allows one to refine a known but large blocking set (e.g., one could apply Algorithm 3 to the blocking set from Example 2 in Appendix B.2.3).

Proposition B.3. *Suppose n' and π are the inputs of Algorithm 3, which returns a blocking set B. If G is n'-separated by $\{\pi_{m+1}, \ldots, \pi_p\}$ for some $m \in [p]$, then $\pi_1, \ldots, \pi_m$ will not be in B.*

Proof. In this proof, we use the equivalence between Algorithm 3 and Algorithm 10. Let $D = \{\pi_{m+1}, \ldots, \pi_p\}$. Without loss of generality, re-index the variables so that $\pi_j = j$ for every $j \in [p]$, and thus $D = \{m+1, \ldots, p\}$. Denote by $N_j^{(h)}$ the set N_j in Algorithm 10 after the hth step, as in the proof of Proposition 3.3. Let W be any of the connected components of the subgraph induced by deleting D, and V be the vertex set of W. Then $V \subseteq \{1, \ldots, m\}$.

Part 1. We first show that $N_j^{(h)} \subseteq V \uplus (I_V \cap D)$ for any $j \in V$ and $h \in [p]$. The proof is similar to Part 1 in the proof of Proposition 3.3. By definition of V, every element of I_V is either in V or D. Define $F = [p] \setminus (V \uplus (I_V \cap D))$. It suffices to show that $k \in F$ will never appear in N_j for any $j \in V$.

Initially, for any $j \in V$, $N_j^{(0)} = I_j$ does not intersect F. Suppose h is the smallest integer such that there exists some $j \in V$ such that $N_j^{(h)}$ contains some $k \in F$. By the construction of the algorithm, $j > h$ and $j \in N_h^{(h-1)}$ (otherwise $N_j^{(h)}$ would not have been altered in the hth step), $k \in N_h^{(h-1)}$ (otherwise k could not have entered $N_j^{(h)}$ at the hth step), and $h \in N_j^{(h-1)}$ (by symmetry of $N_j^{(i)}$ and $N_h^{(i)}$ for $i < \min(h, j)$). The fact that $h < j \leq m$ implies $h \notin D$. Since $h \in N_j^{(h-1)}$, the definition of h guarantees $h \notin F$ (otherwise $h - 1$ would be smaller and satisfy the condition defining h), and thus h is in either V or $I_V \cap D$. But since $h \notin D$, we must have $h \in V$. Now we have shown $k \in N_h^{(h-1)}$, i.e., F intersects N_h before the hth step, and $h \in V$, but this contradicts the definition of h. We conclude that for any $j \in V$ and any $h \in [p]$, $F \cap N_j^{(h)} = \emptyset$ and thus $N_j^{(j-1)} \subseteq (I_V \cap D) \uplus (V \setminus \{j\})$.

Part 2. For any $j \in V$, at the jth step of Algorithm 10, $N_j \cap \{1, \ldots, j-1\} \subseteq V \setminus \{j\}$ by the definition of D. Hence we have

$$\begin{aligned} 3 + |N_j \cap \{1, \ldots, j-1\} \setminus B| + |N_j| &\leq 3 + |V| - 1 + |V \uplus (I_V \cap D)| - 1 \\ &\leq 1 + 2|V| + |(I_V \cap D)| \\ &\leq n', \end{aligned}$$

where the last inequality is because of the condition that G is n'-separated by D. Thus the requirement in Line 4 of Algorithm 10 is satisfied and j is not in the blocking set.

Finally, since j and W are arbitrary, we conclude that any vertex in $\{1, \ldots, m\}$ is not blocked. □

B.2.2 Searching for Blocking Sets

Given any m, Algorithm 11 performs a randomized greedy search for the blocking sets B_i. Although there is no guarantee that the B_i's found by Algorithm 11 satisfy $\bigcap_{i=1}^{m} B_i = \emptyset$, one can subsequently check whether $\eta_j = m$ for any $j \in [p]$, in which case the algorithm can be run again. Inspecting the vertices with $\eta_j = m$ may reveal the difficulties of blocking for this graph. Changing the inputs m and n' may also help.

Algorithm 11 Randomized Greedy Search for Blocking Sets

Input: G, m, n' (by default $n' = \lfloor n/m \rfloor$).
Require: $n' \leq n/m$.
1: Let $\{\eta_j = 0\}_{j=1}^p$ be a sequence counting how often a variable is in a blocking set.
2: **for** $i = 1, \ldots, m$ **do**
3: Set π to be the decreasing order (with ties broken randomly) of η_j's, so $\eta_{\pi_1} \geq \eta_{\pi_2} \geq \cdots \geq \eta_{\pi_p}$.
4: Run Algorithm 3 with π and n' and let $B^{(i)}$ be the returned blocking set.
5: Update $\eta_j \leftarrow \eta_j + \mathbf{1}_{\{j \in B^{(i)}\}}$ for each $j = 1, \ldots, p$.
6: **end for**
7: **return** $B^{(1)}, \ldots, B^{(m)}$.

B.2.3 Examples of (m, n)-Coverable Graphs

Example 1 (Time-inhomogeneous Autoregressive Models)**.** Consider a time-inhomogeneous Gaussian AR(r) model (assuming[1] $r \geq 1$), so that the sparsity pattern $E = \{(i, j) : 1 \leq |i-j| \leq r\}$. Suppose $n \geq 2+8r$. A simple choice of blocking sets is given as follows. Let $d = \lfloor (n-2)/8 \rfloor$, then $d \geq r$. Let $B_1 = [p] \cap \{kd + i : k \text{ odd, and } i = 1, \ldots, d\}$ and $B_2 = [p] \cap \{kd + i : k \text{ even, and } i = 1, \ldots, d\}$. Any connected component W of the subgraph that deletes B_1 is no larger than d and W's vertices V satisfy $|I_V \cap B_1| \leq 2r$, so G is $(2d + 2r + 1)$-separated by B_1 and $2d + 2r + 1 \leq n/2$. The same holds for B_2. Note that $[p] = B_1^c \cup B_2^c$, thus the graph is $(2, n)$-coverable.

Example 2 (d-dimensional Square-lattice Models)**.** Consider a finite subset of the d-dimensional lattice $\mathbb{Z}^d$ where pairs of vertices with distance 1 are adjacent. Suppose $n \geq 6 + 4d$, one could take B_1 as the grid points whose coordinates sum up to an odd number, and B_2 as the complement of B_1. The subgraph that deletes B_1 (or B_2) is isolated and each vertex has a neighborhood of size $2d$, so the graph is $(3 + 2d)$-separated by B_1 (or B_2). Since $3 + 2d \leq n/2$, the graph is $(2, n)$-coverable.

Example 3. Consider a m-colorable graph G. Let each of $V_1, \ldots, V_m$ be the vertex set of the same color. For any $i \in [m]$, the subgraph that deletes $B_i := \cup_{\ell \neq i} V_\ell$ is the subgraph that restricts on V_i, of which each vertex is isolated. Thus G is $(1 + 2 + \max_{v \in V_i} |I_v|)$-separated by B_i. If $n \geq \sum_{i \in [m]} (3 + \max_{v \in V_i} |I_v|)$, the graph is (m, n)-coverable. Note this subsumes Example 2 which has $m = 2$, but also applies to many other graphs such as forests, stars, and circles.

B.3 Discrete Graphical Models

B.3.1 Details about the Algorithms

We begin by proving the computational complexity of Algorithm 5. For each $j \in B^c$, enumerating all nonempty configurations of $\boldsymbol{k}_{I_j}$ takes no more than $\prod_{\ell \in I_j} K_\ell$ operations by checking each $\boldsymbol{k}_{I_j}$ or $n|I_j|$ operations by checking each observed X_{i,I_j}. The random permutation takes no more than n steps in total, so the overall complexity is $O\left(\sum_{j \in B^c} (n + \min(\prod_{\ell \in I_j} K_\ell, n|I_j|))\right)$.

[1] When $r = 0$, the graph is isolated and is $(1, n)$-coverable for any $n \geq 3$.

As mentioned at the beginning of Section 3.3, we can generate knockoffs without assuming the covariate categories being finite. First of all, with infinite K_ℓ's, Algorithm 5 can still be used since in Line 3 it is only needed to enumerate those $\boldsymbol{k}_{I_j}$ actually appearing in the observed data, which is at most n. Furthermore, the proof of Theorem 3.4 does not require the K_ℓ's to be finite.

B.3.2 Graph Expanding

As mentioned in Section 3.3.1, in Algorithm 5 variables in B are blocked and their knockoffs are trivial. One way to mitigate this drawback is to run multiple times of Algorithm 5 with expanded graphs that include the generated knockoff variables.

Specifically, denote by $\bar{G}$ a graph being augmented from G. For each $j \in B^c$, we add an edge between every pair of j's neighbors and add to $\bar{G}$ the 'knockoff vertex' $\tilde{j}$ which has the same neighborhood as j. One can show that $[\boldsymbol{X}, \tilde{\boldsymbol{X}}_{B^c}]$ is locally Markov w.r.t. the new graph. Applying Algorithm 5 to $[\boldsymbol{X}, \tilde{\boldsymbol{X}}_{B^c}]$ with graph $\bar{G}$ but with a different global cut set $\bar{B}$ which pre-includes B^c and also the knockoff vertices, we can generate knockoffs for some of the variables that have been blocked in the first run. One can continue to expand the graph to include the new knockoff variables, although the neighborhoods may become so large that the knockoff variables generated are constrained (through conditioning on these large neighborhoods) to be identical to their corresponding original variables. Algorithm 12 formally describes this process, whose validity is guaranteed by Theorem B.4.

Algorithm 12 Conditional Knockoffs for Discrete Graphical Models with Graph-Expanding

Input: $\boldsymbol{X} \in \mathbb{N}^{n\times p}$, $G = ([p], E)$, Q the maximum number of steps to expand the graph.

1: **Initialization**: the augmented graph $\bar{G} \leftarrow G$, whose vertex set is denoted by $\bar{V}$; $D \leftarrow \emptyset$ contains the variables that have knockoffs ; $q \leftarrow 1$ is the step of the graph expansion.
2: **while** $q \le Q$ and $|D| < p$ **do**
3: Find a global cut set B for $\bar{G}$ such that $B \supseteq D$.
4: Construct knockoffs $\boldsymbol{J}$ for $[\boldsymbol{X}, \tilde{\boldsymbol{X}}_{\bar{V}\setminus[p]}]$ w.r.t. the graph $\bar{G}$ and the cut set B by Algorithm 5.
5: Set $\tilde{\boldsymbol{X}}_{[p]\setminus B} = \boldsymbol{J}_{[p]\setminus B}$.
6: **for** $j \in B^c$ **do**
7: Add $\tilde{j}$ to $\bar{G}$ with the same neighborhood as j.
8: Add an edge between each pair of j's neighbors.
9: Update $D \leftarrow D \uplus \{j\}$.
10: **end for**
11: $q \leftarrow q + 1$.
12: **end while**
13: Set $\tilde{\boldsymbol{X}}_{[p]\setminus D} = \boldsymbol{X}_{[p]\setminus D}$.
14: **return** $\tilde{\boldsymbol{X}}$.

Theorem B.4. *Algorithm 12 generates valid knockoff for model* (3.4).

Proof of Theorem B.4. We will first define some notation to describe the process of graph-expanding, and then write down the joint probability mass function. Finally, we show the p.m.f. remains unchanged when swapping one variable, which suffices to prove the theorem by induction.

Part 1. To streamline the notation, we redefine Q as the number of steps that have actually been taken to expand the graph in Algorithm 5 (rather than the input value). For each $q \in [Q]$, denote by $G^{(q)}$ the augmented graph and by $B^{(q)}$ the blocking set used in the qth run of Algorithm 5. Let $V^{(q)} = [p] \setminus B^{(q)}$. Also denote by $D^{(q)}$ the variables for which knockoffs have already been generated before the qth run of Algorithm 5. Then $D^{(r)} = \bigcup_{q=1}^{r-1} V^{(q)}$ for any $r \geq 2$ and $D^{(1)} = \emptyset$. Let $I_j^{(q)}$ be the neighborhood of j in $G^{(q)}$. For ease of notation, we neglect to write the ranges of k_j and $\boldsymbol{k}_A$ when enumerating them in equations (e.g., taking a product over all their possible values). For a $n \times c$ matrix $\boldsymbol{Z}$ and integers $k_1, \ldots, k_c$, let $\varrho(\boldsymbol{Z}, k_1, \ldots, k_c) := \sum_{i=1}^{n} \mathbf{1}_{\{Z_{i,1}=k_1,\ldots,Z_{i,c}=k_c\}}$, i.e., the number of rows of $\boldsymbol{Z}$ that equal the vector $(k_1, \ldots, k_c)$.

Part 2. For any $q \in [Q]$ and $j \in V^{(q)}$, the neighborhood $I_j^{(q)}$ consists of three parts:

1. $([p] \cap I_j^{(q)}) \setminus D^{(q)}$: the neighbors in $[p]$ for which no knockoffs have been generated,
2. $I_j^{(q)} \cap D^{(q)}$: the neighbors in $[p]$ for which knockoffs have been generated, and
3. $\{\tilde{\ell} : \ell \in I_j^{(q)} \cap D^{(q)}\}$: the neighbors that are knockoffs.

The generation of $\tilde{\boldsymbol{X}}_j$ by Algorithm 5 is to sample uniformly from all vectors in $[K_j]^n$ such that the contingency table for variable j and its neighbors in $I_j^{(q)}$ remains the same if $\boldsymbol{X}_j$ is replaced by any of these vectors. Define

$$\begin{aligned}
&\mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{X}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{X}}_{I_j^{(q)}\cap D^{(q)}}\right) \\
&= \Big\{\mathbf{W}_j \in [K_j]^n : \varrho\left([\mathbf{W}_j, \mathbf{X}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{X}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{X}}_{I_j^{(q)}\cap D^{(q)}}], k_j, \boldsymbol{k}_{I_j^{(q)}}\right) = \\
&\qquad\qquad \varrho\left([\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{X}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{X}}_{I_j^{(q)}\cap D^{(q)}}], k_j, \boldsymbol{k}_{I_j^{(q)}}\right), \\
&\qquad\qquad \forall k_j, \boldsymbol{k}_{I_j^{(q)}}\Big\},
\end{aligned}$$

and then

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\boldsymbol{X}}_j = \tilde{\mathbf{x}}_j \,\middle|\, \boldsymbol{X} = \mathbf{X}, \tilde{\boldsymbol{X}}_{V^{(1)}} = \tilde{\mathbf{X}}_{V^{(1)}}, \ldots, \tilde{\boldsymbol{X}}_{V^{(q-1)}} = \tilde{\mathbf{X}}_{V^{(q-1)}}\right) \\
&= \frac{\mathbf{1}_{\left\{\tilde{\mathbf{x}}_j \in \mathcal{M}_j\left(\mathbf{x}_j, \mathbf{X}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{X}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{X}}_{I_j^{(q)}\cap D^{(q)}}\right)\right\}}}{\left|\mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{X}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{X}}_{I_j^{(q)}\cap D^{(q)}}\right)\right|}.
\end{aligned}$$

Denote the probability mass function of X by $f(\mathbf{x})$. The joint probability mass of the distribution

of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ is

$$
\begin{aligned}
&\mathbb{P}\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}] = [\mathbf{X}, \tilde{\mathbf{X}}]\right) \qquad \text{(B.10)}\\
=&\prod_{i=1}^{n} f(\mathbf{x}_i)\\
&\times \mathbb{P}\left(\tilde{\boldsymbol{X}}_{V^{(1)}} = \tilde{\mathbf{X}}_{V^{(1)}} \mid \boldsymbol{X} = \mathbf{X}\right)\\
&\times \prod_{q=2}^{Q} \mathbb{P}\left(\tilde{\boldsymbol{X}}_{V^{(q)}} = \tilde{\mathbf{X}}_{V^{(q)}} \;\middle|\; \boldsymbol{X} = \mathbf{X}, \tilde{\boldsymbol{X}}_{V^{(1)}} = \tilde{\mathbf{X}}_{V^{(1)}}, \ldots, \tilde{\boldsymbol{X}}_{V^{(q-1)}} = \tilde{\mathbf{X}}_{V^{(q-1)}}\right)\\
&\times \prod_{j \in [p] \setminus (\bigcup_{q=1}^{Q} V^{(q)})} \mathbf{1}_{\{\tilde{\mathbf{x}}_j = \mathbf{x}_j\}}\\
=&\prod_{i=1}^{n} f(\mathbf{x}_i)\\
&\times \prod_{q=1}^{Q} \prod_{j \in V^{(q)}} \frac{\mathbf{1}_{\left\{\tilde{\mathbf{x}}_j \in \mathcal{M}_j\left(\mathbf{x}_j, \mathbf{x}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{x}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{x}}_{I_j^{(q)}\cap D^{(q)}}\right)\right\}}}{|\mathcal{M}_j\left(\mathbf{x}_j, \mathbf{x}_{([p]\cap I_j^{(q)})\setminus D^{(q)}}, \mathbf{x}_{I_j^{(q)}\cap D^{(q)}}, \tilde{\mathbf{x}}_{I_j^{(q)}\cap D^{(q)}}\right)|}\\
&\times \prod_{j \in [p] \setminus (\bigcup_{q=1}^{Q} V^{(q)})} \mathbf{1}_{\{\tilde{\mathbf{x}}_j = \mathbf{x}_j\}},
\end{aligned}
$$

where the product is partitioned into three parts: the distribution of $\boldsymbol{X}$, the distributions of the knockoff columns generated in each step and the indicator functions for the variables that have no knockoffs generated within the Q steps.

Part 3. It suffices to show that for any $\ell \in [p]$,

$$
\mathbb{P}\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}] = [\mathbf{X}, \tilde{\mathbf{X}}]\right) = \mathbb{P}\left([\boldsymbol{X}, \tilde{\boldsymbol{X}}] = [\mathbf{X}, \tilde{\mathbf{X}}]_{\mathrm{swap}(\ell)}\right). \qquad \text{(B.11)}
$$

If both sides of Equation (B.11) equal zero, it holds trivially. Without loss of generality, we will prove this equation under the assumption that the left hand side is non-zero. One can redefine $[\mathbf{X}', \tilde{\mathbf{X}}'] = [\mathbf{X}, \tilde{\mathbf{X}}]_{\mathrm{swap}(\ell)}$ and apply the same proof when assuming the right hand side is non-zero.

First, suppose $\ell \in [p] \setminus (\bigcup_{q=1}^{Q} V^{(q)})$. Since the left hand side is non-zero, by Equation (B.10), $\tilde{\mathbf{X}}_\ell = \mathbf{X}_\ell$ and the p.m.f. does not change when swapping $\mathbf{X}_\ell$ with $\tilde{\mathbf{X}}_\ell$.

Second, suppose $\ell \in V^{(q_\ell)}$ for some $q_\ell \in [Q]$. Since the left hand side of Equation (B.11) is non-zero, then in Equation (B.10), the indicator function in the second part with $q = q_\ell$ and $j = \ell$ being non-zero indicates

$$
\begin{aligned}
&\varrho\left([\tilde{\mathbf{X}}_\ell, \mathbf{X}_{([p]\cap I_\ell^{(q_\ell)})\setminus D^{(q_\ell)}}, \mathbf{X}_{I_\ell^{(q_\ell)}\cap D^{(q_\ell)}}, \tilde{\mathbf{X}}_{I_\ell^{(q_\ell)}\cap D^{(q_\ell)}}], k_\ell, \boldsymbol{k}_{I_\ell^{(q_\ell)}}\right)\\
=&\varrho\left([\mathbf{X}_\ell, \mathbf{X}_{([p]\cap I_\ell^{(q_\ell)})\setminus D^{(q_\ell)}}, \mathbf{X}_{I_\ell^{(q_\ell)}\cap D^{(q_\ell)}}, \tilde{\mathbf{X}}_{I_\ell^{(q_\ell)}\cap D^{(q_\ell)}}], k_\ell, \boldsymbol{k}_{I_\ell^{(q_\ell)}}\right), \forall k_\ell, \boldsymbol{k}_{I_\ell^{(q_\ell)}}. \qquad \text{(B.12)}
\end{aligned}
$$

The only difference between the two sides of Equation B.12 is that $\tilde{\mathbf{X}}_\ell$ is replaced by $\mathbf{X}_\ell$ in the first columns of the matrix. Such a difference will keep appearing in the equations that will be showed below.

In the following, we show that everywhere $\mathbf{X}_\ell$ or $\tilde{\mathbf{X}}_\ell$ appears in the first or second part of the product in Equation (B.10) remains unchanged when swapping $\mathbf{X}_\ell$ and $\tilde{\mathbf{X}}_\ell$.

1. As shown in Section 3.3, there exist some functions ψ_ℓ and $\theta_\ell(k_\ell, \boldsymbol{k}_{I_\ell})$'s such that

$$\prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} \psi_\ell(\mathbf{x}_{i,[p]\setminus\{\ell\}}) \prod_{k_\ell, \boldsymbol{k}_{I_\ell}} \theta_\ell(k_\ell, \boldsymbol{k}_{I_\ell}) \varrho\left([\mathbf{X}_\ell, \mathbf{X}_{I_\ell}], k_\ell, \boldsymbol{k}_{I_\ell}\right)$$

   Note that the initial neighborhood $I_\ell \subseteq I_\ell^{(q_\ell)}$, by summing over Equation (B.12), one can conclude that

$$\varrho\left([\mathbf{X}_\ell, \mathbf{X}_{I_\ell}], k_\ell, \boldsymbol{k}_{I_\ell}\right) = \varrho\left([\tilde{\mathbf{X}}_\ell, \mathbf{X}_{I_\ell}], k_\ell, \boldsymbol{k}_{I_\ell}\right)$$

   for all $k_\ell, \boldsymbol{k}_{I_\ell}$. Thus $\prod_{i=1}^{n} f(\mathbf{x}_i)$ remains unchanged when swapping $\mathbf{X}_\ell$ and $\tilde{\mathbf{X}}_\ell$.

2. For any j such that $j \in V^{(q_j)}$ and $\ell \in I_j^{(q_j)}$, the second part of the product in Equation B.10 involves $\mathbf{X}_\ell$ or $\tilde{\mathbf{X}}_\ell$ with the indices q_j and j. Since $B^{(q_j)}$ is a blocking set, we have $q_j \neq q_\ell$.

   (a) If $q_j > q_\ell$, then $\ell \in I_j^{(q_j)} \cap D^{(q_j)}$. Note that swapping $\mathbf{X}_\ell$ with $\tilde{\mathbf{X}}_\ell$ only changes the order of the dimensions of the contingency table formed by $\mathbf{X}_j$ and $\left[\mathbf{X}_{[p]\cap I_j^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right]$, and thus does not change whether or not the following system of equations (where only the first column of the first argument of ϱ differs between the two lines) holds

$$\begin{aligned}
&\varrho\left([\mathbf{W}_j, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, \boldsymbol{k}_{I_j^{(q_j)}}\right) \\
=&\varrho\left([\mathbf{X}_j,\ \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, \boldsymbol{k}_{I_j^{(q_j)}}\right), \\
&\forall k_j, \boldsymbol{k}_{I_j^{(q_j)}}.
\end{aligned}$$

   Thus the indicator function

$$\mathbf{1}_{\left\{\tilde{\mathbf{x}}_j \in \mathcal{M}_j\left(\mathbf{x}_j, \mathbf{x}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{x}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{x}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)\right\}}$$

   and the cardinal number

$$\left|\mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)\right|$$

   from Equation (B.10) remain unchanged when swapping $\mathbf{X}_\ell$ and $\tilde{\mathbf{X}}_\ell$.

   (b) Now we show the same conclusion for the remaining j values, which will require a few intermediate steps. If $q_j < q_\ell$, then $\ell \in I_j^{(q_j)} \setminus D^{(q_j)}$ and $j \in I_\ell^{(q_\ell)} \cap D^{(q_\ell)}$. By the graph expanding algorithm, we have $I_j^{(q_j)} \setminus \{\ell\} \subseteq I_\ell^{(q_\ell)}$, and $D^{(q_j)} \subseteq D^{(q_\ell)}$. This shows

$$([p] \cap I_j^{(q_j)}) \setminus (\{\ell\} \cup D^{(q_j)}) \subseteq ([p] \cap I_\ell^{(q_\ell)} \setminus D^{(q_\ell)}) \uplus (I_\ell^{(q_\ell)} \cap D^{(q_\ell)}),$$

and

$$I_j^{(q_j)} \cap D^{(q_j)} \subseteq I_\ell^{(q_\ell)} \cap D^{(q_\ell)}.$$

Summing over Equation (B.12) and rearranging the columns of the first argument of ϱ, one can conclude that

$$\begin{aligned} &\varrho\left([\mathbf{X}_j, \tilde{\mathbf{X}}_j, \tilde{\mathbf{X}}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right) \\ &= \varrho\left([\mathbf{X}_j, \tilde{\mathbf{X}}_j, \mathbf{X}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right), \\ &\qquad \forall k_j, k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}, \end{aligned} \tag{B.13}$$

where the two lines only differ in the third column of the first argument of ϱ. Summing over Equation (B.13) w.r.t. $k_{\tilde{j}}$, one can further conclude that

$$\begin{aligned} &\varrho\left([\mathbf{X}_j, \tilde{\mathbf{X}}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right) \\ &= \varrho\left([\mathbf{X}_j, \mathbf{X}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right), \\ &\quad \forall k_j, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}, \end{aligned} \tag{B.14}$$

where the two lines only differ in the second column of the first argument of ϱ, and the first column is $\mathbf{X}_j$. Similarly, summing over Equation (B.13) w.r.t. k_j, we have

$$\begin{aligned} &\varrho\left([\tilde{\mathbf{X}}_j, \tilde{\mathbf{X}}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right) \\ &= \varrho\left([\tilde{\mathbf{X}}_j, \mathbf{X}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right), \\ &\quad \forall k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}, \end{aligned} \tag{B.15}$$

where again the two lines only differ in the second column of the first argument of ϱ, but now the first column is $\tilde{\mathbf{X}}_j$.

Note that $\left|\mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)\right|$ is a product of some multinomial coefficients, and each multinomial coefficient depends on a unique combination of $(k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}})$ and the values of

$$\varrho\left([\mathbf{X}_j, \mathbf{X}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}], k_j, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}\right), \forall k_j.$$

These quantities are the ones on the right hand side of Equation (B.14). Thus we conclude that $\left|\mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)\right|$ remains unchanged when swapping $\mathbf{X}_\ell$ and $\tilde{\mathbf{X}}_\ell$ by checking the terms in Equation (B.14) that appear in the multinomial coefficients.

Note that $\tilde{\mathbf{X}}_j \in \mathcal{M}_j\left(\mathbf{X}_j, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)$ if and only if the right hand sides of Equations (B.14) and (B.15) are equal for all $k_{\tilde{j}}, k_\ell, \boldsymbol{k}_{I_j^{(q_j)}\setminus\{\ell\}}$, which is equivalent to the left hand sides of the equations are equal, which holds if and only if $\tilde{\mathbf{X}}_j \in \mathcal{M}_j\left(\mathbf{X}_j, \tilde{\mathbf{X}}_\ell, \mathbf{X}_{([p]\cap I_j^{(q_j)})\setminus(\{\ell\}\cup D^{(q_j)})}, \mathbf{X}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{X}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)$. Therefore the indicator function $\mathbf{1}_{\left\{\tilde{\mathbf{x}}_j \in \mathcal{M}_j\left(\mathbf{x}_j, \mathbf{x}_{([p]\cap I_j^{(q_j)})\setminus D^{(q_j)}}, \mathbf{x}_{I_j^{(q_j)}\cap D^{(q_j)}}, \tilde{\mathbf{x}}_{I_j^{(q_j)}\cap D^{(q_j)}}\right)\right\}}$ remains unchanged when swapping $\mathbf{X}_\ell$ and $\tilde{\mathbf{X}}_\ell$.

To sum up, Equation (B.11) holds for any $\ell \in [p]$, and the proof is complete. □

B.4 Alternative Knockoff Generations for Discrete Markov Chains

We provide alternative constructions of conditional knockoffs for Markov chains that make use of the simple chain structure. Proofs in this section are deferred to Appendix B.4.3.

We introduce a notation that makes the display clear. For any $n \times c$ matrix $\boldsymbol{Z}$ and any integers $k_1, \ldots, k_c$, let $\varrho(\boldsymbol{Z}, k_1, \ldots, k_c) := \sum_{i=1}^n \mathbf{1}_{\{Z_{i,1}=k_1,\ldots,Z_{i,c}=k_c\}}$, i.e., the number of rows of $\boldsymbol{Z}$ that equal the vector $(k_1, \ldots, k_c)$.

Suppose the components of X follow a general discrete Markov chain, then the joint distribution for n i.i.d. samples is

$$\mathbb{P}(\boldsymbol{X}) = \prod_{k=1}^{K_1} (\pi_k^{(1)})^{\sum_{k'=1}^{K_2} N_{k,k'}^{(2)}} \prod_{j=2}^{p} \prod_{k=1}^{K_{j-1}} \prod_{k'=1}^{K_j} (\pi_{k,k'}^{(j)})^{N_{k,k'}^{(j)}},$$

where $\pi_k^{(1)} = \mathbb{P}(X_{i,1} = k)$ and $\pi_{k,k'}^{(j)} = \mathbb{P}(X_{i,j} = k' \mid X_{i,j-1} = k)$ are model parameters and $N_{k,k'}^{(j)} = \varrho([\boldsymbol{X}_{j-1}, \boldsymbol{X}_j], k, k')$ is the number of samples such that the $(j-1)$th and jth components are k and k', respectively. So all the $N_{k,k'}^{(j)}$'s together form a sufficient statistic which we denote by $T(\boldsymbol{X})$, and although it has some redundant entries (for example, $\sum_{k=1}^{K_{j-1}} \sum_{k'=1}^{K_j} N_{k,k'}^{(j)} = n$), it is nevertheless minimal, and we prefer to keep the redundant entries for notational convenience.

Conditional on $T(\boldsymbol{X})$, $\boldsymbol{X}$ is uniformly distributed on $\mathcal{Q} := \{\mathbf{W} \in \prod_{j=1}^p [K_j]^n : T(\mathbf{W}) = T(\boldsymbol{X})\}$. Hereafter, we distinguish notationally between $\mathbf{X}$'s and $\mathbf{W}$'s (and $\tilde{\mathbf{W}}$'s), with the former denoting realized values of the data in $\boldsymbol{X}$ and the latter denoting hypothetical such values not necessarily observed in the data. The conditional distribution of $\boldsymbol{X}$ can be decomposed as

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} = \mathbf{W} \mid T(\boldsymbol{X})) &= C_0 \prod_{j=2}^{p} \left(\prod_{k=1}^{K_{j-1}} \prod_{k'=1}^{K_j} \mathbf{1}_{\left\{\varrho([\mathbf{W}_{j-1}, \mathbf{W}_j], k, k') = N_{k,k'}^{(j)}\right\}} \right) \\ &= C_0 \prod_{j=2}^{p} \phi_j(\mathbf{W}_{j-1}, \mathbf{W}_j \mid T(\boldsymbol{X})), \end{aligned} \tag{B.16}$$

where C_0 only depends on $T(\boldsymbol{X})$. This decomposition implies that conditional on $T(\boldsymbol{X})$, the columns of $\boldsymbol{X}$ still comprise a vector-valued Markov chain (see Appendix B.4.3). For ease of notation, in what follows we will write $\mathbb{P}(\cdot)$ without '$\mid T(\boldsymbol{X})$' since we always condition on $T(\boldsymbol{X})$ in this section.

B.4.1 SCIP

The sequential conditional independent pairs (SCIP) algorithm from Candès et al. (2018) was introduced in completely general form for any distribution for X with the substantial caveat that actually carrying it out for any given distribution can be quite challenging. Sesia et al. (2018) show how to run SCIP for Markov chains *unconditionally*. When applied to vectors instead of scalars, SCIP can also be adapted to generate *conditional* knockoffs for Markov chains because the conditional distribution of $\boldsymbol{X}$ is uniform on $\mathcal{Q}$, making it a Markov chain, and conditional knockoffs are simply knockoffs for this conditional distribution.

SCIP sequentially samples $\tilde{\boldsymbol{X}}_j \sim \mathcal{L}(\boldsymbol{X}_j|\boldsymbol{X}_{\text{-}j}, \tilde{\boldsymbol{X}}_{1:(j-1)})$[1] independently of $\boldsymbol{X}_j$, for $j = 1, \ldots, p$. For a Markov chain, this sampling can be reduced to $\tilde{\boldsymbol{X}}_j \sim \mathcal{L}(\boldsymbol{X}_j|\boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \tilde{\boldsymbol{X}}_{j-1})$. The main computational challenge is to keep track of the following conditional probabilities:

$$
\begin{aligned}
f_1(\mathbf{W}_1, \mathbf{W}_2) &:= \mathbb{P}\left(\boldsymbol{X}_1 = \mathbf{W}_1 \mid \boldsymbol{X}_2 = \mathbf{W}_2\right), \\
f_j(\mathbf{W}_j, \tilde{\mathbf{W}}_{j-1}, \mathbf{W}_{j+1}) &:= \mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{j-1} = \mathbf{X}_{j-1}, \tilde{\boldsymbol{X}}_{j-1} = \tilde{\mathbf{W}}_{j-1}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}\right), \\
&\qquad\qquad \forall\, j \in \{2, \ldots, p-1\}, \\
f_p(\mathbf{W}_p, \tilde{\mathbf{W}}_{p-1}) &:= \mathbb{P}\left(\boldsymbol{X}_p = \mathbf{W}_p \,\middle|\, \boldsymbol{X}_{p-1} = \mathbf{X}_{p-1}, \tilde{\boldsymbol{X}}_{p-1} = \tilde{\mathbf{W}}_{p-1}\right).
\end{aligned}
$$

Algorithm 13 describes how to generate conditional knockoffs for a discrete Markov Chain with finite states by SCIP, where the functions $f_j(\cdot)$ are computed recursively by the formulas in Proposition B.5. These formulas are different from the ones in Sesia et al. (2018, Proposition 1), in which the authors assume transition probabilities can be evaluated directly.

Algorithm 13 Conditional Knockoffs for Discrete Markov Chains by SCIP

Input: $\boldsymbol{X} \in \prod_{j=1}^{p} [K_j]^n$.

1: Sample $\tilde{\boldsymbol{X}}_1$ uniformly from $\{\mathbf{W}_1 \in [K_1]^n : (\mathbf{W}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) \in \mathcal{Q}\}$.
2: Compute $f_1(\tilde{\boldsymbol{X}}_1, \mathbf{W}_2)$ for all $\mathbf{W}_2 \in [K_2]^n$.
3: **for** j from 2 to $p-1$ **do**
4: Compute $f_j(\mathbf{W}_j, \tilde{\boldsymbol{X}}_{j-1}, \mathbf{W}_{j+1})$ for all $\mathbf{W}_j \in [K_j]^n$ and $\mathbf{W}_{j+1} \in [K_{j+1}]^n$.
5: Sample $\tilde{\boldsymbol{X}}_j$ from $f_j(\cdot, \tilde{\boldsymbol{X}}_{j-1}, \boldsymbol{X}_{j+1})$.
6: **end for**
7: Compute $f_p(\mathbf{W}_j, \tilde{\boldsymbol{X}}_{p-1})$ for all $\mathbf{W}_p \in [K_p]^n$.
8: Sample $\tilde{\boldsymbol{X}}_p$ from $f_p(\cdot, \tilde{\boldsymbol{X}}_{p-1})$.
9: **return** $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_p]$.

Proposition B.5. *Define* $\frac{0}{0} = 0$. *We formally write* $f_1(\mathbf{W}_1, \tilde{\mathbf{X}}_0, \mathbf{W}_2)$ *for* $f_1(\mathbf{W}_1, \mathbf{W}_2)$. *Suppose* $\tilde{\mathbf{X}}$ *is a realization of* $\tilde{\boldsymbol{X}}$ *generated by Algorithm 2. Then the following equations hold*

$$
f_1(\mathbf{W}_1, \mathbf{W}_2) = \frac{\mathbf{1}_{\{(\mathbf{W}_1, \mathbf{W}_2, \mathbf{X}_3, \ldots, \mathbf{X}_p) \in \mathcal{Q}\}}}{\#\{\mathbf{W}_1' \in [K_1]^n : (\mathbf{W}_1', \mathbf{W}_2, \mathbf{X}_3, \ldots, \mathbf{X}_p) \in \mathcal{Q}\}},
$$

$$
\forall\, 1 < j < p,\ f_j(\mathbf{W}_j, \tilde{\mathbf{X}}_{j-1}, \mathbf{W}_{j+1}) = \frac{\mathbf{1}_{\{(\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{W}_j, \mathbf{W}_{j+1}, \ldots, \mathbf{X}_p) \in \mathcal{Q}\}} f_{j-1}(\tilde{\mathbf{X}}_{j-1}, \tilde{\mathbf{X}}_{j-2}, \mathbf{W}_j)}{\sum_{\mathbf{W}_j' \in [K_j]^n} \mathbf{1}_{\left\{(\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{W}_j', \mathbf{W}_{j+1}, \ldots, \mathbf{X}_p) \in \mathcal{Q}\right\}} f_{j-1}(\tilde{\mathbf{X}}_{j-1}, \tilde{\mathbf{X}}_{j-2}, \mathbf{W}_j')},
$$

[1] $\mathcal{L}(\boldsymbol{X}_j|\boldsymbol{X}_{\text{-}j}, \tilde{\boldsymbol{X}}_{1:(j-1)})$ is the conditional distribution of $\boldsymbol{X}_j$ given $(\boldsymbol{X}_{\text{-}j}, \tilde{\boldsymbol{X}}_{1:(j-1)})$.

$$f_p(\mathbf{W}_p, \tilde{\mathbf{X}}_{p-1}) = \frac{\mathbf{1}_{\{(\mathbf{x}_1,\ldots,\mathbf{x}_{p-1},\mathbf{w}_p)\in\mathcal{Q}\}} f_{p-1}(\tilde{\mathbf{X}}_{p-1}, \tilde{\mathbf{X}}_{p-2}, \mathbf{W}_p)}{\sum_{\mathbf{W}'_p\in[K_p]^n} \mathbf{1}_{\{(\mathbf{x}_1,\ldots,\mathbf{x}_{p-1},\mathbf{W}'_p)\in\mathcal{Q}\}} f_{p-1}(\tilde{\mathbf{X}}_{p-1}, \tilde{\mathbf{X}}_{p-2}, \mathbf{W}'_p)}.$$

Computing $f_j(\mathbf{W}_j, \tilde{\mathbf{X}}_{j-1}, \mathbf{W}_{j+1})$ in Proposition B.5 requires enumerating all possible configurations of $\mathbf{W}_j \in [K_j]^n$ and $\mathbf{W}_{j+1} \in [K_{j+1}]^n$, making the total computational complexity of SCIP $O(\sum_{j\le p-1}(K_jK_{j+1})^n)$. Due to the n in the exponent, SCIP quickly becomes intractable as the sample size grows, even for binary states and $n \gtrsim 10$. A simple remedy is to first randomly divide the rows of $\boldsymbol{X}$ into disjoint folds of small size around n_0, say $n_0 = 10$, and then run SCIP for each fold separately. This construction conditions on a statistic which is n/n_0 times as large as that conditioned on before dividing into folds, but the former's computation time scales linearly with n, instead of exponentially. Conditioning on more should tend to degrade the quality of the knockoffs, but is necessary to enable computation. Still, compared to Algorithm 6, SCIP does not block any variables and thus has the potential to generate better knockoffs.

B.4.2 Refined Blocking

One can apply Algorithm 6 to Markov Chains, as a 2-colorable graph, with two blocking sets, one with all even numbers and the other with all odd numbers. Instead of running Algorithm 5 in Line 3, a refined blocking algorithm, Algorithm 14, can be used for $1 < j < p$. It introduces more variability in the knockoff generation because it first draws a new contingency table $\tilde{\boldsymbol{H}}$ that is conditionally exchangeable with the observed contingency table $\boldsymbol{H}$ of $(\boldsymbol{X}_{j-1}, \boldsymbol{X}_j, \boldsymbol{X}_{j+1})$, and then samples $\tilde{\boldsymbol{X}}_j$ given $\tilde{\boldsymbol{H}}$. This algorithm constructs a reversible Markov Chain by proposing random walks on the space of contingency tables, moved by $\Delta\boldsymbol{H}$ and corrected by acceptance ratio α. In the following, we discuss how to sample $\Delta\boldsymbol{H}$ and compute α, and provide a detailed version of Algorithm 14 at the end of this section.

Algorithm 14 Improved Conditional Knockoffs for $\boldsymbol{X}_j$ in a Discrete Markov Chain (Pseudocode)

Input: $\boldsymbol{X}_j$, columns to condition on: $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$.

1: Initialize a chain of contingency table $\tilde{\boldsymbol{H}} = \boldsymbol{H}(\boldsymbol{X}_j)$.
2: **for** t from 1 to $t_{\max}$ **do**
3: Draw $\Delta\boldsymbol{H}$ independent of $\tilde{\boldsymbol{H}}$, and propose $\tilde{\boldsymbol{H}}^* = \tilde{\boldsymbol{H}} + \Delta\boldsymbol{H}$.
4: **if** all elements of $\tilde{\boldsymbol{H}}^*$ are nonnegative **then**
5: Calculate the Metropolis–Hastings acceptance ratio α.
6: With probability α, update $\tilde{\boldsymbol{H}} \leftarrow \tilde{\boldsymbol{H}}^*$.
7: **end if**
8: **end for**
9: Sample $\tilde{\boldsymbol{X}}_j$ uniformly on all possible vectors that match $\tilde{\boldsymbol{H}}$.
10: **return** $\tilde{\boldsymbol{X}}_j$.

Conditional on $T(\boldsymbol{X})$ and $\boldsymbol{X}_B$, $\boldsymbol{X}_j$ is uniformly distributed on all $\mathbf{W}_j \in [K_j]^n$ such that

$$\varrho\left([\boldsymbol{X}_{j-1}, \mathbf{W}_j], k_{j-1}, k_j\right) = N^{(j)}_{k_{j-1},k_j} \text{ and } \varrho\left([\mathbf{W}_j, \boldsymbol{X}_{j+1}], k_j, k_{j+1}\right) = N^{(j+1)}_{k_j,k_{j+1}}$$

for all (k_{j-1}, k_j, k_{j+1}). In the following, we view $T(\boldsymbol{X})$ and $\boldsymbol{X}_B$ as fixed and only $\boldsymbol{X}_j$ as being random.

We begin with some notation. To avoid burdensome subscripts, we write (k_-, k, k_+) for (k_{j-1}, k_j, k_{j+1}). Let $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{X}_j)$ be the three-dimensional array with elements $\boldsymbol{H}_{k_-,k,k_+} := \varrho\left(\boldsymbol{X}_{(j-1):(j+1)}, k_-, k, k_+\right)$ for all (k_-, k, k_+). The statistic $\boldsymbol{H}$ is essentially a three-way contingency table and its three-dimensional marginals satisfy

$$\begin{aligned}
&\sum_{k_-\in[K_-]} \boldsymbol{H}_{k_-,k,k_+} = N^{(j+1)}_{k,k_+}, \quad \forall\, k, k_+; \qquad \sum_{k_+\in[K_+]} \boldsymbol{H}_{k_-,k,k_+} = N^{(j)}_{k_-,k}, \quad \forall\, k_-, k;\\
&\sum_{k\in[K]} \boldsymbol{H}_{k_-,k,k_+} = M_{k_-,k_+}, \qquad \text{where} \quad M_{k_-,k_+} := \varrho\left([\boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}], k_-, k_+\right), \quad \forall\, k_-, k_+.
\end{aligned}$$

Here M_{k_-,k_+} is a function of $\boldsymbol{X}_B$ and thus fixed. Conditional on $\boldsymbol{H}$, $\boldsymbol{X}_j$ is uniform on all vectors in $[K_j]^n$ that match the three-way contingency table. The probability function for $\boldsymbol{H}$ is

$$\mathbb{P}\left(\forall\, k_-, k, k_+, \boldsymbol{H}_{k_-,k,k_+} = \mathsf{H}_{k_-,k,k_+} \mid \boldsymbol{X}_B\right) \propto \prod_{k_-\in[K_{j-1}], k_+\in[K_{j+1}]} \binom{M_{k_-,k_+}}{\mathsf{H}_{k_-,1,k_+}, \ldots, \mathsf{H}_{k_-,K_j,k_+}}, \quad \text{(B.17)}$$

where the counts H_{k_-,k,k_+} satisfy $\sum_{k\in[K_j]} \mathsf{H}_{k_-,k,k_+} = M_{k_-,k_+}$ for each pair of $(k_-, k_+) \in [K_{j-1}] \times [K_{j+1}]$.

The construction of the Markov chain on contingency tables begins with defining the basic moves: suppose there are L different three-way tables $\{\boldsymbol{\Delta}^{(\ell)}\}_{\ell=1}^L \subseteq \mathbb{Z}^{K_{j-1}\times K_j \times K_{j+1}}$ such that the marginals of each table $\boldsymbol{\Delta}^{(\ell)}$ are 0's:[1]

$$\begin{aligned}
&\forall\, k, k_+, \sum_{k_-} \boldsymbol{\Delta}^{(\ell)}_{k_-,k,k_+} = 0, \qquad \forall\, k_-, k, \sum_{k_+} \boldsymbol{\Delta}^{(\ell)}_{k_-,k,k_+} = 0,\\
&\forall\, k_-, k_+, \sum_{k} \boldsymbol{\Delta}^{(\ell)}_{k_-,k,k_+} = 0.
\end{aligned}$$

A simple set of basic moves, indexed by $\boldsymbol{\ell}$, can be constructed as follows: for each $\boldsymbol{\ell} = (r_1, r_2, c_1, c_2, d_1, d_2)$ where $r_1, r_2 \in [K_{j-1}]$, $c_1, c_2 \in [K_{j+1}]$, $d_1, d_2 \in [K_j]$ and $r_1 \neq r_2$, $c_1 \neq c_2$, $d_1 \neq d_2$, define

$$\boldsymbol{\Delta}^{(\boldsymbol{\ell})}_{k_-,k,k_+} = \begin{cases} (-1)^{\mathbf{1}_{\{k_-=r_1\}}+\mathbf{1}_{\{k_+=c_1\}}+\mathbf{1}_{\{k=d_1\}}}, & \text{if } k_- \in \{r_1, r_2\}, k_+ \in \{c_1, c_2\}, z_j \in \{d_1, d_2\} \\ 0, & \text{otherwise} \end{cases}$$

Algorithm 15 is a detailed sampling procedure, whose validity is guaranteed by Proposition B.6.

Proposition B.6. *For $j \in B^c$, if $\tilde{\boldsymbol{X}}_j$ is drawn from Algorithm 15, then*

$$(\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{X}}_j, \boldsymbol{X}_j) \mid \boldsymbol{X}_B.$$

A final remark is that one can generalize the refined blocking algorithm to Ising models. By Equation (3.5), the sufficient statistic is the vector that includes all the counts of configurations of adjacent pairs, i.e., $\sum_{i=1}^n \mathbf{1}_{\{X_{i,s}=k, X_{i,t}=k'\}}$ for all $k, k' \in \{-1, 1\}$ and $(s, t) \in E$. Instead of sampling a three-way contingency table in Algorithm 15, now one has to construct a reversible Markov Chain for the $(2d+1)$-way contingency table for each vertex and its neighborhood. The basic moves can be constructed similarly as the $\Delta^{(\boldsymbol{\ell})}$ given before.

[1]The set $\{\boldsymbol{\Delta}^{(\ell)}\}_{\ell=1}^L$ is similar to the Markov bases used in algebraic statistics (see Diaconis et al. (1998)), but it does not need to connect every two possible contingency tables.

Algorithm 15 Improved Conditional Knockoffs for $\boldsymbol{X}_j$ in a Discrete Markov Chain

Input: $\boldsymbol{X}_j$, columns to condition on: $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$.

Initialize $\tilde{\boldsymbol{H}}^0$ with $\tilde{\boldsymbol{H}}^0_{k_{j-1},k_j,k_{j+1}} := \sum_{i=1}^n \mathbf{1}_{\{X_{i,j-1}=k_{j-1},X_{i,j}=k_j,X_{i,j+1}=k_{j+1}\}}$ for all k_{j-1}, k_j and k_{j+1}.
for t from 1 to $t_{\max}$ **do**
  Sample uniformly without replacement the pair (R_1, R_2) from $\{1, \dots, K_{j-1}\}$, the pair (C_1, C_2) from $\{1, \dots, K_{j+1}\}$, and the pair (D_1, D_2) from $\{1, \dots, K_j\}$.
  Define $\tilde{\boldsymbol{H}}^* = \tilde{\boldsymbol{H}}^{t-1} + \boldsymbol{\Delta}^{(R_1,R_2,C_1,C_2,D_1,D_2)}$.
  if all elements of $\tilde{\boldsymbol{H}}^*$ are nonnegative **then**
    Calculate the Metropolis-Hastings acceptance ratio

$$\alpha = \min\left(\frac{\prod_{k_{j-1},k_{j+1}} \binom{M_{k_{j-1},k_{j+1}}}{\tilde{\boldsymbol{H}}^*_{k_{j-1},1,k_{j+1}},\dots,\tilde{\boldsymbol{H}}^*_{k_{j-1},K_j,k_{j+1}}}}{\prod_{k_{j-1},k_{j+1}} \binom{M_{k_{j-1},k_{j+1}}}{\tilde{\boldsymbol{H}}^{t-1}_{k_{j-1},1,k_{j+1}},\dots,\tilde{\boldsymbol{H}}^{t-1}_{k_{j-1},K_j,k_{j+1}}}}, 1\right).$$

    Sample $U \sim \text{Unif}(0,1)$.
    if $U \le \alpha$ **then**
      Set $\tilde{\boldsymbol{H}}^t = \tilde{\boldsymbol{H}}^*$.
    else
      Set $\tilde{\boldsymbol{H}}^t = \tilde{\boldsymbol{H}}^{t-1}$.
    end if
  else
    Set $\tilde{\boldsymbol{H}}^t = \tilde{\boldsymbol{H}}^{t-1}$.
  end if
end for
Set $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{H}}^{t_{\max}}$.
Sample $\tilde{\boldsymbol{X}}_j$ uniformly on all possible vectors that match $\tilde{\boldsymbol{H}}$. In other words, for each $(k_{j-1}, k_{j+1}) \in [K_{j-1}] \times [K_{j+1}]$, set the subvector of $\tilde{\boldsymbol{X}}_j$ with indices $\{i \in [n] : X_{i,j-1} = k_{j-1}, X_{i,j+1}(i) = k_{j+1}\}$ as a random uniform permutation of

$$\Big(\underbrace{1, \dots, 1}_{\tilde{\boldsymbol{H}}(k_{j-1},1,k_{j+1})}, \dots, \underbrace{K_j, \dots, K_j}_{\tilde{\boldsymbol{H}}(k_{j-1},K_j,k_{j+1})} \Big).$$

return $\tilde{\boldsymbol{X}}_j$.

Remark 5. The calculation of α is simply

$$\prod_{k_{j-1}\in\{R_1,R_2\},k_{j+1}\in\{C_1,C_2\}} \frac{\tilde{\boldsymbol{H}}^{t-1}_{k_{j-1},D_1,k_{j+1}}!}{\tilde{\boldsymbol{H}}^*_{k_{j-1},D_1,k_{j+1}}!} \frac{\tilde{\boldsymbol{H}}^{t-1}_{k_{j-1},D_2,k_{j+1}}!}{\tilde{\boldsymbol{H}}^*_{k_{j-1},D_2,k_{j+1}}!}$$
$$= \frac{\tilde{\boldsymbol{H}}^{t-1}_{R_1,D_1,C_1}}{\tilde{\boldsymbol{H}}^{t-1}_{R_1,D_2,C_1}+1} \times \frac{\tilde{\boldsymbol{H}}^{t-1}_{R_1,D_2,C_2}}{\tilde{\boldsymbol{H}}^{t-1}_{R_1,D_1,C_2}+1} \times \frac{\tilde{\boldsymbol{H}}^{t-1}_{R_2,D_2,C_1}}{\tilde{\boldsymbol{H}}^{t-1}_{R_2,D_1,C_1}+1} \times \frac{\tilde{\boldsymbol{H}}^{t-1}_{R_2,D_1,C_2}}{\tilde{\boldsymbol{H}}^{t-1}_{R_2,D_2,C_2}+1},$$

where all quantities can be read off directly from $\tilde{\boldsymbol{H}}^{t-1}$.

B.4.3 Proofs

Conditional Markov Chains We first show that conditional on $T(\boldsymbol{X})$, the sequence of $\boldsymbol{X}_j$'s forms a Markov chain, i.e., Equation (B.16) describes a Markov chain. We still write '$\mid T(\boldsymbol{X})$' in the probability here to emphasize the dependence on $T(\boldsymbol{X})$.

Summing Equation (B.16) over $\mathbf{W}_p$, we have

$$\mathbb{P}\left(\boldsymbol{X}_{1:(p-1)} = \mathbf{W}_{1:(p-1)} \mid T(\boldsymbol{X})\right) = C_0 \prod_{j=2}^{p-1} \phi_j(\mathbf{W}_{j-1}, \mathbf{W}_j \mid T(\boldsymbol{X})) \sum_{\mathbf{W}_p} \phi_p(\mathbf{W}_{p-1}, \mathbf{W}_p \mid T(\boldsymbol{X})), \quad \text{(B.18)}$$

thus

$$\mathbb{P}\left(\boldsymbol{X}_p = \mathbf{W}_p \,\middle|\, \boldsymbol{X}_{1:(p-1)} = \mathbf{W}_{1:(p-1)}, T(\boldsymbol{X})\right) = \frac{\phi_p(\mathbf{W}_{p-1}, \mathbf{W}_p \mid T(\boldsymbol{X}))}{\sum_{\mathbf{W}'_p} \phi_p(\mathbf{W}_{p-1}, \mathbf{W}'_p \mid T(\boldsymbol{X}))}.$$

Since the right hand side of the last equation does not involve $\mathbf{W}_{1:(p-2)}$, we conclude

$$\mathbb{P}\left(\boldsymbol{X}_p \,\middle|\, \boldsymbol{X}_{1:(p-1)}, T(\boldsymbol{X})\right) = \mathbb{P}\left(\boldsymbol{X}_p \mid \boldsymbol{X}_{p-1}, T(\boldsymbol{X})\right).$$

In addition, let $\phi'_{p-1}(\mathbf{W}_{p-2}, \mathbf{W}_{p-1} \mid T(\boldsymbol{X})) = \phi_{p-1}(\mathbf{W}_{p-2}, \mathbf{W}_{p-1} \mid T(\boldsymbol{X})) \sum_{\mathbf{W}'_p} \phi_p(\mathbf{W}_{p-1}, \mathbf{W}'_p \mid T(\boldsymbol{X}))$ and for $j < p-1$, let $\phi'_j = \phi_j$, (B.18) can be rewritten as

$$\mathbb{P}\left(\boldsymbol{X}_{1:(p-1)} = \mathbf{W}_{1:(p-1)} \mid T(\boldsymbol{X})\right) = C_0 \prod_{j=2}^{p-1} \phi'_j(\mathbf{W}_{j-1}, \mathbf{W}_j \mid T(\boldsymbol{X})), \quad \text{(B.19)}$$

which has the same form as Equation (B.16). Continuing the same reasoning for $\boldsymbol{X}_{p-1}, \boldsymbol{X}_{p-2}, \ldots, \boldsymbol{X}_2$, we conclude

$$\mathbb{P}\left(\boldsymbol{X}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)}, T(\boldsymbol{X})\right) = \mathbb{P}\left(\boldsymbol{X}_j \mid \boldsymbol{X}_{j-1}, T(\boldsymbol{X})\right),\ 2 \le j \le p,$$

that is, the sequence of $\boldsymbol{X}_j$'s is a Markov chain conditional on $T(\boldsymbol{X})$.

SCIP

Proof of Proposition B.5. The first equation follows from the uniform distribution and the Markovian property

$$\mathbb{P}\left(\boldsymbol{X}_1 = \mathbf{W}_1 \mid \boldsymbol{X}_2 = \mathbf{W}_2\right) = \mathbb{P}\left(\boldsymbol{X}_1 = \mathbf{W}_1 \mid \boldsymbol{X}_2 = \mathbf{W}_2, \boldsymbol{X}_{3:p} = \mathbf{X}_{3:p}\right).$$

Next we prove the second equation, except for the case of $j = 2$. However, the second equation with $j = 2$ and also the third equation both follow the same proof, by allowing $k_1 : k_2$ for $k_1 > k_2$ to denote the empty set.

Before the proof, we show an implication of Bayes' rule. For any $\mathbf{W}_j$ and $\mathbf{W}_{j+1}$,

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
\propto\ &\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
&\times \mathbb{P}\left(\tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)} \,\middle|\, \boldsymbol{X}_j = \mathbf{W}_j, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \qquad \text{(B.20)} \\
\propto\ &\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
&\times \mathbb{P}\left(\tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)} \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}\right) \qquad \text{(B.21)} \\
\propto\ &\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \qquad \text{(B.22)} \\
\propto\ &\mathbf{1}_{\left\{(\mathbf{X}_{1:(j-1)}, \mathbf{W}_j, \mathbf{W}_{j+1}, \mathbf{X}_{(j+2):p}) \in \mathcal{Q}\right\}}, \qquad \text{(B.23)}
\end{aligned}$$

where Equation (B.20) is due to Bayes' rule, Equation (B.21) is due to the Markovian property and the fact that SCIP sampling of $\tilde{\boldsymbol{X}}_{1:(j-2)}$ only depends on $\boldsymbol{X}_{1:(j-1)}$, and Equation (B.22) is because the conditional probability of $\tilde{\boldsymbol{X}}_{1:(j-2)}$ does not depend on $\mathbf{W}_j$. Note that the normalizing constant in Equation (B.23) depends on $\mathbf{W}_{j+1}$ but not on $\mathbf{W}_j$.

Now the second equation of Proposition B.5 can be shown as follows

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{j-1} = \mathbf{X}_{j-1}, \tilde{\boldsymbol{X}}_{j-1} = \tilde{\mathbf{X}}_{j-1}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}\right) \\
=\ &\mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)}, \tilde{\boldsymbol{X}}_{j-1} = \tilde{\mathbf{X}}_{j-1}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
&\qquad \text{(Since } \boldsymbol{X}_j, \tilde{\boldsymbol{X}}_{1:(j-2)}, \boldsymbol{X}_{(j+2):p} \text{ are conditionally independent given } \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}) \\
\propto\ &\mathbb{P}\left(\tilde{\boldsymbol{X}}_{j-1} = \tilde{\mathbf{X}}_{j-1} \,\middle|\, \boldsymbol{X}_j = \mathbf{W}_j, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
&\times \mathbb{P}\left(\boldsymbol{X}_j = \mathbf{W}_j \,\middle|\, \boldsymbol{X}_{1:(j-1)} = \mathbf{X}_{1:(j-1)}, \tilde{\boldsymbol{X}}_{1:(j-2)} = \tilde{\mathbf{X}}_{1:(j-2)}, \boldsymbol{X}_{j+1} = \mathbf{W}_{j+1}, \boldsymbol{X}_{(j+2):p} = \mathbf{X}_{(j+2):p}\right) \\
&\qquad \text{(By Bayes' rule)} \\
\propto\ &f_{j-1}(\tilde{\mathbf{X}}_{j-1}, \tilde{\mathbf{X}}_{j-2}, \mathbf{W}_j) \mathbf{1}_{\left\{(\mathbf{X}_{1:(j-1)}, \mathbf{W}_j, \mathbf{W}_{j+1}, \mathbf{X}_{(j+2):p}) \in \mathcal{Q}\right\}}, \\
&\qquad \text{(By the Markovian property and Equation (B.22))}
\end{aligned}$$

where the normalizing constant does not depend on $\mathbf{W}_j$. Hence we have

$$f_j(\mathbf{W}_j, \tilde{\mathbf{X}}_{j-1}, \mathbf{W}_{j+1}) = \frac{f_{j-1}(\tilde{\mathbf{X}}_{j-1}, \tilde{\mathbf{X}}_{j-2}, \mathbf{W}_j) \mathbf{1}_{\left\{(\mathbf{X}_{1:(j-1)}, \mathbf{W}_j, \mathbf{W}_{j+1}, \mathbf{X}_{(j+2):p}) \in \mathcal{Q}\right\}}}{\sum_{\mathbf{W}'_j \in [K_j]^n} f_{j-1}(\tilde{\mathbf{X}}_{j-1}, \tilde{\mathbf{X}}_{j-2}, \mathbf{W}'_j) \mathbf{1}_{\left\{(\mathbf{X}_{1:(j-1)}, \mathbf{W}'_j, \mathbf{W}_{j+1}, \mathbf{X}_{(j+2):p}) \in \mathcal{Q}\right\}}}$$

□

Refined Blocking

Proof of Proposition B.6. In the following, we view $T(\boldsymbol{X})$ and $\boldsymbol{X}_B$ as fixed and only $\boldsymbol{X}_j$ being random, and denote this conditional probability by $\mathbb{P}_j(\cdot)$.

We first show $(\boldsymbol{H}, \tilde{\boldsymbol{H}}^{t_{\max}}) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{H}}^{t_{\max}}, \boldsymbol{H})$. Denote the probability mass function in (B.17) as $g(\mathbf{H})$. Since $\tilde{\boldsymbol{H}}^0 = \boldsymbol{H} \sim g$ and the transition kernel constructed in Algorithm 15 is in detailed balance

with density $g(\mathbf{H})$, $(\tilde{\boldsymbol{H}}^t)_{t=0}^{t_{\max}}$ is reversible. Thus

$$(\tilde{\boldsymbol{H}}^0, \tilde{\boldsymbol{H}}^{t_{\max}}) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{H}}^{t_{\max}}, \tilde{\boldsymbol{H}}^0). \tag{B.24}$$

By the sampling of $\tilde{\boldsymbol{X}}_j$ in the algorithm, we have

$$\tilde{\boldsymbol{X}}_j \perp\!\!\!\perp \boldsymbol{X}_j \mid \tilde{\boldsymbol{H}}^{t_{\max}}, \boldsymbol{H}. \tag{B.25}$$

and

$$\mathbb{P}_j\left(\tilde{\boldsymbol{X}}_j = \mathbf{X}_j \,\middle|\, \tilde{\boldsymbol{H}}^{t_{\max}} = \mathbf{H}\right) = \mathbb{P}_j\left(\boldsymbol{X}_j = \mathbf{X}_j \mid \boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}\right). \tag{B.26}$$

Hence

$$\begin{aligned}
&\mathbb{P}_j\left(\boldsymbol{X}_j = \mathbf{W}_j, \boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}, \tilde{\boldsymbol{X}}_j = \tilde{\mathbf{W}}_j, \tilde{\boldsymbol{H}}^{t_{\max}} = \tilde{\mathbf{H}}\right) \\
&= \mathbb{P}_j\left(\boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}, \tilde{\boldsymbol{H}}^{t_{\max}} = \tilde{\mathbf{H}}\right)\mathbb{P}_j\left(\boldsymbol{X}_j = \mathbf{W}_j, \tilde{\boldsymbol{X}}_j = \tilde{\mathbf{W}}_j \,\middle|\, \boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}, \tilde{\boldsymbol{H}}^{t_{\max}} = \tilde{\mathbf{H}}\right) \\
&= \mathbb{P}_j\left(\boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}, \tilde{\boldsymbol{H}}^{t_{\max}} = \tilde{\mathbf{H}}\right)\mathbb{P}_j\left(\boldsymbol{X}_j = \mathbf{W}_j \mid \boldsymbol{H}(\boldsymbol{X}_j) = \mathbf{H}\right)\mathbb{P}_j\left(\tilde{\boldsymbol{X}}_j = \tilde{\mathbf{W}}_j \,\middle|\, \tilde{\boldsymbol{H}}^{t_{\max}} = \tilde{\mathbf{H}}\right) \\
&= \mathbb{P}_j\left(\tilde{\boldsymbol{H}}^{t_{\max}} = \mathbf{H}, \boldsymbol{H}(\boldsymbol{X}_j) = \tilde{\mathbf{H}}\right)\mathbb{P}_j\left(\tilde{\boldsymbol{X}}_j = \mathbf{W}_j \,\middle|\, \tilde{\boldsymbol{H}}^{t_{\max}} = \mathbf{H}\right)\mathbb{P}_j\left(\boldsymbol{X}_j = \tilde{\mathbf{W}}_j \,\middle|\, \boldsymbol{H}(\boldsymbol{X}_j) = \tilde{\mathbf{H}}\right) \\
&= \mathbb{P}_j\left(\boldsymbol{X}_j = \tilde{\mathbf{W}}_j, \boldsymbol{H}(\boldsymbol{X}_j) = \tilde{\mathbf{H}}, \tilde{\boldsymbol{X}}_j = \mathbf{W}_j, \tilde{\boldsymbol{H}}^{t_{\max}} = \mathbf{H}\right),
\end{aligned}$$

where the second equality is due to (B.25) and the third equality is due to Equations (B.24) and (B.26). Summing over all $\mathbf{H}, \tilde{\mathbf{H}}$, we conclude that $(\boldsymbol{X}_j, \tilde{\boldsymbol{X}}_j) \overset{\mathcal{D}}{=} (\tilde{\boldsymbol{X}}_j, \boldsymbol{X}_j) \mid \boldsymbol{X}_B$. □

C Conditional Hypothesis

This section concerns the hypotheses actually tested by conditional knockoffs. Suppose $(\boldsymbol{x}_i, Y_i) \overset{i.i.d.}{\sim} (X, Y)$ and $T(\boldsymbol{X})$ is a statistic of $\boldsymbol{X}$. The knockoff procedure using conditional knockoffs treats the variables in $\mathcal{H}_{0,T} = \{j \,:\, \boldsymbol{y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\}$ as null. It is of interest to compare $\mathcal{H}_{0,T}$ with $\mathcal{H}_0$, the original set of null variables defining the variable selection problem we actually care about.

Proposition C.1. $\mathcal{H}_0 \subseteq \mathcal{H}_{0,T}$.

Proof. Suppose $j \in \mathcal{H}_0$. For i.i.d. data, $j \in \mathcal{H}_0$ implies $Y_i \perp\!\!\!\perp X_{i,j} \mid X_{i,\text{-}j}$, which together with the independence among $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$ implies $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}$. Note that $\boldsymbol{y} \perp\!\!\!\perp T(\boldsymbol{X}) \mid \boldsymbol{X}$ (since $T(\boldsymbol{X})$ is deterministic given $\boldsymbol{X}$), which together with $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}$ implies $\boldsymbol{y} \perp\!\!\!\perp (\boldsymbol{X}_j, T(\boldsymbol{X})) \mid \boldsymbol{X}_{\text{-}j}$ by the contraction property of conditional independence. And by the weak union property of conditional independence, $\boldsymbol{y} \perp\!\!\!\perp (\boldsymbol{X}_j, T(\boldsymbol{X})) \mid \boldsymbol{X}_{\text{-}j}$ implies $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})$. Thus $j \in \mathcal{H}_{0,T}$. This holds for arbitrary $j \in \mathcal{H}_0$ and thus $\mathcal{H}_0 \subseteq \mathcal{H}_{0,T}$. □

The converse is not true in general, for instance if $T(\boldsymbol{X}) = \boldsymbol{X}$ and $\mathcal{H}_0 = \emptyset$, then all variables are automatically null conditional on $T(\boldsymbol{X})$ and thus $\mathcal{H}_0 \subsetneq \mathcal{H}_{0,T}$. In general, when $T(\boldsymbol{X})$ does not allow full reconstruction of $\boldsymbol{X}_j$ it should be rare for a non-null variable $\boldsymbol{X}_j$ to be null conditional on $T(\boldsymbol{X})$, as this can only happen if there is a perfect synergy of $F_{Y|X}$ and F_X so that $F_{Y|X}$ is only a function of X_j through a transformation computable from the sufficient statistic $T(\boldsymbol{X})$ of F_X. For most problems of interest, Theorem C.2 provides a sufficient condition for $\boldsymbol{y} \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})$, i.e., $\mathcal{H}_0 = \mathcal{H}_{0,T}$: the conditional mean of Y_i (or some transformation of Y_i) given $\boldsymbol{x}_i$, say $\phi(\boldsymbol{x}_i)$, should not be deterministic after conditioning on $\boldsymbol{X}_{\text{-}j}$ and $T(\boldsymbol{X})$.

Theorem C.2. *Suppose for a bounded function $g(y)$ and $\phi(\boldsymbol{x}) := \mathbb{E}[g(Y) \mid X = \boldsymbol{x}]$, there exist two disjoint Borel sets $B_1, B_2 \subset \mathbb{R}^p$ such that $\inf_{\boldsymbol{x}\in B_1} \phi(\boldsymbol{x}) > \sup_{\boldsymbol{x}\in B_2} \phi(\boldsymbol{x})$. If for each $j \in [p]$, it holds with positive probability that*

$$\mathbb{P}(\boldsymbol{x}_1 \in B_i \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})) > 0, i = 1, 2,$$

then $\mathcal{H}_0 = \mathcal{H}_{0,T}$.

This theorem is based on Proposition C.1 and the following Proposition C.3. By Proposition C.3, for each $j \notin \mathcal{H}_0$, it holds that $j \notin \mathcal{H}_{0,T}$; hence $\mathcal{H}_{0,T} \subseteq \mathcal{H}_0$. In addition, Proposition C.1 shows $\mathcal{H}_0 \subseteq \mathcal{H}_{0,T}$, thus $\mathcal{H}_0 = \mathcal{H}_{0,T}$.

Proposition C.3. *Suppose $Y \not\perp\!\!\!\perp X_j \mid X_{\text{-}j}$, and $g(y)$ is a bounded function. Define $K := \mathbb{E}[g(Y_1) \mid \boldsymbol{x}_1]$, and $M := \mathbb{E}[K \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})]$.*

(a) If K is different from M, then

$$\boldsymbol{y} \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X}).$$

(b) If K can be written as $\phi(\boldsymbol{x}_1)$ and ϕ is not constant on the support of the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{X}_{\text{-}j}$ and $T(\boldsymbol{X})$, i.e., there exist two disjoint Borel sets B_1, B_2 such that $\inf_{\boldsymbol{x}\in B_1} \phi(\boldsymbol{x}) > \sup_{\boldsymbol{x}\in B_2} \phi(\boldsymbol{x})$, and

$$0 < \mathbb{P}(\mathbb{P}(\boldsymbol{x}_1 \in B_i \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})) > 0, i = 1, 2),$$

then K is different from M.

To prove this proposition, we need the following lemma.

Lemma C.4. *Suppose $Y \not\perp\!\!\!\perp X$ and T is a function of X. Furthermore, if there exists a bounded function g such that $K := \mathbb{E}[g(Y) \mid X]$ is not conditionally deterministic in the following sense:*

$$0 < \mathbb{P}(K \neq \mathbb{E}[K \mid T]),$$

then $Y \not\perp\!\!\!\perp X \mid T$.

Proof. Let $M := \mathbb{E}[K \mid T]$. Then M is $\sigma(T)$-measurable. Since $\mathbb{P}(K \neq M) = \mathbb{P}(K > M) + \mathbb{P}(K < M)$, without loss of generality, we assume $0 < \mathbb{P}(K > M)$.

We compute $\mathbb{E}\left[g(Y)\mathbf{1}_{\{K>M\}} \,\middle|\, T\right]$ in two different ways. On the one hand,

$$\begin{aligned}
&\mathbb{E}\left[g(Y)\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] \\
\overset{a.s.}{=}\ &\mathbb{E}\left[\mathbb{E}\left[g(Y)\mathbf{1}_{\{K>M\}} \,\middle|\, X\right] \,\middle|\, T\right] && (\because \sigma(T) \subseteq \sigma(X)) \\
\overset{a.s.}{=}\ &\mathbb{E}\left[\mathbb{E}[g(Y) \mid X]\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] && (\because \{K > M\} \in \sigma(X)) \\
=\ &\mathbb{E}\left[K\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] && (\because \text{definition of } K)
\end{aligned}$$

On the other hand, if $Y \perp\!\!\!\perp X \mid T$ then

$$\begin{aligned}
&\mathbb{E}\left[g(Y)\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] \\
&\overset{a.s.}{=} \mathbb{E}\left[g(Y) \mid T\right]\mathbb{E}\left[\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] && (\because Y \perp\!\!\!\perp X \mid T) \\
&\overset{a.s.}{=} \mathbb{E}\left[\mathbb{E}\left[g(Y) \mid X\right] \mid T\right]\mathbb{E}\left[\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] && (\because \text{law of total expectation}) \\
&\overset{a.s.}{=} M\mathbb{E}\left[\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] && (\because \text{definition of } M) \\
&\overset{a.s.}{=} \mathbb{E}\left[M\mathbf{1}_{\{K>M\}} \,\middle|\, T\right]. && (\because M \in \sigma(T))
\end{aligned}$$

Combining these two expressions shows that if $Y \perp\!\!\!\perp X \mid T$ then $\mathbb{E}\left[(K-M)\mathbf{1}_{\{K>M\}} \,\middle|\, T\right] \overset{a.s.}{=} 0$, and thus $\mathbb{E}\left[(K-M)\mathbf{1}_{\{K>M\}}\right]=0$. However, this implies $\mathbb{P}(K>M)=0$ and contradicts the condition; hence $Y \not\perp\!\!\!\perp X \mid T$. □

Proof of Proposition C.3.

(a) The condition that $Y \not\perp\!\!\!\perp X_j \mid X_{\text{-}j}$ implies $Y_1 \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}$ (see Lemma C.5 below). Because $\boldsymbol{x}_1 \perp\!\!\!\perp \boldsymbol{x}_{2:n}$, we have $K = \mathbb{E}\left[g(Y_1) \mid \boldsymbol{x}_1, \boldsymbol{x}_{2:n}\right] = \mathbb{E}\left[g(Y_1) \mid \boldsymbol{X}_j, \boldsymbol{X}_{\text{-}j}\right]$. The condition $\mathbb{P}(K \neq M) > 0$ implies that $\mathbb{P}\left(K \neq M \mid \boldsymbol{X}_{\text{-}j}\right) > 0$ holds with positive probability.

To apply Lemma C.4, $\boldsymbol{X}_{\text{-}j}$ is treated as fixed, and $\boldsymbol{X}_j$ (resp. Y_1) is treated as X (resp. Y). Then we have $Y_1 \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})$, which immediately implies $\boldsymbol{y} \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})$.

(b) The existence of B_1 and B_2 implies that there exists a real number s such that

$$\sup_{x \in B_2} \phi(x) < s < \inf_{x \in B_1} \phi(x).$$

We will prove by contradiction that K is different from M. Suppose $\mathbb{P}(K \neq M) = 0$, then $\mathbb{P}\left(K \neq M \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \overset{a.s.}{=} 0$. Thus a.s. we have

$$\begin{aligned}
&\mathbb{P}\left(X_1 \in B_1 \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \\
&= \mathbb{P}\left(X_1 \in B_1, K = M \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \\
&\leq \mathbb{P}\left(X_1 \in B_1, M > s \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) && (\because s < \inf_{x \in B_1} \phi(x)) \\
&\leq \mathbf{1}_{\{M>s\}}. && (\because M \in \sigma(\boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})))
\end{aligned}$$

Similarly $\mathbb{P}\left(X_1 \in B_2 \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \overset{a.s.}{\leq} \mathbf{1}_{\{M<s\}}$. Since $\mathbf{1}_{\{M>s\}} \cdot \mathbf{1}_{\{M<s\}} = 0$, it follows that

$$\mathbb{P}\left(X_1 \in B_1 \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \cdot \mathbb{P}\left(X_1 \in B_2 \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) \overset{a.s.}{\leq} 0,$$

which contradicts the condition that $0 < \mathbb{P}\left(\mathbb{P}\left(X_1 \in B_i \mid \boldsymbol{X}_{\text{-}j}, T(\boldsymbol{X})\right) > 0, i = 1, 2\right)$. Hence we conclude $\mathbb{P}(K \neq M) > 0$.

□

Lemma C.5. *If $Y \not\perp\!\!\!\perp X \mid U$ and $(X, Y, U) \perp\!\!\!\perp (V, W)$, then $Y \not\perp\!\!\!\perp (X, V) \mid (U, W)$.*

Proof. Suppose $Y \perp\!\!\!\perp (X, V) \mid (U, W)$. Then

$$Y \perp\!\!\!\perp X \mid (U, W), \tag{C.1}$$

The condition that $(X, Y, U) \perp\!\!\!\perp (V, W)$ implies $(X, Y, U) \perp\!\!\!\perp W$, and thus by weak union property of conditional independence, we have

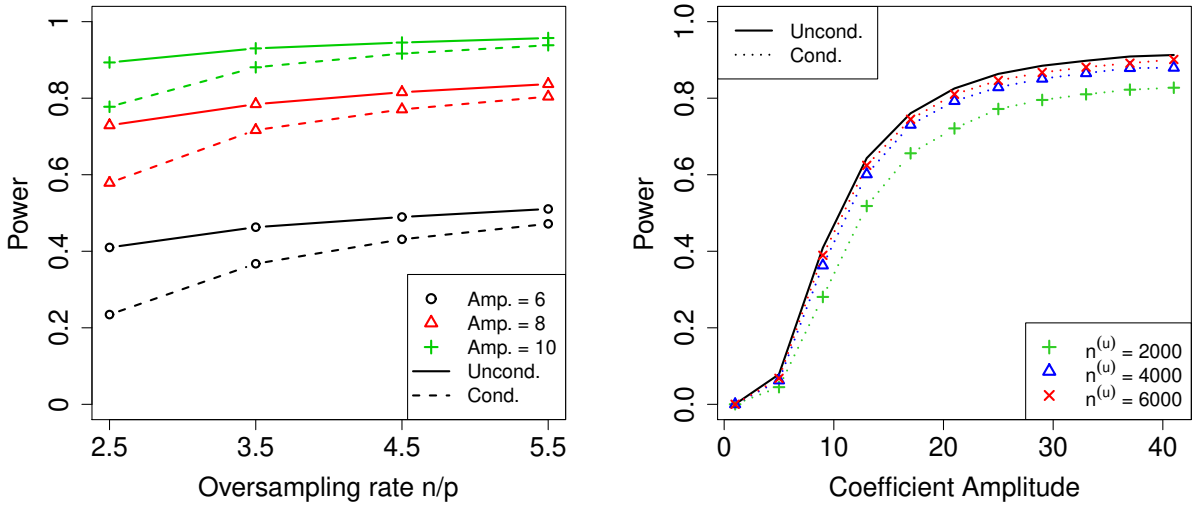
$$(X, Y) \perp\!\!\!\perp W \mid U. \tag{C.2}$$

Equations (C.1) and (C.2) together with the contraction property of conditional independence imply $Y \perp\!\!\!\perp X \mid U$. This contradicts the condition, so we conclude that $Y \not\perp\!\!\!\perp (X, V) \mid (U, W)$. □

D Supplementary Simulations

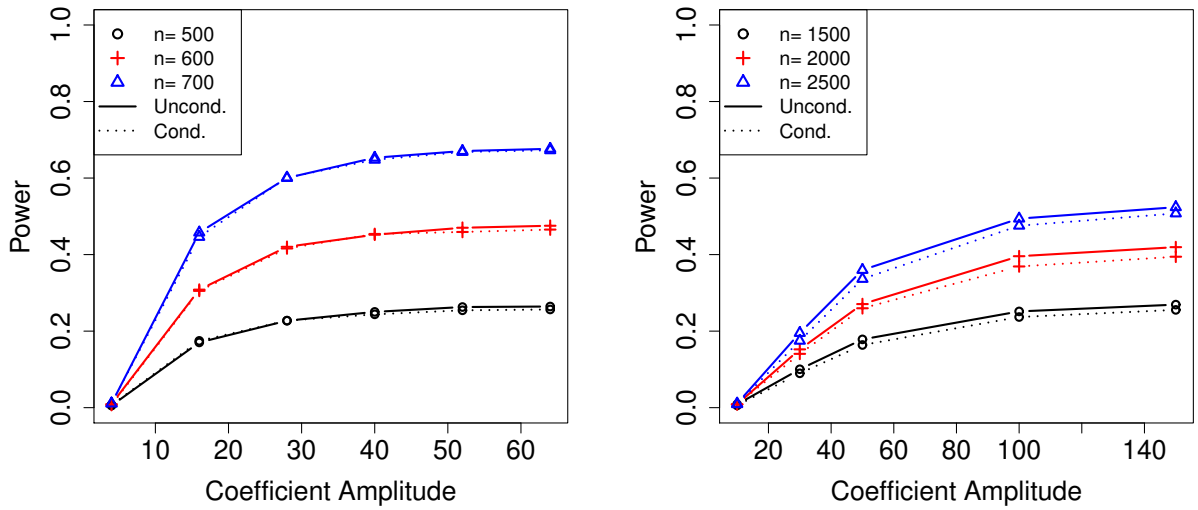
D.1 Nonlinear Response Models

We re-conduct the same the simulations in Section 3 on logistic regression, confirming that the variable selection by using conditional knockoff allows for general response dependence. The experiments follow the same designs as in Sections 3.1.2, 3.2.2 and 3.3.3, but with binary responses sampled as $Y_i \mid \boldsymbol{x}_i \sim \text{Bernoulli}(\varsigma(\boldsymbol{x}_i^\top \boldsymbol{\beta}/\sqrt{n}))$, where $\varsigma(t) = e^t/(1+e^t)$ is the logistic function, and slightly larger sample sizes n for the re-conducted simulations of Sections 3.2.2 and 3.3.3.



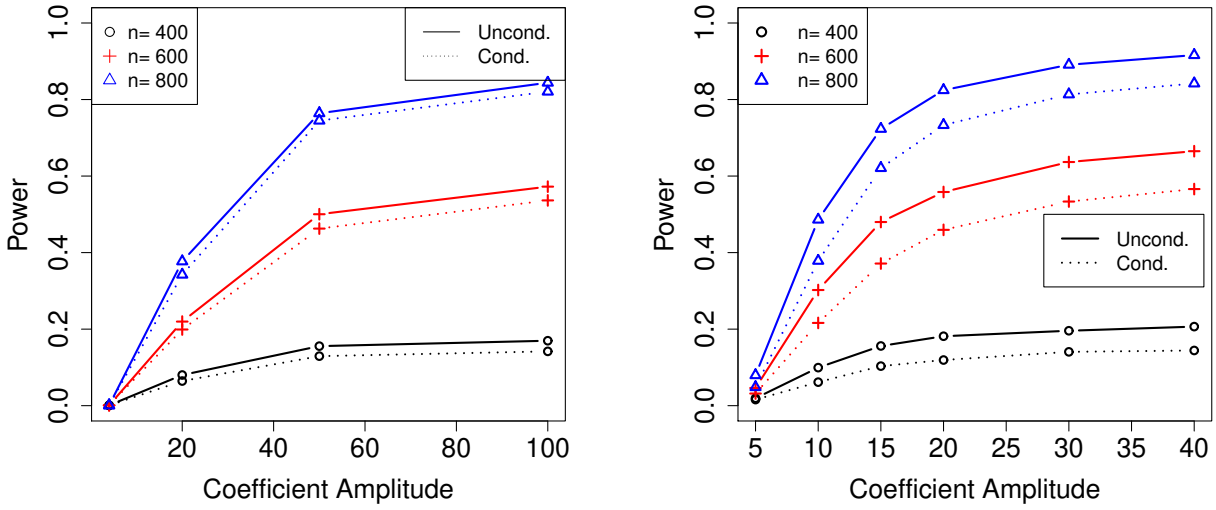
(a) (b)

Figure 6: (Logistic regression version of Figure 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 $n = 800$ fixed. Standard errors are all below 0.006.



(a) (b)

Figure 7: (Logistic regression version of Figure 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) (b)

Figure 8: (Logistic regression version of Figure 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 0.006.

.

D.2 Varying the Sparsity and Magnitude of the Regression Coefficients

The following simulations reproduce Figure 2a but with varying sparsities and magnitudes. Specifically, the sparsity level k varies between 30, 60, and 90, and the nonzero entries are randomly sampled from Unif$(1, 2)$. The message from these experiments is the same as those in the main pa-

per, that is, the power of conditional knockoffs is almost the same as that of unconditional knockoffs even though it does not know the exact distribution of $\boldsymbol{X}$.

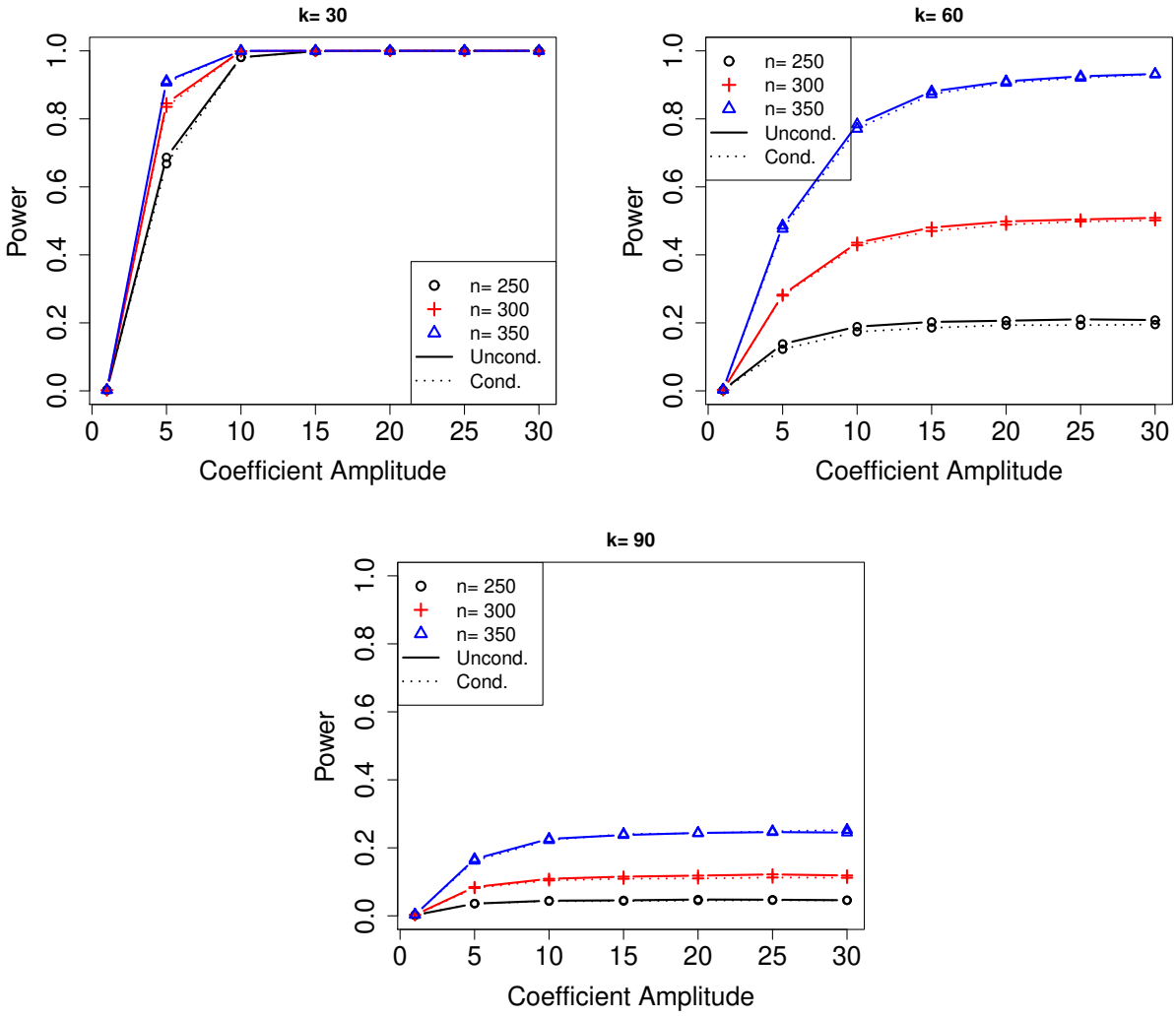


Figure 9: Reproducing Figure 2(a) of the paper with different $\boldsymbol{\beta}$. Power curves of conditional and unconditional knockoffs for $p = 2000$ and a range of n for an AR(1) model.

D.3 Power of Different Sufficient Statistics

We provide the following experiment to examine the power performance of conditional knockoffs that are generated using different sufficient statistics. It confirms our intuition that conditioning on more generally leads to lower power.

Specifically, for a Gaussian graphical model, we have run Algorithm 4 for a sequence of nested sufficient statistics to see how this choice affects the power. In the following simulation, X is sampled from an AR(1) distribution with autocorrelation coefficient 0.3, and models of (nonstationary) AR(q) with various $q \geq 1$ are used to model X, i.e., each model assume a banded precision matrix with bandwidth q, and we increase q beyond 1 to study the effect of more conditioning. Since the models are nested, all of them lead to valid conditional knockoffs. As q grows, the graphical model

gets denser and the sufficient statistic conditioned on in Algorithm 4 contains more elements (and always contains all the elements of the sufficient statistic conditioned on for all smaller q), which can be done by choosing two increasing sequences of blocking sets for the two split data folds and making sure that these two sequences never intersect with each other. Thus, we expect to see some loss of power when q increases. We chose $n = 400$, $p = 2000$, and the algorithmic parameter n' to be set to 160, and produced results for a range of $Y \mid X$'s linear model coefficient amplitudes and for q ranging from 1–30; see Figure 10. Although a larger value of q indeed lowers the power, the loss is relatively small in this example despite conditional knockoffs with $q = 30$ conditioning on far more than with $q = 1$.

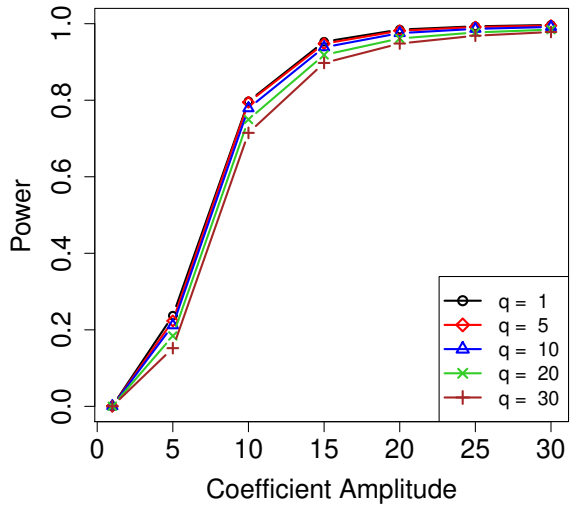


Figure 10: Power curves of conditional knockoffs for AR(q) models with $p = 2000$ as coefficient amplitude varies for various q. The nominal FDR level is 0.2. Standard errors are all below 0.007.

D.4 Gaussian Graphical Models with Unknown Edge Supersets

The conditional knockoff generation in Section 3.2 requires knowing a superset of the true edge set of the Gaussian graphical model. We present the following experiment to preliminarily examine the idea mentioned in Remark 3 for when such a superset is not known a priori and one has to estimate the edge set (or its superset) using the data.

Suppose the true covariance matrix $\mathbf{\Sigma}$ is a rescaled (to have diagonal entries equal to 1) version of $\mathbf{\Sigma}^{(0)}$, where $(\mathbf{\Sigma}^{(0)})^{-1}_{j,k} = \mathbf{1}_{j=k} - \frac{1}{7}\mathbf{1}_{1\le|j-k|\le3}$. In other words, every node in the true graph is connected to its 6 nearest neighbors. In the following simulations, we set $p = 400$ and $n = 200$.

We can estimate the edge set E by the nonzero entries $\hat{E}$ of the estimated precision matrix using the *graphical Lasso* (Friedman et al., 2008), which is implemented via the `R` package `huge`. The tuning parameter of the graphical Lasso is selected by the standard method *StARS* (Liu et al., 2010). Once $\hat{E}$ is computed, we can then construct conditional knockoffs as if $\hat{E}$ were given. The blocking set used in our algorithms is obtained by Algorithm 2 with input $n' = 80$. We additionally consider the case where a set of $n_u = 1,600$ unlabeled data points is available and is used together with the labeled covariates to estimate the edge set.

The FDR and power curves are shown in Figure 11. "Label Cond." refers to conditional knockoffs generated with $\hat{E}$ estimated using only labeled data, and "Unlabel. Cond." refers to conditional knockoffs with $\hat{E}$ estimated additionally with the unlabeled data. Both methods control the FDR and in fact are conservative. One might attribute the FDR control to some over-conservative choice of graphical Lasso tuning parameter, but in fact $\hat{E}$ estimated with just the labeled data, although it tended to find a larger graph than the truth (its maximal degree was often above 20), also missed around 40 true edges on average. Unsurprisingly, the use of unlabeled data improves the power by improving the estimate of the edge set, with much fewer false negative and false positive edges.

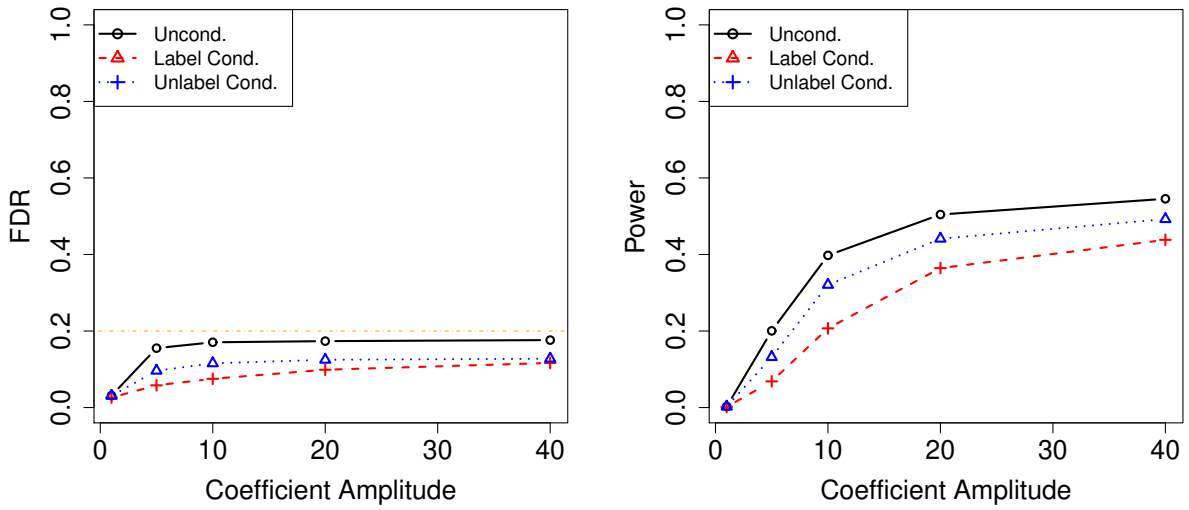


Figure 11: FDR and Power curves of unconditional knockoffs and (approximate) conditional knockoffs using an estimated edge set either from purely labeled or from both labeled and unlabeled data. $p = 400$ and $n = 200$. Standard errors are all below 0.007.

D.5 Robustness to Model Misspecification

The current paper focuses on the cases where the models for the covariates are known and well-specified. In practice, practitioners may not know what the true model is. Here we provide an experiment to examine the robustness of Gaussian conditional knockoffs ($n > 2p$). The following robustness experiment constructs a set of distributions that approximate a multivariate Gaussian by discretizing it at different resolutions by varying a parameter K.

We first generate $X^{(0)} \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{i,j} = 0.3^{|i-j|}$, and then discretize each coordinate as follows

$$X_j = \frac{\lfloor X_j^{(0)} \times K + 1/2 \rfloor}{K}, \quad j = 1, \ldots, p.$$

In other words, $X_j^{(0)}$ is rounded to the nearest $\frac{1}{K}$-grid value. Since $|X_j - X_j^{(0)}| \leq \frac{1}{K}$, the larger K, the closer X is to a multivariate Gaussian vector, and indeed as $K \to \infty$, $X \to X^{(0)}$ and becomes multivariate Gaussian. However, for small K, the distribution is very far from Gaussian. For conditional knockoffs, we pretend that X is drawn from a multivariate Gaussian distribution and directly apply Algorithm 3.1. To get a baseline for power (since changing K not only affects

the model misspecification, but also changes the nature of the data-generating distribution and thus the power of any procedure), we also generate exactly-valid unconditional knockoffs for X by discretizing an unconditional knockoff of $X^{(0)}$ with the same K (of course this procedure would be impossible in practice, since $X^{(0)}$ is unobserved).

We fix $p = 1,000$, linear model coefficient amplitude at 4, vary $n \in \{2001, 3000, 4000\}$ and vary $K \in \{1/2, 1, 2, 3, \infty\}$. The other details of the experiment are the same as in Figure 1a of the paper, where the response is drawn from $Y_i \mid X_i \sim N(X_i^\top \boldsymbol{\beta}/\sqrt{n}, 1)$. The result is shown in Figure 12.

Note that $K = 1/2$ produces a distribution that is almost entirely concentrated on just three values $\{-2, 0, 2\}$, making it extremely non-Gaussian, yet the FDR is controlled quite well for all values of n at this K value and all others. The power difference between conditional knockoffs and unconditional knockoffs is also quite insensitive to K and, as seen in all other simulations, quite small for all n except when $n \approx 2p$ (in the $n \approx 2p$ setting the power gap is substantial, although conditional knockoffs still has quite a bit of power).

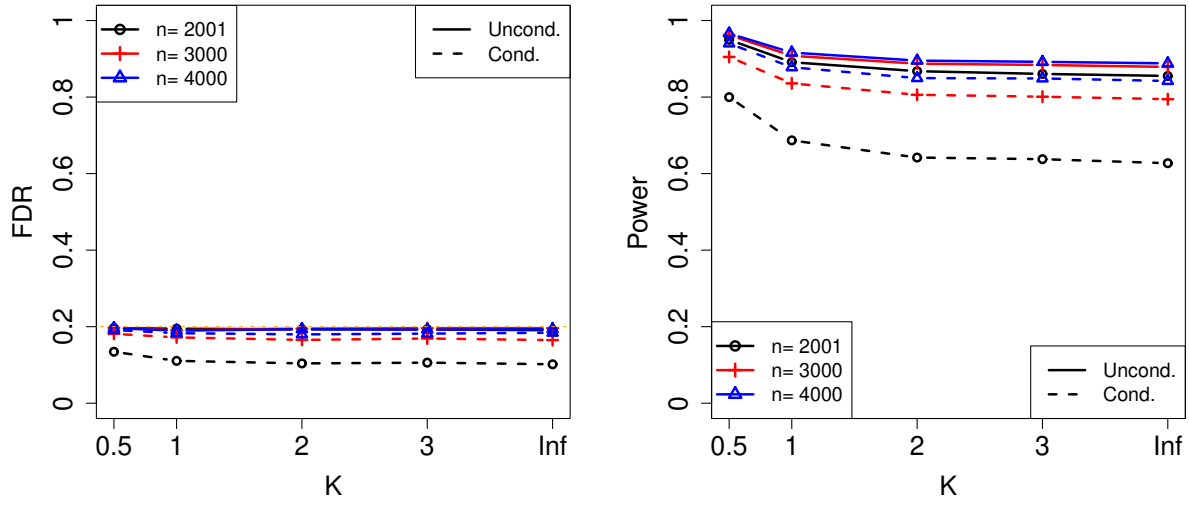


Figure 12: FDR and Power curves of unconditional and conditional knockoffs for a discretized AR(1) model with $p = 1000$ and linear model coefficient amplitude 4. The nominal FDR level is 0.2. Standard errors are all below 0.004.