

Powerful Knockoffs via Minimizing Reconstructability

Asher Spector and Lucas Janson

Department of Statistics, Harvard University

Abstract

Model-X knockoffs (Candès et al., 2018) allows analysts to perform feature selection using almost any machine learning algorithm while still provably controlling the expected proportion of false discoveries. To apply model-X knockoffs, one must construct synthetic variables, called knockoffs, which effectively act as controls during feature selection. The gold standard for constructing knockoffs has been to minimize the mean absolute correlation (MAC) between features and their knockoffs, but, surprisingly, we prove this procedure can be powerless in extremely easy settings, including Gaussian linear models with correlated exchangeable features. The key problem is that minimizing the MAC creates strong joint dependencies between the features and knockoffs, which allow machine learning algorithms to partially or fully *reconstruct* the effect of the features on the response using the knockoffs. To improve the power of knockoffs, we propose generating knockoffs which *minimize the reconstructability* (MRC) of the features, and we demonstrate our proposal for Gaussian features by showing it is computationally efficient, robust, and powerful. We also prove that certain MRC knockoffs minimize a natural definition of estimation error in Gaussian linear models. Furthermore, in an extensive set of simulations, we find many settings with correlated features in which MRC knockoffs dramatically outperform MAC-minimizing knockoffs and no settings in which MAC-minimizing knockoffs outperform MRC knockoffs by more than a very slight margin. We implement our methods and a host of others from the knockoffs literature in a new open source python package `knockpy`.¹

1 Introduction

Model-X (MX) knockoffs (Candès et al., 2018) has recently emerged as a powerful and flexible method to perform *controlled variable selection*. Informally, given a set of features (X_1, \dots, X_p) and an outcome of interest Y , knockoffs allows one to leverage almost any regression method to discover relationships between the features and the outcome. Notably, knockoffs exactly control the expected proportion of false positives in finite samples provided that the distribution of the features X is known, while assuming nothing about the conditional distribution $Y | X$.

The knockoffs framework accomplishes this task by constructing synthetic variables, called knockoffs, which mimic the correlation structure of the original features. In principle, there are many possible ways to construct valid knockoff variables. However, the knockoffs literature has largely converged to a single measure of knockoff quality, namely that one should minimize the mean absolute correlation (MAC) between features and their knockoffs in order to maximize statistical power. Such knockoffs can be constructed via semidefinite programming (SDP) when the features are multivariate Gaussian, and almost the entire knockoffs literature has treated them as the “gold standard” of knockoff quality (see, e.g., Barber and Candès (2015); Candès et al. (2018); Li and Maathuis (2019); Bates et al. (2020); Askari et al. (2020)).

1.1 Contribution

This paper describes an improved heuristic for generating powerful knockoffs. With this goal in mind, our work makes two key contributions.

¹See <https://github.com/amspector100/knockpy>.

MVR, ME, and SDP Knockoffs on Various Gaussian Designs

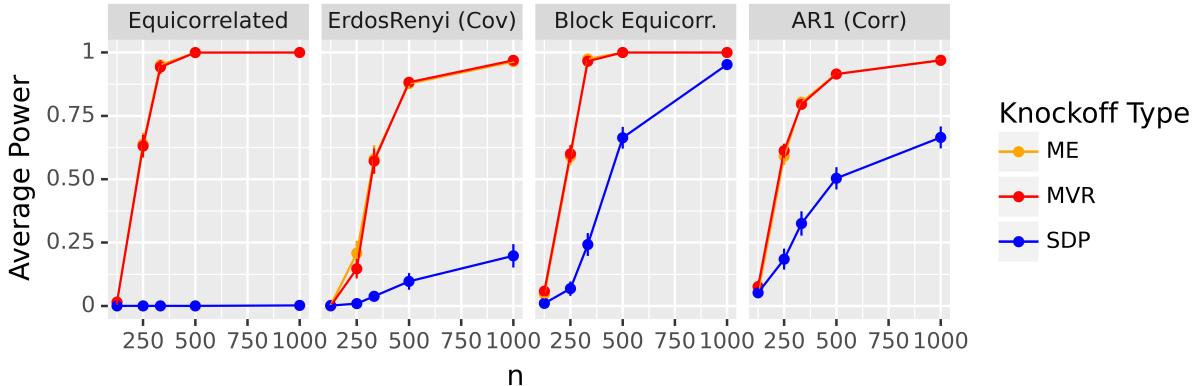


Figure 1: This figure illustrates the main contributions of our work, namely that SDP knockoffs can have extraordinarily low power and the proposed MRC framework (MVR and ME knockoffs) resolves this issue. In all cases, we sample $X \sim \mathcal{N}(0, \Sigma)$ with the different horizontal facets corresponding to different Σ , as defined in Section 4.2. We let $Y | X \sim \mathcal{N}(X\beta, 1)$ and use lasso coefficient differences as feature statistics with $p = 500$ and 50 non-nulls. The non-nulls are sampled independently from $\text{Unif}([-\delta, -\delta/2] \cup [\delta/2, \delta])$ where $\delta = 2$ for the AR1 panel and 1 for the others. We control the FDR at level $q = 0.1$.

Identifying the reconstruction effect. We prove that existing knockoff generators often create strong joint dependencies between the features and knockoffs. This allows predictive algorithms like the lasso to *reconstruct* the effect of non-null features on the response using the knockoffs, which can substantially reduce statistical power.

- We prove that in a very simple setting—an exchangeable Gaussian design with correlation $\rho \geq 0.5$ and a Gaussian linear model for the response—almost every feature statistic has asymptotically zero power when used with knockoffs that minimize the MAC.
- We argue via both theory and simulations that minimizing the MAC frequently causes the reconstruction effect for correlated designs, and that this phenomenon will reduce power. We also identify several examples of the reconstruction effect in the previous knockoffs literature.

More powerful knockoffs via minimizing reconstructability (MRC). We introduce a novel framework which generates powerful knockoffs by minimizing one's ability to reconstruct a feature using the other features and the knockoffs. We consider two concrete instantiations of this framework based on two measures of reconstructability: minimum variance-based reconstructability (MVR) knockoffs and maximum entropy (ME) knockoffs. These methods are well-defined for all design distributions, although they are particularly easy to analyze in the case when the features are Gaussian.

- We prove that when the features and response jointly follow a multivariate Gaussian distribution, MVR knockoffs exactly minimize the estimation error when using ordinary least squares (OLS) coefficients as feature importances. This means that even if MAC-minimizing knockoffs are perturbed to prevent the features from being exactly reconstructable, we still expect MVR knockoffs to have higher power.
- We demonstrate via simulations that MVR and ME knockoffs often have dramatically higher power than MAC-minimizing knockoffs for correlated Gaussian features. As the features become less dependent, the performances of all three methods equalize, but crucially, we have not observed any examples where SDP knockoffs dramatically outperform either MVR or ME knockoffs. We provide simulation results in both high and low dimensions for a wide variety of design distributions, response distributions, and feature statistics. Notably, this same conclusion holds even in our simulations with highly non-linear responses and feature statistics.

- We also apply the MRC framework in a variety of simulation settings beyond the Gaussian model-X case. We demonstrate MRC knockoffs can increase the power of fixed-X knockoffs (Barber and Candès, 2015), second-order knockoffs (Candès et al., 2018), and the general Metropolized knockoff sampler for non-Gaussian features (Bates et al., 2020).

Figure 1 encapsulates both of these contributions, namely that when the features are correlated, minimizing the MAC (via SDP when the features are Gaussian) often produces much-lower-power knockoffs than the MVR and ME knockoffs considered in this paper.

1.2 Notation

Let $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ be the set of p features and let $Y \in \mathbb{R}$ be the response. We stack n i.i.d. observations of the features and response into the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the response vector $\mathbf{y} \in \mathbb{R}^n$. We use the non-bolded notation (X, Y) to refer to an arbitrary single observation. For $p \in \mathbb{N}$, let $[p]$ denote the set $\{1, \dots, p\}$. For any subset $J \subset [p]$, we will denote \mathbf{X}_J as the matrix of columns of \mathbf{X} whose indices belong to J . \mathbf{X}_{-J} denotes all of the columns of \mathbf{X} whose indices do *not* belong to J . \mathbf{X}_j denotes the j th column of \mathbf{X} , and \mathbf{X}_{-j} denotes all of the columns of \mathbf{X} except column j . We let I_p refer to the $p \times p$ identity. For a pair of square matrices $M_1, M_2 \in \mathbb{R}^{d \times d}$, we say that $M_1 \succcurlyeq M_2$ if and only if $M_1 - M_2$ is positive semi-definite. We will let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimum and maximum eigenvalues of a square matrix M , respectively, and $\lambda_j(M)$ denotes the j th smallest eigenvalue of M . For two vectors $v \in \mathbb{R}^k, v' \in \mathbb{R}^{k'}$ with $k, k' \geq 1$, we let $(v, v') \in \mathbb{R}^{k+k'}$ denote their concatenation. For two matrices $M_1, M_2 \in \mathbb{R}^{n \times p}$, we will define $[M_1, M_2] \in \mathbb{R}^{n \times 2p}$ to represent the column-wise concatenation of M_1 and M_2 . For $j \in [p]$, we will let $[M_1, M_2]_{\text{swap}(j)}$ denote the concatenation of M_1 and M_2 except that the j th column of M_1 has been swapped with the j th column of M_2 . For $J \subset [p]$, $[M_1, M_2]_{\text{swap}(J)}$ refers to the concatenation of M_1 and M_2 except with all columns $j \in J$ swapped. For any permutation $\sigma : [p] \rightarrow [p]$ and any matrix $M \in \mathbb{R}^{k \times p}$, let $\sigma(M)$ denote the same matrix as M but with the columns permuted according to the permutation σ . E.g., if $\sigma(1) = 3$, then the first column of M becomes the third column of $\sigma(M)$. For $J \subset [p]$, we let $-\mathbf{1}_J$ refer to the vector $v \in \mathbb{R}^p$ such that $v_j = -1$ for $j \in J$ and $v_j = 1$ when $j \notin J$. For a scalar k , we let $\mathbf{1}_k$ represent the vector of all ones in \mathbb{R}^k . For a vector $v \in \mathbb{R}^p$, $\text{diag}(v)$ represents the $p \times p$ diagonal matrix with diagonal v . If M_1, \dots, M_k are square matrices, $\text{blockdiag}(M_1, \dots, M_k)$ denotes the square block-diagonal matrix with blocks M_1 through M_k .

1.3 Review of model-X knockoffs

MX knockoffs aims to simultaneously test the hypotheses $H_j : X_j \perp\!\!\!\perp Y | X_{-j}$, for $j \in [p]$. If $\mathcal{H}_0 = \{j : H_j\}$ is the set of null hypotheses, the knockoffs procedure selects a set of features \hat{S} and provably controls the false discovery rate (FDR) when the distribution of X is known:

$$\text{FDR} \equiv \mathbb{E} \left(\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|} \right) \leq q$$

for some prespecified $q \in [0, 1]$. Note that when $\hat{S} = \emptyset$, we use the convention that $0/0 = 0$. The knockoffs procedure consists of three steps: constructing knockoffs, computing feature importances and feature statistics, and applying a data-dependent threshold.

Step 1: Constructing knockoffs. First, given a feature vector $X \in \mathbb{R}^p$, we define knockoffs $\tilde{X} \in \mathbb{R}^p$ as random variables satisfying

$$[X, \tilde{X}]_{\text{swap}(J)} \stackrel{d}{=} [X, \tilde{X}] \text{ and } \tilde{X} \perp\!\!\!\perp Y | X \quad (1)$$

for all $J \subset [p]$. This pairwise exchangeability condition implies that

$$\text{Cov}([X, \tilde{X}]) = G_S \equiv \begin{bmatrix} \Sigma & \Sigma - S \\ \Sigma - S & \Sigma \end{bmatrix} \quad (2)$$

for some diagonal matrix S such that $G_S \succcurlyeq 0$, where $\Sigma = \text{Cov}(X)$. Note in the case where $X \sim \mathcal{N}(0, \Sigma)$, Candès et al. (2018) proved that when $[X, \tilde{X}] \sim \mathcal{N}(0, G_S)$, \tilde{X} are valid knockoffs for X for any S , as long as $G_S \succcurlyeq 0$. If one chooses to generate \tilde{X} such that $[X, \tilde{X}]$ are multivariate Gaussian, then the choice of S uniquely determines any valid knockoff-generation mechanism.

Step 2: Feature importances and feature statistics. The next step is to compute feature importances $Z \in \mathbb{R}^{2p}$ where for $j \in [p]$, Z_j and Z_{j+p} measure the importances of X_j and its knockoff \tilde{X}_j , respectively. We can use any function of the data $Z = z([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \in \mathbb{R}^{2p}$ to generate Z under the restriction that swapping a feature with its knockoff also swaps the feature importances Z_j, Z_{j+p} . This allows one to use almost any feature importance measure to create Z , from cross-validated lasso coefficients to neural networks (Lu et al., 2018).

We combine Z_j, Z_{j+p} into a *feature statistic* $W_j = f(Z_j, Z_{j+p})$, where f is an antisymmetric function such that $f(x, y) = -f(y, x)$. For example, if Z are absolute lasso coefficients, we might set $f(Z_j, Z_{j+p}) = Z_j - Z_{j+p}$. Here, W_j represents the lasso (absolute) coefficient difference (LCD), where a high value of W_j indicates that feature X_j is more important than its knockoff in predicting Y . The antisymmetric property of f implies that swapping X_j with its knockoff flips the sign of W_j . We let $W = w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \in \mathbb{R}^p$ be the vector of feature statistics. Note Z, W are random variables, and z, w are functions.

Step 3: The data-dependent threshold. Finally, we define the data-dependent threshold

$$T = \min \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\} + 1}{\#\{j : W_j \geq t\}} \leq q \right\}. \quad (3)$$

Formally, to ensure this minimum is well-defined, the minimum is only over $t \in \{|W_j| : j \in [p]\} \setminus \{0\}$. By convention, $T = \infty$ if that set is empty. Selecting features $\hat{S} = \{j : W_j \geq T\}$ guarantees FDR control at level q .

1.4 Related literature

The MX knockoffs framework (Candès et al., 2018) has received significant attention recently because it guarantees exact FDR control in feature selection even when the response $Y | X$ is nonlinear and the features X are arbitrarily correlated. In particular, knockoffs has been applied successfully in genome-wide association studies (GWAS), with promising empirical performance (Sesia et al. (2018, 2019, 2020)). Although the model-X framework does assume that the distribution of X is known, prior work has shown through both theory (Barber et al., 2020) and simulations (for example, Sesia et al. (2018)) that knockoffs is fairly robust to misspecification of the distribution of X . Additionally, a new line of research has relaxed the assumptions of knockoffs, such that knockoffs can control the false discovery rate as long as the distribution of X is known up to a parametric model (Huang and Janson, 2020). Since a main advantage of knockoffs is that they control the false discovery rate under feature dependence, the main goal of our paper is to improve the *power* of knockoffs under feature dependence. We pause to briefly compare our contribution in this respect to two relevant strands of the wider knockoffs literature.

First, our work draws heavily from the literature on sampling model-X knockoffs. Initially, Barber and Candès (2015) introduced the fixed-X knockoff framework and suggested creating knockoffs which minimize the MAC. Candès et al. (2018) built on Barber and Candès (2015) to introduce the model-X knockoffs filter, demonstrate how to sample knockoffs for Gaussian designs, and prove the existence of nontrivial knockoffs for general designs. Notably, Candès et al. (2018) also suggested generating knockoffs which minimize the MAC metric, although in their discussion they note the possibility of increased power from alternate knockoff constructions. Since then, several works have developed techniques to sample knockoffs for non-Gaussian designs. Sesia et al. (2018) developed a method to sample knockoffs for discrete Markov chains, and Gimenez et al. (2019) demonstrated how to sample knockoffs for some Bayesian networks. When the true model is unknown, Romano et al. (2018); Jordon et al. (2019) showed how to use deep generative networks and generative adversarial networks (GANs), respectively, to generate approximate knockoff constructions. Most recently, Bates et al. (2020) characterized all knockoff distributions and developed efficient algorithms to sample exact knockoffs in great generality, using tools from Markov Chain Monte Carlo (MCMC). However, very little of the existing literature has discussed *which* knockoffs to generate, and most works have assumed

implicitly that weaker feature-knockoff correlations improve power. Indeed, Bates et al. (2020) did not even simulate the power of their general knockoff sampler—instead, they reported the MAC as a proxy for knockoff quality. The point of our paper is to demonstrate that the MAC heuristic can fail spectacularly and to propose a solution to this problem.

Second, four recent papers (Chen et al. (2019); Gimenez and Zou (2019); Liu and Rigollet (2019); Ke et al. (2020)) have observed that knockoffs can sometimes lose power in correlated settings, and Fan et al. (2020) prove the consistency of lasso-based knockoffs under certain conditions on the feature-knockoff distribution. Our work differs substantially from these, but we will be better able to explain why after we have formally introduced the reconstruction effect. See Section 2.5 for detailed comparisons to each of these works.

1.5 Outline

The outline of the rest of the paper is as follows. In Section 2, we describe how minimizing the MAC causes the reconstruction effect for general Gaussian designs. We also demonstrate that reconstructability reduces the power of knockoffs for all designs. In Section 3, we define two types of MRC knockoffs: MVR knockoffs and ME knockoffs. We prove that MVR knockoffs are optimal in a sense for OLS coefficients in Gaussian linear models, and we further prove that MVR knockoffs are consistent in low dimensions, unlike SDP knockoffs. We also give intuition as to why MRC knockoffs are likely to improve power for nonlinear responses and discuss efficient algorithms for computing MRC knockoffs when the features are Gaussian. Finally, in Section 4, we present an extensive set of simulations comparing the power of MRC and MAC-minimizing knockoffs.

2 The reconstruction effect reduces knockoffs’ power

The MAC heuristic applied in correlated settings can lead to significant power loss due to the reconstruction effect we identify in this section.

2.1 Minimizing the MAC results in reconstructability for Gaussian designs

In this section, we demonstrate that minimizing the MAC often ensures that many X_j can be reconstructed from X_{-j}, \tilde{X} when X is Gaussian and correlated. To begin with, note that when $[X, \tilde{X}] \sim \mathcal{N}(0, G_S)$, choosing a type of knockoffs to generate is equivalent to choosing the S -matrix from equation (2). When X is Gaussian, we will refer to MAC-minimizing knockoffs as SDP knockoffs, as in the literature. When $X \sim \mathcal{N}(0, \Sigma)$, we will also assume X is scaled such that $\Sigma_{jj} = 1$ for $j \in [p]$.

Definition 2.1 (SDP Knockoffs). *Set $S_{\text{SDP}} = \text{diag}(s)$ where $s \in \mathbb{R}^p$ is the solution to the semidefinite program*

$$\begin{aligned} & \text{minimize} \quad \sum_{j=1}^p |1 - s_j| \\ & \text{s.t. } 0 \preceq \text{diag}(s) \preceq 2\Sigma. \end{aligned}$$

A more computationally-efficient version of this procedure, called the equicorrelated method, minimizes the same objective under the constraint that s is a constant vector.

This SDP formulation will continue to increase the diagonal values of S until it hits the boundary condition $\lambda_{\min}(2\Sigma - S) = 0$ or attains the optimal $S = I_p$. Since the eigenvalues of G_S are those of S and $2\Sigma - S$, $G_{\text{SDP}} \equiv G_{S_{\text{SDP}}}$ will be low rank whenever $S_{\text{SDP}} \neq I_p$.

Lemma 2.1. *Suppose $X \sim \mathcal{N}(0, \Sigma)$ and $\lambda_{\min}(\Sigma) \leq 0.5$. Then $\text{rank}(G_{\text{SDP}}) < 2p$. Furthermore, if Σ is block-diagonal with b blocks, each with an eigenvalue below 0.5, then $\text{rank}(G_{\text{SDP}}) \leq 2p - b$.*

As a running example throughout this section, we will often analyze the simple case where Σ is “equicorrelated” to enable more explicit analysis of the power of SDP knockoffs. This means $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{jk} = \rho$ if and only if $j \neq k$ and 1 otherwise, which implies X is exchangeable.

Lemma 2.2. *In the equicorrelated case when $\rho \geq 0.5$, let \tilde{X} be generated according to the SDP procedure. Then G_{SDP} has rank $p + 1$, and $X_j + \tilde{X}_j = X_k + \tilde{X}_k$ for all $1 \leq j, k \leq p$.*

2.2 Reconstructability and power

Why does the rank of G_{SDP} affect power? As an intuitive example, consider the equicorrelated case, where Lemma 2.2 tells us that one could reconstruct all information contained in the features X simply by looking at \tilde{X} and one other feature. This makes it difficult to assign feature importances to X versus \tilde{X} . For example, suppose Y follows a single-index model, meaning we can represent $Y = f(X\beta, U)$ for some deterministic function f and $U \sim \text{Unif}(0, 1)$ independent of X . Then Lemma 2.2 implies that for any $J \subset [p]$, we can write

$$X_J\beta_J = (X_1 + \tilde{X}_1) \left(\sum_{j \in J} \beta_j \right) - \tilde{X}_J\beta_J. \quad (4)$$

To see why this is a problem, initially assume $\sum_{j \in J} \beta_j = 0$, which guarantees that $X_J\beta_J = -\tilde{X}_J\beta_J$. In this setting, the feature importances $Z = z([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$ will have difficulty distinguishing between two different models, the correct model where $Y = f([X_J, X_{-J}](\beta_J, \beta_{-J}), U)$ and an incorrect model where $Y = f([\tilde{X}_J, X_{-J}](-\beta_J, \beta_{-J}), U)$. Note that in the correct model, we would expect $\{Z_j\}_{j \in J}$ to be large and $\{Z_{j+p}\}_{j \in J}$ to be near zero, leading to large positive $\{W_j\}_{j \in J}$, but in the incorrect model, we would expect the opposite, leading to highly negative $\{W_j\}_{j \in J}$. Since both models fit the data equally well and the models differ only in some of the signs of their coefficients, $\{W_j\}_{j \in J}$ is equally likely to be positive or negative unless the feature statistic function w incorporates prior information about the signs of β_J . For example, one could constrain the lasso to only assign nonnegative coefficient values to each feature. With enough data, the lasso would then correctly assign high feature importances to the set of features $\{j \in J : \beta_j > 0\}$. On the other hand, it would also ensure that the lasso would assign zero importance to features in the set $\{j \in J : \beta_j < 0\}$ and high importances to their knockoff counterparts, precluding the discovery of those features.

When $\sum_{j \in J} \beta_j \neq 0$, the incorrect model becomes

$$Y = f \left(\left[X_1, \tilde{X}_1, \tilde{X}_J, X_{-J} \right] \left(\sum_{j \in J} \beta_j, \sum_{j \in J} \beta_j, -\beta_J, \beta_{-J} \right), U \right).$$

In this case, regularized feature statistics like the lasso, with enough data, may eventually identify the correct model, whose coefficients have smaller ℓ_1 norm than the coefficients in the incorrect model. However, when the ℓ_1 norms of both options are similar, it will take a very large amount of data for the lasso to identify the true model (see Theorem 2.4). This exemplifies a broader theme, which is that the power of common sparsity-inducing feature statistics will decline further when it is possible to reconstruct features using a *sparse* subset of the other features and knockoffs. We refer to this phenomenon as *sparse reconstructability*, and in the general Gaussian case, it will occur when many eigenvalues of G_{SDP} are close to zero.

Gaussian equicorrelated designs with a single-index response are only an example of the broader reconstruction effect, which occurs quite generally: whenever X_j can be reconstructed from X_{-j}, \tilde{X} , feature importances such as those derived from a random forest are likely to confuse the contribution of a non-null X_j with other features or knockoffs. In the worst case, however, Theorem 2.3 indicates that if Y depends on X_J through a statistic $g(X_J)$ which can be reconstructed using some function g^* of \tilde{X}_J , no feature statistic will be able to use the data to distinguish between X_J and \tilde{X}_J . This theorem does not assume X to be Gaussian, nor does it assume Y to follow a single-index model.

Theorem 2.3. Suppose we can represent $Y = f(g(X_J), X_{-J}, U)$ for some set $J \subset [p]$, functions f and g , and independent noise $U \sim \text{Unif}(0, 1)$. Equivalently, this means $Y \perp\!\!\!\perp X_J \mid g(X_J), X_{-J}$. Suppose a function g^* exists such that

$$g(X_J) = g^*(\tilde{X}_J) \quad (5)$$

holds almost surely. If $Y^* = f(g^*(X_J), X_{-J}, U)$, then

$$([X, \tilde{X}], Y) \stackrel{d}{=} ([X, \tilde{X}]_{\text{swap}(J)}, Y^*) \quad \text{and} \quad ([X, \tilde{X}], Y^*) \stackrel{d}{=} ([X, \tilde{X}]_{\text{swap}(J)}, Y). \quad (6)$$

Furthermore, if we set $W = w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$, and $W^* = w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}^*)$, then for all $j \in J$,

$$\mathbb{P}(W_j > 0) + \mathbb{P}(W_j^* > 0) \leq 1. \quad (7)$$

For instance, in the equicorrelated case when $Y = f(X\beta, U)$ is single-index, SDP knockoffs satisfy the assumption $g(X_J) = X_J\beta_J = -\tilde{X}_J\beta_J = g^*(\tilde{X}_J)$ when $\sum_{j \in J} \beta_j = 0$. In this setting, equation (7) says that SDP knockoffs produce a no free lunch situation, where no feature statistic can have non-trivial power to select any $\{X_j\}_{j \in J}$ for both Y and Y^* . Indeed, equation (6) tells us that even a limitless amount of data gives us no way to distinguish between the features and the knockoffs in J , since any statistic which consistently selects X_J over \tilde{X}_J when the response comes from g faces *exactly* the opposite choice when the response comes from g^* . Therefore, any feature statistic which has nontrivial power to select X_J when the response follows g will have “negative” power (less power than if all features in J were null) when the response follows g^* . For example, any feature statistic which uses X_J to predict Y (using g) can do equally well using \tilde{X}_J to predict Y (using g^*). If we have no a priori bias towards g or g^* , then we will be equally likely to select the features X_J or their knockoffs \tilde{X}_J , making us powerless to detect any of $\{X_j\}_{j \in J}$, no matter the signal-to-noise ratio. Alternatively, a feature statistic which a priori biases towards (for example) g over g^* may gain power when the true model uses g , but as equation (7) indicates, it will correspondingly lose power when the true model uses g^* . Furthermore, a preference for sparsity will not help in general. For example, when $g(X_J) = X_J\beta_J = -\tilde{X}_J\beta_J = g^*(\tilde{X}_J)$, both models have the same sparsity.

To gain intuition about when equation (5) might hold, it may again be helpful to let $g(X_J) = X_J\beta_J$ for $\beta_J \in \mathbb{R}^{|J|}$. This includes (but is not limited to) the set of models where $Y \mid X$ is partially linear in X_J . In this case, equation (5) would be satisfied if a β_J^* exists such that $X_J\beta_J = \tilde{X}_J\beta_J^*$ holds almost surely. Although this condition may initially seem pathological, several remarks are in order here. First, even if this relation only holds approximately, reconstructability will still reduce the power of knockoffs in finite samples. For example, if one perturbed SDP knockoffs to prevent G_{SDP} from being *exactly* low rank, we can still have $X_J\beta_J \approx \tilde{X}_J\beta_J^*$ as long as some of the eigenvalues of G_{SDP} are small. Second, as the approximate rank of G_{SDP} decreases, the approximate null space of $[X_J, \tilde{X}_J]$ will grow larger, which makes it more likely that for any β_J , there may be *some* β_J^* such that $X_J\beta_J \approx \tilde{X}_J\beta_J^*$. This corresponds to the sparse reconstructability phenomenon we identified earlier, where it is possible to reconstruct each feature X_j using a small subset of the other features and knockoffs. Third, note that (5) does not require X_J to be reconstructable from \tilde{X}_J alone—for example, in the equicorrelated case, X_J is not reconstructable from \tilde{X}_J , but $X_J\beta_J = -\tilde{X}_J\beta_J$ anyway when $\sum_{j \in J} \beta_j = 0$. In general, certain classes of statistics $g(X_J)$ can be reconstructable from \tilde{X}_J even when reconstructing X_J requires joint information from X_{-J}, \tilde{X} . For this reason, it is important to prevent X_J from being reconstructable from any of X_{-J}, \tilde{X} . Lastly, as p grows larger, there are many more subsets J for which (5) can hold, especially when the distribution of (X, Y) does not have any special structure. As a result, we expect this phenomenon to be more frequent in higher dimensions.

To briefly summarize, so far, we have shown that the MAC heuristic often causes the diagonal entries of the S matrix to become so large that many eigenvalues of the joint covariance matrix G_S become quite small. Unfortunately, this creates strong joint dependencies in the distribution of $[X, \tilde{X}]$. In practice, this means that many features X_j can largely be reconstructed from the knockoffs \tilde{X} and the other features X_{-j} in spite of low marginal correlations between X_j and \tilde{X}_j . As a result, knockoff feature importances (such as lasso coefficients) may ignore a non-null feature X_j and instead use X_{-j}, \tilde{X} to reconstruct its effect on Y , dramatically reducing the power of knockoffs. In Section 2.3, we show that the reconstruction effect can actually render knockoffs powerless in some settings, and as we will see in Section 4, this phenomenon occurs quite broadly, even when X is not Gaussian.

2.3 Main result for equicorrelated Gaussian designs

Our main result for equicorrelated Gaussian designs and single-index response models largely follows from the intuition that as $p \rightarrow \infty$, there will often be a very large number of subsets J such that $\sum_{j \in J} \beta_j \approx 0$, unless β has very special structure. Theorem 2.3 tells us that when Σ is equicorrelated, the presence of many such J will dramatically reduce the power of knockoffs unless the feature statistic function w encodes specific information about each of the signs of the non-null coefficients. To exclude this case, we impose a mild condition on w which prevents it from treating any of the features of X differently based on their position, and therefore prevents w from incorporating different a priori information about features with positive and negative signs. In particular, we define a knockoff feature statistic function w to be *permutation invariant* if and only if for any permutation $\sigma : [p] \rightarrow [p]$,

$$\sigma(w([\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})) = w([\sigma(\mathbf{X}), \sigma(\tilde{\mathbf{X}})], \mathbf{y}). \quad (8)$$

Since some feature statistics are randomized functions, our proofs also allow for equation (8) to hold in distribution conditional on the data. Note almost all feature statistics in the knockoffs literature, including those derived from lasso coefficients and even neural networks, satisfy this property. To quickly define notation, we denote the average power by:

$$\text{Power}(w, \beta) = \frac{\mathbb{E} [\left| \hat{S} \cap \{j : \beta_j \neq 0\} \right|]}{|\{j : \beta_j \neq 0\}|}$$

where the expectation above is over the data for a fixed feature statistic function w and a fixed vector of coefficients β . As a reminder, \hat{S} is the set of selected features.

Theorem 2.4. *Let X be an equicorrelated Gaussian design with correlation $\rho \geq 0.5$. Assume we can represent $Y = f(X\beta + \zeta, U)$ for $\zeta \sim \mathcal{N}(0, \sigma_0^2)$ and $U \sim \text{Unif}(0, 1)$ independent of the data. Let \mathcal{W} be the class of all permutation invariant feature statistic functions and suppose we aim to control the FDR at level $q \leq 0.1$ using SDP knockoffs.*

Sample $\beta^{(n)} \in \mathbb{R}^p$ uniformly from a p -dimensional hypercube centered at 0 with a fixed side-length. Let n be the number of data points and suppose there exists some $\epsilon > 0$ such that $n = o(p^{2-\epsilon})$. Then as $n, p \rightarrow \infty$,

$$\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \xrightarrow{P} 0.$$

Theorem 2.4 follows from a more general statement which applies to block-diagonal Σ whose blocks are equicorrelated, which we prove in Appendix A. Note that neither the precise distribution of β nor the assumption that $q \leq 0.1$ are essential, as we discuss in Appendix A.5.

This theorem tells us that even in low-dimensional regimes, SDP knockoffs may have asymptotically zero average power, even when the feature statistic w is chosen with oracle knowledge of the response model $Y | X$. This theorem applies to a variety of familiar single-index models, including Gaussian linear models and probit regression, which satisfy the assumption that $Y = f(X\beta + \zeta, U)$. Furthermore, since $\text{Var}(\zeta) = \sigma_0^2$ can be arbitrarily small, we expect this conclusion to hold for most single-index and generalized linear models, which take the form $Y = f(X\beta, U)$.

Of course, real datasets will rarely be exactly equicorrelated. The point of Theorem 2.4 is to show that the reconstruction effect can prevent SDP knockoffs from discovering many non-null features, no matter the signal size or feature statistic. However, as we discuss at length in Section 2.2, the reconstruction effect can occur without equicorrelated features, and indeed, our empirical results in Section 4 show that SDP knockoffs lose power in a variety of practical settings.

2.4 SDP knockoffs infer signal magnitudes but not signs

To help understand why SDP knockoffs fail in the equicorrelated setting, we briefly analyze what SDP knockoffs do *right*. In particular, SDP knockoffs can easily be used to detect that *either* X_j or \tilde{X}_j has an

effect on Y , but not to choose between them. This means that when using SDP knockoffs, non-null W_j will often have the “right” magnitude but a negative sign.

Proposition 2.1. *Assume X is equicorrelated and $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$ with $n > 2p$ and correlation $\rho \geq 0.5$. Suppose we generate knockoffs such that the feature-knockoff correlation is constant and the MAC equals v . If $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$, then the mean squared error of $\hat{\beta}_{1:p}^{(\text{ext})} - \hat{\beta}_{(p+1):2p}^{(\text{ext})}$ as an estimator of β is increasing in v .*

This theorem proves that for OLS coefficients, the differences $\hat{\beta}_j^{(\text{ext})} - \hat{\beta}_{j+p}^{(\text{ext})}$ proxy β_j better as the marginal correlations between X_j and \tilde{X}_j get smaller, apparently in support of the MAC heuristic. Unfortunately, setting $W_j = \hat{\beta}_j^{(\text{ext})} - \hat{\beta}_{j+p}^{(\text{ext})}$ is a poor feature statistic when β_j may be negative, since W_j will likely be negative when β_j is negative, preventing the selection of features with negative coefficients. On the other hand, if we set $W_j = |\hat{\beta}_j^{(\text{ext})}| - |\hat{\beta}_{j+p}^{(\text{ext})}|$ to fix this problem, knockoffs will still have extremely low power as the MAC v gets smaller because $\hat{\beta}_j^{(\text{ext})}$ and $\hat{\beta}_{j+p}^{(\text{ext})}$ individually will have high variance, making W_j (approximately) equally likely to be positive or negative.

Penalized regressions like lasso coefficients $\hat{\beta}^{(\ell_1, \text{ext})}$ exhibit the same behavior. Consider how the lasso assigns coefficient values to a pair of features $\hat{\beta}_j^{(\ell_1, \text{ext})}, \hat{\beta}_k^{(\ell_1, \text{ext})}$ and their knockoffs $\hat{\beta}_{j+p}^{(\ell_1, \text{ext})}, \hat{\beta}_{k+p}^{(\ell_1, \text{ext})}$ when $\beta_j + \beta_k = 0$. As $n \rightarrow \infty$, the lasso will have to choose between at least two options: setting $(\hat{\beta}_j^{(\ell_1, \text{ext})}, \hat{\beta}_{j+p}^{(\ell_1, \text{ext})}) \approx (\beta_j, 0)$ and $(\hat{\beta}_k^{(\ell_1, \text{ext})}, \hat{\beta}_{k+p}^{(\ell_1, \text{ext})}) \approx (\beta_k, 0)$, or alternatively setting $(\hat{\beta}_j^{(\ell_1, \text{ext})}, \hat{\beta}_{j+p}^{(\ell_1, \text{ext})}) \approx (0, \beta_k)$ and $(\hat{\beta}_k^{(\ell_1, \text{ext})}, \hat{\beta}_{k+p}^{(\ell_1, \text{ext})}) \approx (0, \beta_j)$. When using SDP knockoffs, both options deterministically have the same empirical mean-squared error as well as the same ℓ_1 norm, so the lasso genuinely cannot choose between these two options. Moreover, this applies to *every* j, k such that $\beta_j + \beta_k = 0$. Note that since $W_j = |\hat{\beta}_j^{(\ell_1, \text{ext})}| - |\hat{\beta}_{j+p}^{(\ell_1, \text{ext})}|$ are symmetric but have large magnitudes, they actually make it more difficult to discover other features, since knockoffs can only make discoveries when the W_j with large magnitudes have consistently positive signs (see equation (3) or Figure 2 for an illustration of this).

2.5 Relationship with the literature

At this point, we pause to put our results on the reconstruction effect in context with the rest of the literature. To start, we note that the reconstruction effect appears to differ fundamentally from the “alternating sign effect” introduced by Chen et al. (2019), who give heuristic evidence that lasso-path statistics may lose power when two features are positively correlated and their coefficients have opposite or “alternating” signs. For instance, the alternating sign effect only plagues certain feature statistics such as lasso-path statistics, and the initial version of Chen et al. (2019) shows that simple statistics like OLS statistics do not suffer from this effect. In contrast, Theorem 2.4 shows that minimizing the MAC creates an *unidentifiability* problem which makes nearly *every* feature statistic asymptotically powerless in the equicorrelated case. Second, the reconstruction effect does not even rely on the presence of alternating signs in β —for example, we have shown in Theorem 2.3 that it can occur outside of single-index models, where no coefficient vector β exists.

Furthermore, Theorem 2.4 may seem to contradict Theorem 1 of Fan et al. (2020), which proves that when X is Gaussian with a Gaussian linear response Y , the power of lasso-based knockoffs approaches one asymptotically. However, that theorem assumes the regularity condition that $\lambda_{\min}(\Sigma)$ and $\lambda_{\min}(2S - S\Sigma^{-1}S)$ must be bounded away from zero. Yet Lemma 2.1 demonstrates that for SDP knockoffs, G_{SDP} and therefore its Schur complement $2S - S\Sigma^{-1}S$ will often be low rank, meaning the regularity conditions used by Fan et al. (2020) will not hold. This suggests that we ought to use a different heuristic than minimizing the MAC, one that will automatically ensure $2S - S\Sigma^{-1}S$ is full rank whenever Σ is. In the next section, we will define two heuristics which satisfy this property. Furthermore, we will prove that one of these heuristics is consistent under only the assumption that $\lambda_{\min}(\Sigma)$ is bounded away from zero. This result parallels Theorem 1 of Fan et al. (2020), except we use different technical tools to avoid requiring the problematic assumption that $\lambda_{\min}(2S - S\Sigma^{-1}S)$ is bounded above zero.

Like Fan et al. (2020), Liu and Rigolet (2019) observed that the consistency of knockoffs may depend

on properties of the joint covariance matrix G_S . Although Liu and Rigollet (2019) did not analyze the power of SDP knockoffs, they proposed generating “conditional independence” (CI) knockoffs such that $X_j \perp\!\!\!\perp \tilde{X}_j | X_{-j}$, and they proved that some lasso-based feature statistics applied to CI knockoffs are consistent in Gaussian linear models. Unfortunately, the CI condition says little about whether X_j is (approximately) reconstructable using *joint* information from X_{-j}, \tilde{X} , and as a result, we show in Appendix F.2 that CI knockoffs actually can suffer from the same reconstructability problems that SDP knockoffs suffer from. This means that CI knockoffs may have low power in finite samples even if they are consistent asymptotically. Furthermore, CI knockoffs only exist for a restricted set of Gaussian designs, outside of which they violate the pairwise exchangeability condition (1). While Ke et al. (2020) recently suggested an extension of CI knockoffs to the general Gaussian case, their method involves computing S_{CI} without constraining $G_S \succcurlyeq 0$ and then performing a binary search to find the maximum $\gamma \in [0, 1]$ such that $G_{\gamma \cdot S_{\text{CI}}} \succcurlyeq 0$. Unfortunately, this new matrix $\gamma \cdot S_{\text{CI}}$ lacks any conditional independence properties, and furthermore, γ can be extremely small in practice, substantially reducing power. We discuss this more in Appendix F.2.

Lastly, Theorem 5.4 of Ke et al. (2020) proves that SDP (fixed-X) knockoffs match the power of an “oracle” procedure for a positive, block-equicorrelated correlation structure. However, Ke et al. (2020) assume a fixed block-size of $\ell = 2$ and vanishing sparsity, meaning that asymptotically, a vanishing proportion of the non-nulls lie in the same equicorrelated block as other non-nulls. Since the reconstruction effect for block-equicorrelated designs only occurs when multiple non-null features are correlated, it will asymptotically not occur in this regime.

3 Minimum reconstructability knockoffs

In Section 2, we demonstrated that SDP knockoffs lack power in the equicorrelated case because they make it possible to “reconstruct” a feature X_j using X_{-j}, \tilde{X} . To fix this problem, we suggest constructing \tilde{X} in order to minimize the reconstructability (MRC) of each feature X_j given the other features X_{-j} and the knockoffs \tilde{X} . In this section, we describe two instantiations of this framework which can be efficiently computed when X is Gaussian and perform well in a variety of settings.

3.1 Minimum variance-based reconstructability (MVR) knockoffs

One intuitive way to minimize “reconstructability” is to maximize the conditional variance $\text{Var}(X_j | X_{-j}, \tilde{X})$ for each $j \in [p]$. This motivates the minimum variance-based reconstructability (MVR) knockoff construction.

Definition 3.1 (Minimum Variance-Based Reconstructability (MVR) Knockoffs). *To sample MVR knockoffs, sample $\tilde{X} | X$ so as to minimize*

$$L_{\text{MVR}} = \sum_{j=1}^p \frac{1}{\mathbb{E}[\text{Var}(X_j | X_{-j}, \tilde{X})]}. \quad (9)$$

MVR knockoffs minimize the *inverse* expected conditional variances $\frac{1}{\mathbb{E}[\text{Var}(X_j | X_{-j}, \tilde{X})]}$ in order to harshly penalize high levels of reconstructability and ensure $\mathbb{E}[\text{Var}(X_j | X_{-j}, \tilde{X})] > 0$ for all j . This is important because high levels of reconstructability often cause some feature statistics to have large magnitudes but negative signs, which dramatically reduces the power of knockoffs. For example, if X_j is non-null and highly reconstructable, then feature statistics can easily reconstruct the effect of X_j on Y using X_{-j}, \tilde{X} , especially when the number of data points is small. When this happens, X_j will likely be assigned a low feature importance and, more importantly, a knockoff variable such as \tilde{X}_k or \tilde{X}_j may be assigned a *high* feature importance corresponding to the effect of X_j on Y . This implies that the feature statistic W_k (resp. W_j) may have a large magnitude but a negative sign. Ultimately, the knockoff filter can only make rejections if there exists some t such that approximately $\frac{t}{1+q}$ of the t feature statistics with the largest absolute values have positive signs. Thus, when even a few feature statistics have large magnitudes but negative signs,

knockoffs may have very low power. We illustrate this effect in Figure 2 for equicorrelated designs. We discuss this phenomenon more in Appendix C.3, where we also show that the same phenomenon occurs for a diverse range of design matrices, including Gaussian Markov chains and Gaussian designs where the covariance matrix is 80% sparse.

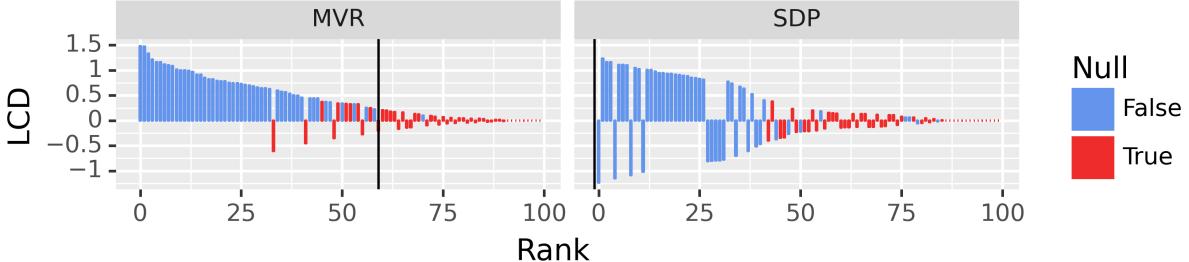


Figure 2: W -statistic values for MVR vs. SDP knockoffs. We plot LCD feature statistics sorted in descending order of absolute value for an exchangeable Gaussian design with correlation $\rho = 0.6$, $n = 190$, $p = 100$, and $Y | X \sim \mathcal{N}(X\beta, 1)$ with 50 non-null coefficients sampled as independent random signs ± 1 . The black line denotes the data dependent threshold T . SDP knockoffs have low power because many feature statistics have large magnitudes but negative signs.

In the Gaussian case when $[X, \tilde{X}] \sim \mathcal{N}(0, G_S)$, we can write L_{MVR} as a simple and convex function of the S -matrix. In particular, $L_{\text{MVR}}(S) = \frac{1}{2} \text{Tr}(G_S^{-1})$, as $1/\text{Var}(X_j | X_{-j}, \tilde{X}) = (G_S^{-1})_{j,j}$ for Gaussian features (see Anderson (2009), Section 2.5). See Appendix D.1 for a proof of convexity. This convenient formulation allows us to prove two appealing properties of MVR knockoffs when $X \sim \mathcal{N}(0, \Sigma)$, $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$, and we use the absolute value of OLS coefficients as feature importances.

First, we prove a type of optimality result for MVR knockoffs. Although we would like to know which S -matrix maximizes power to detect non-nulls, the exact answer depends on the unknown coefficients β . As a proxy for this, however, we note that the power of knockoffs almost entirely depends on the accuracy of the OLS coefficients $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$. As $n \rightarrow \infty$, we hope that $\hat{\beta}_{1:p}^{(\text{ext})}$ will converge to β and the knockoff feature importances $\hat{\beta}_{(p+1):2p}^{(\text{ext})}$ will converge to 0. In finite samples, it turns out that for *any* Σ and β , MVR knockoffs minimize the mean-squared error between $\hat{\beta}^{(\text{ext})}$ and its target.

Proposition 3.1. *Suppose $X \sim \mathcal{N}(0, \Sigma)$ for any Σ and $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$. Let $\beta^{(\text{ext})} \in \mathbb{R}^{2p}$ be the concatenation of $\beta \in \mathbb{R}^p$ with p zeros. Suppose $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$ and $n > 2p + 1$. Then*

$$S_{\text{MVR}} = \arg \min_S \mathbb{E}[||\hat{\beta}^{(\text{ext})} - \beta^{(\text{ext})}||_2^2].$$

Second, we prove that the MVR knockoff filter is consistent in low-dimensions for OLS feature statistics. This result would not be surprising except that we showed in Section 2 that in some low-dimensional settings, SDP knockoffs is inconsistent for *every* feature statistic.

Theorem 3.1. *Suppose $X \sim \mathcal{N}(0, \Sigma^{(n)})$, $Y | X \sim \mathcal{N}(X\beta^{(n)}, \sigma^2)$, and \tilde{X} is generated using S_{MVR} . Suppose $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$, and set $w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) = |\hat{\beta}_{1:p}^{(\text{ext})}| - |\hat{\beta}_{(p+1):2p}^{(\text{ext})}|$.*

Let $n, p \rightarrow \infty$ such that $p = o(n)$ and consider a sequence of covariance matrices $\Sigma^{(n)} \in \mathbb{R}^{p \times p}$ such that the minimum eigenvalue of $\Sigma^{(n)}$ is bounded above a fixed constant $\gamma \in \mathbb{R}^+$. Suppose we sample a sequence of random $\beta^{(n)}$ as follows. Let all but a uniformly drawn subset of $\lceil s_0 p \rceil$ entries of $\beta^{(n)}$ equal zero, for a fixed constant $s_0 \in (0, 1]$, and then sample the remaining (non-null) entries of $\beta^{(n)}$ from a $\lceil s_0 p \rceil$ -dimensional hypercube centered at 0 with any fixed side-length. Then

$$\text{Power}(w, \beta^{(n)}) \xrightarrow{P} 1.$$

Note that the equicorrelated case discussed in Section 2 satisfies these regularity conditions on $\Sigma^{(n)}$, since its eigenvalues are bounded uniformly above $1 - \rho$. As noted in Section 2.5, this theorem proves a similar consistency result to Fan et al. (2020), except our proof uses properties of MVR knockoffs to avoid assumptions about the minimum eigenvalue of G_S .

This result should also make intuitive sense in the context of Liu and Rigollet (2019), who show that knockoffs are consistent if and only if (informally) the diagonals of $\frac{1}{n}G_S^{-1}$ converge to 0. Technically, Liu and Rigollet (2019)'s theory does not directly apply to our setting, as it assumes a different asymptotic regime and relies on regularity conditions which the authors themselves admit are “highly nontrivial” to verify unless $\Sigma^{(n)}$ is block-diagonal. We use different technical tools to avoid these assumptions. Conceptually, however, the key novelty of Theorem 3.1 is that it shows for the first time that a concrete method—MVR knockoffs—can achieve Liu and Rigollet (2019)'s condition with minimal assumptions on $\Sigma^{(n)}$. To summarize, Liu and Rigollet (2019)'s theory established the importance of making the diagonals of G_S^{-1} small, and MVR knockoffs explicitly follows this principle by minimizing $\text{Tr}(G_S^{-1})$, allowing us to prove its consistency in more generality than has been established for any other knockoff generation method.

Lastly, we note that the convexity of L_{MVR} in the Gaussian case allows us to develop an algorithm to compute S_{MVR} in $O(n_{\text{iter}}p^3)$, which is the same time complexity or faster than the methods used to compute SDP knockoffs Askari et al. (2020). Our method is inspired by Askari et al. (2020), who introduced a coordinate-descent algorithm to compute S_{SDP} efficiently. The key idea is to use rank-one updates to maintain a running Cholesky decomposition of $2\Sigma - S$. Our algorithm uses this same strategy, although extending their ideas to the MVR loss requires some nontrivial additional analysis. For brevity, we defer the details to Appendix D. Since the algorithms used to compute MVR and SDP knockoffs are extremely similar, we show in simulations in Appendix D.6 that not only do they have the same computational complexity, but they have very similar runtimes in practice. Additionally, like the SDP formulation (2.1), our algorithm can take advantage of block-diagonal approximations (Candès et al., 2018) or low-rank factor structure in Σ (Askari et al., 2020) to dramatically speed up computations to be linear in p —see Appendix D.5 for additional details. The overall point is that there is no computational reason to prefer either MVR or SDP knockoffs.

3.2 Maximum entropy (ME) knockoffs

An information-theoretic alternative to MVR knockoffs is to minimize the mutual information between X and \tilde{X} ; as we shall see in a moment, this is equivalent to maximizing the entropy of $[X, \tilde{X}]$. Gimenez and Zou (2019) have previously considered ME knockoffs, but only in the Gaussian case, and they introduce this method for a very different reason than we do. In particular, the authors motivate ME knockoffs by observing that the SDP knockoff construction can induce sparsity in the diagonal of S_{SDP} , meaning that some values of the diagonal of S_{SDP} become too small to distinguish between features and knockoffs. Although they demonstrate that ME knockoffs will not suffer from this problem, they only compare the power of the two methods in a single case where the distributions of both X and $Y | X$ are chosen adversarially against SDP knockoffs. In contrast, we advocate the use of ME knockoffs to solve almost exactly the opposite problem: that the values of S_{SDP} are frequently *too large*. Indeed, our most dramatic empirical results come in the equicorrelated case, where $S_{\text{SDP}} = \gamma \cdot I_p$ for γ as large as 1. Clearly, in this case, the diagonal of S_{SDP} is not sparse at all.

Definition 3.2 (Maximum Entropy (ME) Knockoffs). *Suppose X is absolutely continuous on \mathcal{X} with respect to a base measure μ with density $p(x)$. To sample ME knockoffs, sample $\tilde{X} | X$ so as to minimize*

$$L_{\text{ME}} = \int_{x \in \mathcal{X}} \int_{\tilde{x} \in \mathcal{X}} p(x, \tilde{x}) \log(p(x, \tilde{x})) d\mu(\tilde{x}) d\mu(x), \quad (10)$$

where \mathcal{X} is the support of X , and $p(x, \tilde{x})$ is the joint density of $[X, \tilde{X}]$. Note L_{ME} corresponds to the negative entropy of $[X, \tilde{X}]$. If $[X, \tilde{X}]$ admits no joint density with respect to the product measure $\mu \times \mu$, we adopt the convention that $L_{\text{ME}} = \infty$.

Given this definition, it is not immediately obvious that ME knockoffs minimize any notion of reconstructability. Of course, the entropy of $[X, \tilde{X}]$ equals twice the entropy of X minus the mutual information between X and \tilde{X} , so ME knockoffs also minimize the mutual information between X and \tilde{X} . Since mutual information can account for joint dependencies between X and \tilde{X} , we might expect it to perform better than the MAC metric, which only looks marginally at dependencies between X_j and \tilde{X}_j . However, mutual information still

may not necessarily capture the right notion of reconstructability. To perform feature selection with knockoffs, one must assign a feature importance to each individual feature and knockoff. To do this powerfully, we have argued that *each* non-null feature X_j must contain information that cannot be reconstructed using X_{-j} and \tilde{X} . Although mutual information captures some notion of the aggregate dependencies between X and \tilde{X} , it is not clear it coincides with this feature-level definition of reconstructability.

To better connect ME knockoffs with reconstructability, we turn to the setting where X is Gaussian, where L_{ME} has a simple formulation. In particular, we can take \tilde{X} to be jointly Gaussian with X , and then the entropy of $[X, \tilde{X}]$ equals $\log \det(G_S)$ up to a constant. Thus, maximizing the entropy of $[X, \tilde{X}]$ corresponds to minimizing $\log \det(G_S^{-1})$. This loss is actually quite similar to the MVR loss for Gaussian X , as both losses are convex, elementwise-decreasing functions of the eigenvalues of G_S . In particular,

$$L_{\text{MVR}}(S) \propto \text{Tr}(G_S^{-1}) = \sum_{j=1}^{2p} \frac{1}{\lambda_j(G_S)} \quad \text{and} \quad L_{\text{ME}}(S) = \log \det(G_S^{-1}) = \sum_{j=1}^{2p} \log \left(\frac{1}{\lambda_j(G_S)} \right), \quad (11)$$

where we express the L_{ME} loss up to an additive constant. As a result, S_{ME} and S_{MVR} are quite similar in the Gaussian case, suggesting that L_{ME} captures a similar notion of reconstructability in this setting. This also indicates that ME and MVR knockoffs will have very similar power in the Gaussian setting, which we confirm in Section 4. In Appendix D.3, we modify the algorithm which computes S_{MVR} to compute S_{ME} in the same time complexity.

Outside the Gaussian case, we cannot write L_{ME} so simply, but we do prove in Lemma 3.2 that when X has finite support, ME knockoffs will not allow any feature to be perfectly reconstructable as long as any knockoff procedure can accomplish this. Intuitively, this result holds because if X_j is nearly a deterministic function of X_{-j}, \tilde{X} , then almost all of the probability density of $[X, \tilde{X}]$ will lie along a $(2p - 1)$ dimensional subset of $\mathcal{X} \times \mathcal{X}$. To ensure the density $p(x, \tilde{x})$ integrates to one over $\mathcal{X} \times \mathcal{X}$, the average log value of $p(x, \tilde{x})$ must become larger than is optimal. See Appendix E for a more detailed discussion of the discrete case as well as a proof of Lemma 3.2.

Lemma 3.2. *Let X have finite support \mathcal{X} . For any $j \in [p]$ and any $x, \tilde{x} \in \mathcal{X} \times \mathcal{X}$, ME knockoffs satisfy $\text{Var}(X_j | X_{-j} = x_{-j}, \tilde{X} = \tilde{x}) > 0$, so long as this property does not contradict the definition of valid knockoffs.*

At this point, one might wonder whether there are clear grounds to prefer MVR knockoffs over ME knockoffs or vice versa. Our empirical results in Section 4 suggest MVR and ME knockoffs perform very similarly in the Gaussian case. There is more theoretical work to be done to better understand their relationship in general, but we can make two comments summarizing what we know (and do not know) so far. First, ME knockoffs seem to minimize some notion of reconstructability in all examples we can tractably analyze, but it is not clear that our arguments generalize beyond the Gaussian case. We include ME knockoffs in this section for the reader's consideration because they seem promising and they perform well empirically in Section 4. Second, MVR knockoffs are appealing because they enjoy exact optimality properties with OLS feature importances and are more readily proved to be consistent, as shown in Section 3.1. Note that since ME knockoffs are not identical to MVR knockoffs, they do not enjoy this exact optimality, and it is difficult to prove the consistency of ME knockoffs for general sequences of Σ . That said, the solutions to MVR and ME knockoffs are asymptotically identical in the case where Σ is equicorrelated, so consistency and (approximate) optimality do hold for ME knockoffs for exchangeable Gaussian designs—see Appendix D.4 for a precise statement.

4 Empirical results

In this section, we run an extensive set of simulations to demonstrate the power of MVR and ME knockoffs. To do this, we developed a new open source python package `knockpy`, which implements a host of methods from the knockoffs literature, including a wide variety of feature statistics and knockoff sampling mechanisms for both fixed-X and model-X knockoffs. For example, `knockpy` includes a fully general Metropolized knockoff sampler (Bates et al., 2020) which can be multiple orders of magnitude faster than previous implementations.

Most importantly, `knockpy` is written to be modular, so researchers can easily tweak its functionality or add other features on top of it. We discuss `knockpy` further in Appendix G. All additional code for our simulations is available at <https://github.com/amspector100/mrcrep>.

We aim to demonstrate that the MRC framework offers real advantages over MAC-minimizing knockoffs in very practical settings, even when minimizing the MAC does not result in exact reconstructability. At the outset, we highlight some key conclusions.

Power for linear responses: Both MVR and ME knockoffs generally outperform SDP knockoffs in the setting where the design X is Gaussian and $Y | X$ is linear. This result holds for a range of highly correlated covariance matrices and feature statistics, although as the features become less correlated, the performances of all three methods tend to equalize. We note further that MRC knockoffs frequently outperform their SDP counterparts by very large margins (as much as 100 percentage points). Although there are examples where SDP knockoffs outperform MVR and ME knockoffs, these are rare, and even in such cases, SDP knockoffs usually outperform MVR and ME knockoffs by only a small margin.

Power for nonlinear responses: The preceding paragraph's conclusions *also* hold when the conditional distribution $Y | X$ is highly nonlinear, even in cases where Y does not follow a single-index model and linear feature statistics such as the lasso have zero power. Even very complex feature statistics like random forests with swap importances (Gimenez et al., 2019) or DeepPINK (Lu et al., 2018) perform better with MVR and ME knockoffs.

MVR vs. ME knockoffs: As our theory predicts, MVR and ME knockoffs perform similarly in the Gaussian setting, although there are a few cases where ME knockoffs slightly outperform MVR knockoffs. From now on, we use “MRC knockoffs” to refer to both MVR and ME knockoffs at once.

Robustness: Although we did not study robustness theoretically, we found empirically that MRC knockoffs can be both powerful and robust in the setting where X is Gaussian but the covariance matrix Σ is unknown and estimated using the data. In particular, in the high-dimensional setting where $\hat{\Sigma}$ is estimated with a shrinkage estimator, MRC knockoffs appear to violate FDR control less severely than SDP knockoffs, although we have no theoretical reason to believe that MRC knockoffs are in general more robust than SDP knockoffs.

Power for fixed-X (FX) knockoffs: It is straightforward to define MRC knockoffs for the fixed-X knockoff filter of Barber and Candès (2015). We demonstrate that MRC FX knockoffs are generally more powerful than their SDP counterparts.

Power for non-Gaussian designs: MRC-inspired knockoffs even perform well in the setting where the features X are not Gaussian, both for second-order knockoffs and as a guide for the Metropolized knockoff sampler. We will carefully define these generalizations in Section 4.6.

In general, we will only plot average power, since knockoffs provably control the FDR. The exception to this, of course, is when we discuss the robustness of the knockoffs procedure. Unless otherwise specified, we use knockoffs to control the FDR at level $q = 0.1$. In all examples, we use Algorithms 1 and 2, detailed in Appendix D, to compute MVR and ME knockoffs. In all instances, our plots include two standard deviation error bars, although in some cases the error bars are so small they are difficult to see. We also always plot both MVR and ME knockoffs, although often the two methods have nearly identical performance, causing their power curves to entirely overlap.

4.1 Simulations on equicorrelated designs

In this subsection, we investigate the performance of MRC and SDP knockoffs in the equicorrelated setting studied in Section 2. In Figure 3, we run simulations where $X \sim \mathcal{N}(0, \Sigma)$ with Σ equicorrelated and ρ varied between 0 and 0.9 for both a Gaussian and logistic linear response (see the caption for details). We compare four types of knockoffs: MVR, ME, SDP, and a perturbed version of SDP knockoffs where we set

$$S_{\text{TOL}} = \gamma \cdot S_{\text{SDP}}. \quad (12)$$

We refer to this version as the “sdp_tol” option because it ensures that the minimum eigenvalue of G_{TOL} is above a tolerance $1 - \gamma$. When $\rho \geq 0.5$, we set $\gamma = 0.99$ to ensure that $\lambda_{\min}(G_{\text{TOL}}) \geq 0.01$. We only do this in the case where $\rho \geq 0.5$, since otherwise G_{SDP} will be full rank anyway. As our feature statistics, we use cross-validated lasso and ridge absolute coefficient differences of the form $W_j = |\hat{\beta}_j^{(\text{ext})}| - |\hat{\beta}_{j+p}^{(\text{ext})}|$, where $\hat{\beta}_k^{(\text{ext})}$ refers to the k th estimated lasso or ridge coefficient, respectively, for $k \in [2p]$.

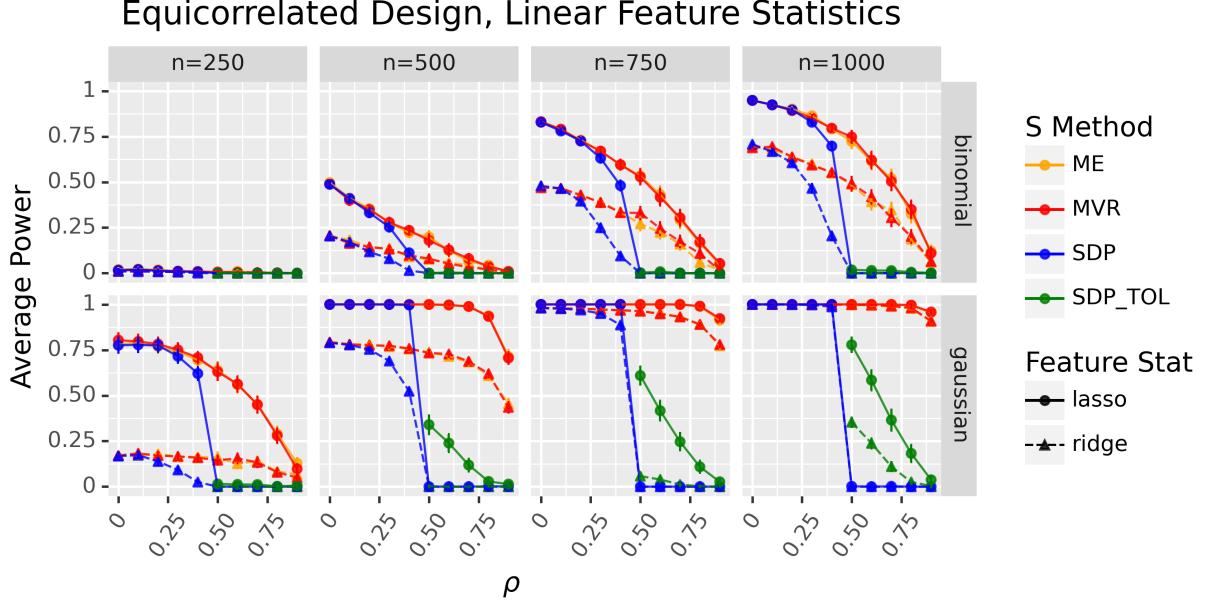


Figure 3: Empirical powers when X is an exchangeable Gaussian design. We set $p = 500$ and vary the correlation ρ between 0 and 0.9 with the number of data points n on the x-facets. In the lower panel, $Y | X \sim \mathcal{N}(X\beta, 1)$ and in the upper panel $Y | X \sim \text{Bin}(\pi(X\beta))$ where π is the sigmoid function. In all cases the number of non-nulls is 50 and the non-null values are sampled independently from $\text{Unif}([-1, -1/2] \cup [1/2, 1])$. We use lasso and ridge coefficient differences (and the logistic versions in the binomial case) and compare the performance of MVR, ME, and SDP knockoffs. Note SDP_TOL knockoffs modify the SDP algorithm so that $\lambda_{\min}(G_S) \geq 0.01$ —see equation (12).

Figure 3 demonstrates that MRC knockoffs substantially outperform SDP knockoffs, even when G_{SDP} is not low rank. First, both MRC knockoff types uniformly outperform the perturbed SDP knockoffs, even though the perturbed SDP knockoffs outperform the exact SDP knockoffs. Moreover, *even when $\rho < 0.5$ and G_{SDP} is full rank, MRC knockoffs can have much more power than SDP knockoffs*: for example, when $\rho = 0.4$ and $n = 750$, logistic ridge has over twice the power when applied to MRC knockoffs versus SDP knockoffs. We note that this power gain in highly correlated settings does not come at the cost of power in settings with weaker correlations, since MRC knockoffs outperform or match the power of SDP knockoffs for every value of ρ .

Note that MVR and ME knockoffs provably yield the same solution asymptotically in this setting (see Appendix D.4), which is why their power curves almost entirely overlap in Figure 3.

Since the perturbed SDP knockoffs outperform the SDP knockoffs, one might wonder how the choice of γ affects power. To analyze this question, we perform a line search over all S -matrices which can be represented as a scaled identity matrix. In particular, we set

$$S_\gamma = \gamma \cdot (2 - 2\rho) \cdot I_p. \quad (13)$$

Here, we can think of γ as interpolating between the minimum possible S -matrix where $S = 0 \cdot I_p$ and the “maximum” S -matrix, where $S = 2\lambda_{\min}(\Sigma) \cdot I_p = (2 - 2\rho) \cdot I_p$. The results indicate that up to Monte Carlo error, MVR and ME knockoffs have the highest power among all such S -matrices in this setting for both lasso and ridge feature statistics. Notably, this result holds for all $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, whereas SDP knockoffs lose power compared to MVR knockoffs for all ρ except $\rho = 0.1$, where all methods have nearly the same performance. For brevity, we present these results in Appendix F.1.

Lastly, we note that two previous papers (Xing et al., 2019; Dai et al., 2020) have observed that in the equicorrelated case, knockoffs have surprisingly low power when compared to other feature selection methods. Since these papers used the SDP formulation, in Appendix F.3, we exactly replicate these simulations but use MRC knockoffs instead. Our results show that MRC knockoffs have comparable or higher power than all of the other methods used in these papers.

4.2 Simulations on Gaussian designs

In the previous section, we investigated the performance of MRC and SDP knockoffs on equicorrelated Gaussian designs with linear responses. In this section, we let $X \sim \mathcal{N}(0, \Sigma)$ but vary Σ extensively. Furthermore, we allow the conditional distribution $Y | X$ to be highly nonlinear.

We defer a precise description of the covariance matrices Σ in our simulations to Appendix F.4. However, we will give a brief overview of the types of Σ in order to emphasize that these Σ differ substantially from the equicorrelated case. In the “AR1” setting, for example, X is a standardized Gaussian Markov chain with correlations $\text{Cov}(X_j, X_{j+1})$ sampled from Beta(3, 1). The “AR1 (Corr)” is the same setting except we cluster the non-nulls together along the chain, as we might expect in, e.g., genetic studies. Then, in the “ER (Cov)” and “ER (Precision)” settings, the covariance and precision matrices (respectively) are 80% sparse, where the nonzero entries are chosen uniformly at random, in accordance with an ErdosRenyi (ER) procedure. Lastly, we include simulations when Σ is block-equicorrelated with a block size of 5 and within-block correlations of $\rho = 0.5$.

First, in Figure 4, we compare the power of MVR, ME, and SDP knockoffs when $Y | X \sim \mathcal{N}(X\beta, 1)$ for sparse β (see the caption for details). We use cross-validated lasso and ridge coefficient differences as feature statistics, and once again, we find that both MVR and ME knockoffs tend to substantially outperform SDP knockoffs.

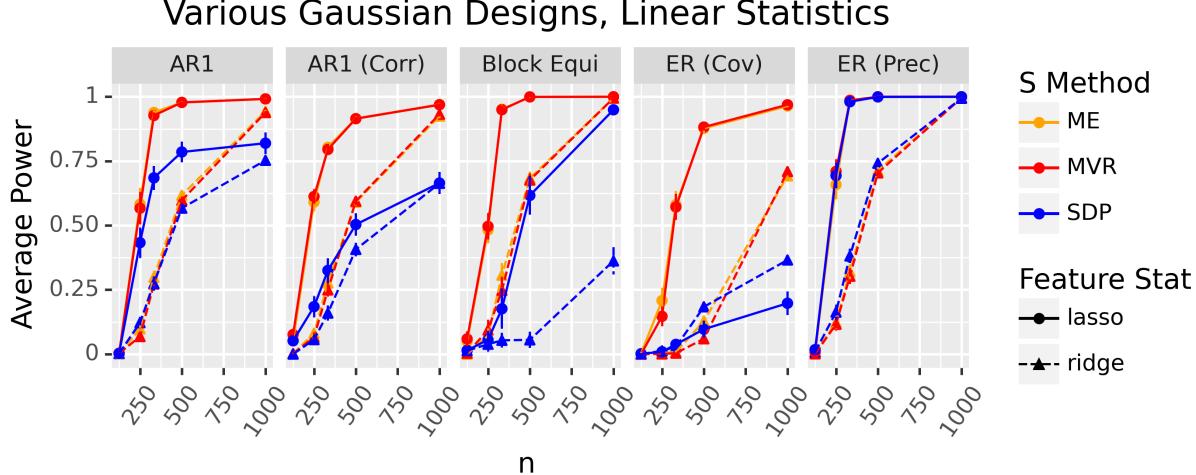


Figure 4: Gaussian designs with a linear response. We set $X \sim \mathcal{N}(0, \Sigma)$ with Σ as defined in Appendix F.4. and let $Y | X \sim \mathcal{N}(X\beta, 1)$. In all cases $p = 500$ and there are 50 non-null with coefficients sampled independently from $\text{Unif}([-δ, -δ/2] \cup [δ/2, δ])$, with $δ = 2$ for the AR1 plots and $δ = 1$ for the others.

In Appendix F.5, we also explore the effect of varying the between-feature correlations in the AR1 setting. As expected, higher correlations improve the performance of MRC knockoffs relative to SDP knockoffs, although interestingly, when the correlation $\text{Cov}(X_j, X_{j+1})$ is constant over all j , the performances of all methods are quite similar. This result is broadly consistent with our theory, although we defer discussion to Appendix F.5.

Next, we consider the case where $Y | X \sim \mathcal{N}(\mu(X), 1)$ where μ is a nonlinear sparse model. We defer precise descriptions of these μ to Appendix F.4, but we emphasize that they are highly nonlinear and $Y | X$ never

follows a single-index model. For example, in the “pairint” setting, $\mu(X) = \sum_{j,k \in [p]} \beta_{j,k} X_j X_k$ for a sparse β . Similarly, in the “cos” setting, $\mu(X) = \cos(X)\beta$, and since cos is an even function, this ensures the features have no linear effect on Y . For this reason, linear feature statistics like the lasso frequently have zero power even when n is as large as $15p$, as demonstrated in Appendix F.6.

Since the conditional distribution $Y | X$ is very complicated, knockoffs will not have much power except in fairly low-dimensional settings. For example, when Y is a linear response to pairwise interactions among the features, even a parametric feature statistic that searches explicitly for pairwise interactions must estimate $O(p^2)$ coefficients. In contrast, a single-index model only requires estimation of $O(p)$ parameters. As a result, we set $p = 200$ with 30 non-null values and vary n between 200 and 3000.

We consider three feature statistics: a random forest feature statistic using the swap importance suggested by Gimenez et al. (2019), the DeepPINK feature statistic from Lu et al. (2018), and, as a baseline, the lasso coefficient difference. We report results for the random forest statistic in Figure 5. Results for the DeepPINK feature statistic and the lasso feature statistic are in Appendix F.6. In all three cases, MVR and ME knockoffs consistently outperform SDP knockoffs (although there are again a few exceptions to this, such as the truncated linear conditional mean on equicorrelated designs).

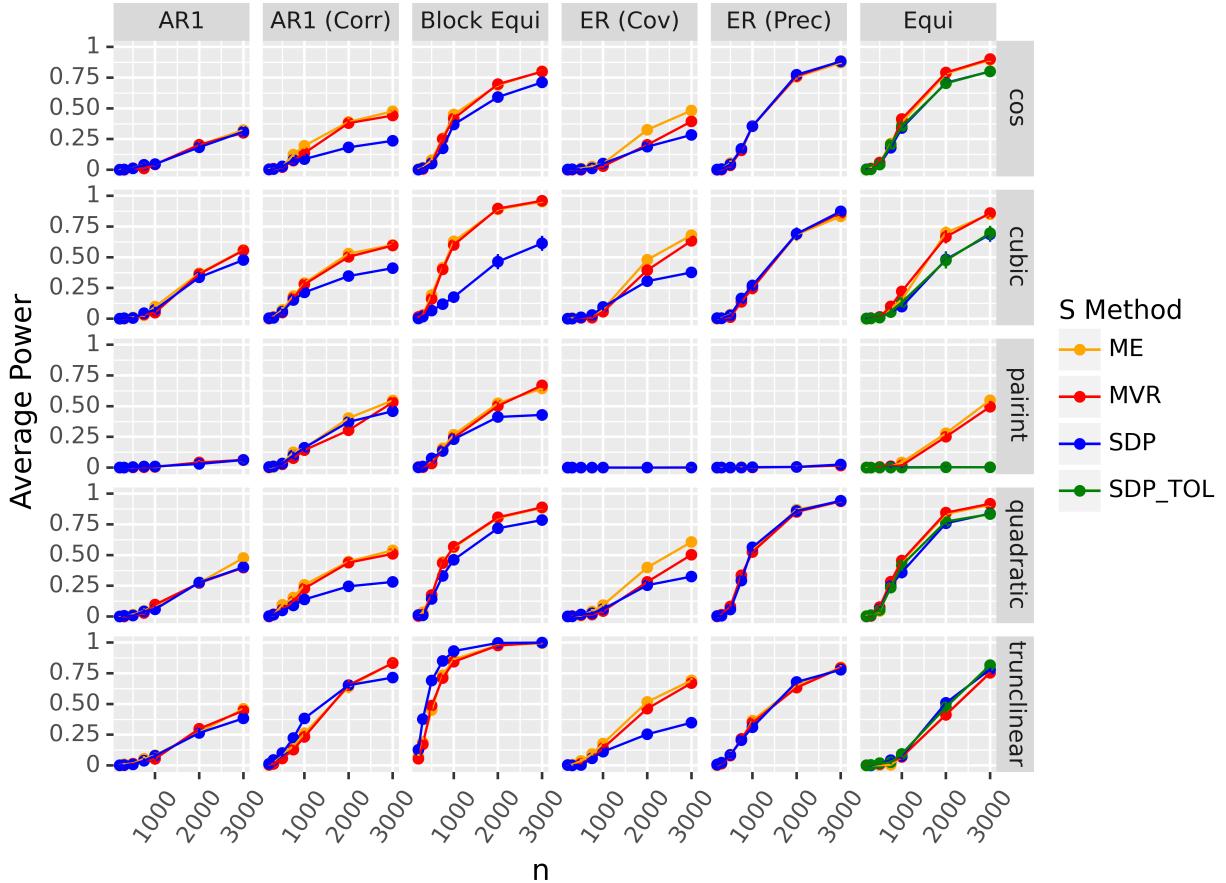


Figure 5: Random forest statistic with swap importances on nonlinear responses: we let $X \sim \mathcal{N}(0, \Sigma)$ for various Σ and $Y | X \sim \mathcal{N}(\mu(X), 1)$. We vary Σ on the x-facets and μ on the y-facets. The precise definitions of the covariance matrices and conditional responses are presented in Appendix F.4. We let $p = 200$ with 30 non-nulls.

4.3 Robustness on Gaussian designs

In Sections 2 through 4.2, we have assumed that we know the true covariance matrix Σ . In this section, we analyze the robustness of MVR, ME, and SDP knockoffs when Σ is not known and is estimated using the same data used to run knockoffs, using one of three methods: Ledoit–Wolf estimation (Ledoit and Wolf, 2004), the graphical lasso algorithm (Friedman et al., 2007), and the maximum likelihood estimate of Σ when $n > p$. We let $Y | X \sim \mathcal{N}(X\beta, 1)$, and we use lasso coefficient difference statistics. Figures 6 (power) and 7 (FDR) show that MVR and ME knockoffs control the FDR better than SDP knockoffs, especially in high-dimensional settings where the covariance has been estimated using shrinkage methods. None of our theory predicted this, but it at least indicates that the power improvement of MVR/ME knockoffs over SDP knockoffs does not come at the expense of robustness. See the figure caption for more simulation details.

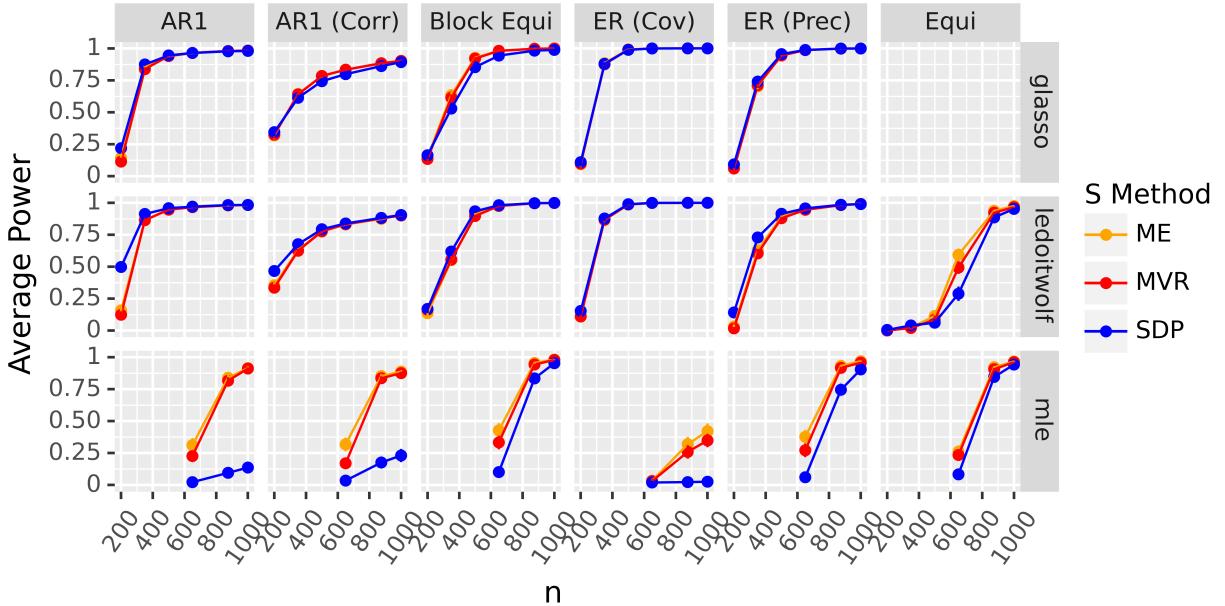


Figure 6: Power of LCD statistics for estimated Σ : we let $X \sim \mathcal{N}(0, \Sigma)$ for various Σ and $Y | X \sim \mathcal{N}(X\beta, 1)$. The definitions of Σ follow Section 4.2 and are formally defined in Appendix F.4. We let $p = 500$ with 50 non-nulls with coefficients sampled independently from $\text{Unif}([\delta, -\delta/2] \cup [\delta/2, \delta])$, where $\delta = 1$ in “ER (Cov)” and “ER (Prec)” panels and $\delta = 0.5$ otherwise. Note that we do not apply the graphical lasso in the case where Σ is equicorrelated, since in this case, the precision matrix is fully dense and the graphical lasso algorithm failed to converge.

4.4 Application to fixed-X knockoffs

So far, we have worked with the MX knockoffs framework, where we assume we know the distribution of X and we control the FDR in expectation over the distribution of X . However, the idea behind MVR and ME knockoffs extends straightforwardly to the fixed-X (FX) knockoff filter, which treats the features X as fixed and controls the FDR when $Y | X$ follows a Gaussian linear model. In the fixed-X setting, we assume the Gram matrix $\Sigma = \mathbf{X}^\top \mathbf{X}$ has diagonals equal to 1 and construct knockoffs $\tilde{\mathbf{X}}$ such that

$$[\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}] = G_S \equiv \begin{bmatrix} \Sigma & \Sigma - S \\ \Sigma - S & \Sigma \end{bmatrix}, \quad (14)$$

where as previously, S is a diagonal matrix such that $S \succcurlyeq 0$ and $2\Sigma - S \succcurlyeq 0$. $\tilde{\mathbf{X}}$ can be efficiently constructed when $n \geq 2p$ as outlined in Barber and Candès (2015). From here on, the FX-knockoffs procedure is quite similar to the MX-knockoffs procedure, except that the feature statistics must obey a *sufficiency* constraint that they only depend on the matrix $[\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}]$ and the empirical covariances $[\mathbf{X}, \tilde{\mathbf{X}}]^\top \mathbf{y}$. Then, when $Y | X \sim \mathcal{N}(X\beta, 1)$, the FX knockoff filter provably controls the FDR in finite samples.

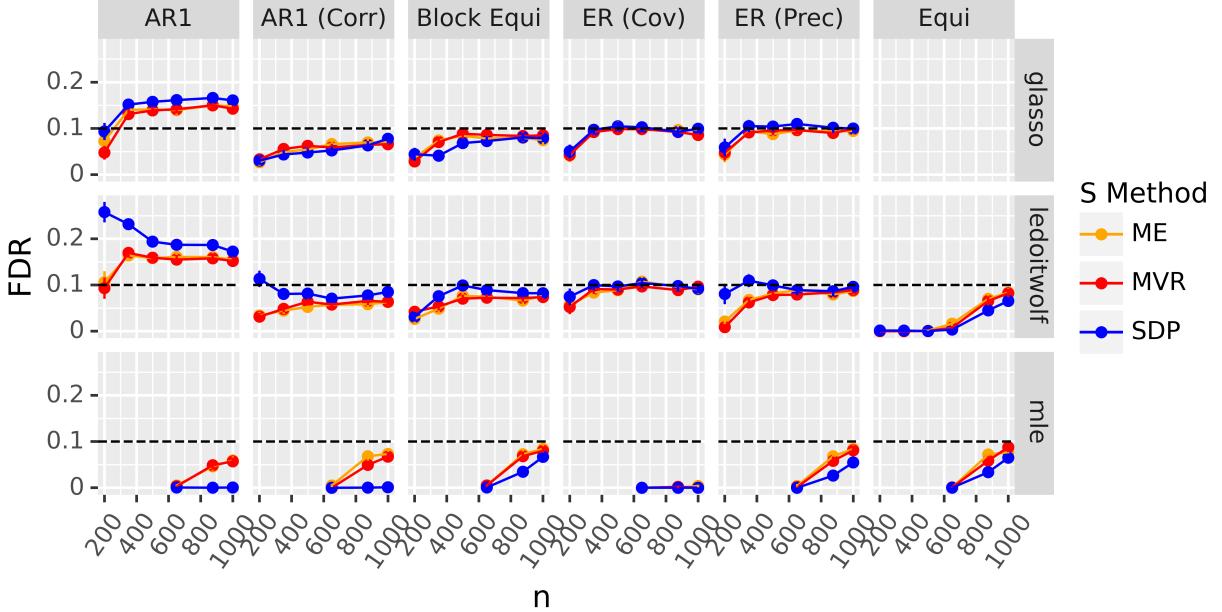


Figure 7: The corresponding FDR plot for Figure 6.

Although the features do not have conditional variances or entropy in this setting since they are treated as fixed, the MVR and ME losses in terms of G_S naturally generalize to this setting. In particular, we set

$$S_{\text{MVR}} = \arg \min_S \text{Tr}(G_S^{-1}) \quad \text{and} \quad S_{\text{ME}} = \arg \min_S \log \det(G_S^{-1}). \quad (15)$$

Proposition 3.1 naturally extends to the FX setting, where the S_{MVR} matrix minimizes the estimation error of OLS feature importances, except this time conditional on \mathbf{X} . See the proof of Proposition 3.1 for details.

To study the power of MVR and ME FX knockoffs, we simulate $X \sim \mathcal{N}(0, \Sigma)$ for the same covariance matrices as in Section 4.2 and let $Y | X \sim \mathcal{N}(X\beta, 1)$ for sparse β (see the caption of Figure 8 for details). In order to obey the sufficiency property, FX feature statistics cannot use cross-validation, so we use the lasso signed max (LSM) statistic introduced in Barber and Candès (2015) instead of cross-validated lasso and ridge coefficient differences. Figure 8 shows that MRC knockoffs generally outperform SDP knockoffs, although the power differential can be lower than in the MX case.

4.5 Application to group knockoffs

In this section, we show that the MRC framework can increase the power of group knockoffs Dai and Barber (2016). In settings with highly correlated features, such as GWAS, some previous work has suggested partitioning the features into disjoint groups $G_1, \dots, G_m \subset [p]$. After clustering the features such that the between-group correlations are suitably low, Dai and Barber (2016) showed how to construct “group knockoffs” which test the group-level hypotheses $H_{G_j} : Y \perp\!\!\!\perp X_{G_j} | X_{-G_j}$. Such group knockoffs need only satisfy a relaxed pairwise-exchangeability constraint, where the distribution of $[X, \tilde{X}]$ is invariant to swaps of entire groups, but not necessarily swaps of individual features. Although grouping the features can reduce correlations between the groups, it generally will not remove all dependence from the data (see Candès et al. (2018); Dai and Barber (2016)), motivating the construction of MRC group knockoffs. One can construct MRC group knockoffs by minimizing the MRC loss functions L_{MVR} and L_{ME} subject to this relaxed constraint.

Figure 9 compares the powers of MRC and SDP group knockoffs in the AR1 settings studied in Section 4.2. To partition the features, we hierarchically cluster the features using correlations as a similarity measure and

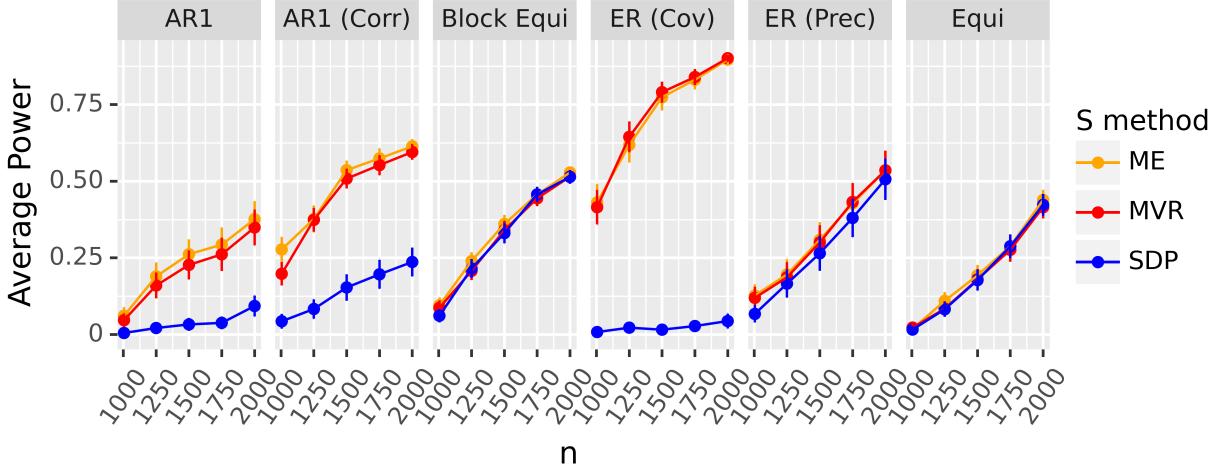


Figure 8: Power for FX Knockoffs for Gaussian designs with lasso signed max statistics. We set $X \sim \mathcal{N}(0, \Sigma)$ with Σ as defined in Appendix F.4 and let $Y | X \sim \mathcal{N}(X\beta, 1)$. In all cases $p = 500$ and there are 50 non-nulls with coefficients sampled independently from $\text{Unif}(-\delta, -\delta/2] \cup [\delta/2, \delta])$. For the AR1 plots, we let $\delta = 0.45$. For the Equi and Block-Equi plots, we let $\delta = 0.15$. For the ER (Prec) and ER (Cov) panels, we let $\delta = 0.25$ and $\delta = 0.5$, respectively.

a single-linkage cutoff of c_{corr} , where we vary c_{corr} between 1.0, which recovers the ungrouped procedure, and 0.7. The results show that MRC group knockoffs consistently outperform SDP group knockoffs.

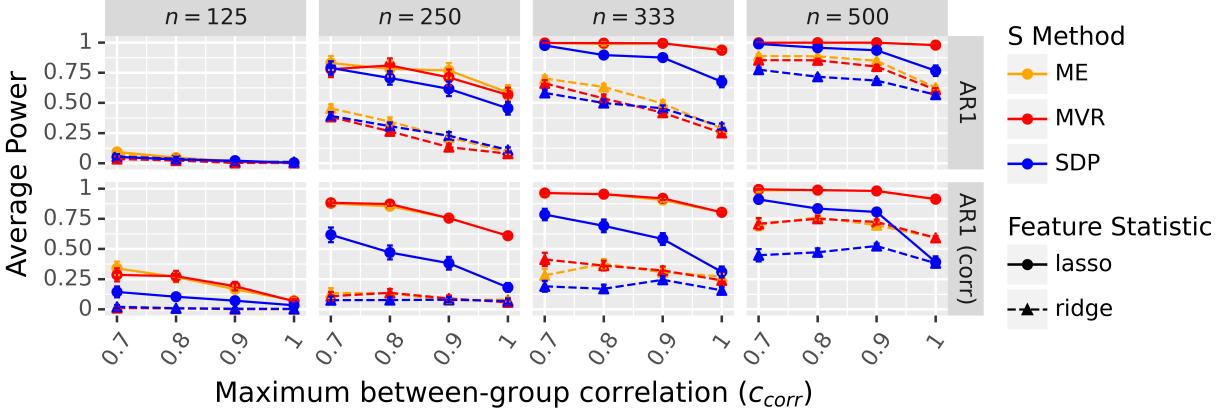


Figure 9: Power for Group MX Knockoffs for Gaussian AR1 designs. We set $X \sim \mathcal{N}(0, \Sigma)$ with Σ as defined in Appendix F.4 and let $Y | X \sim \mathcal{N}(X\beta, 1)$. In all cases $p = 500$ and there are 50 non-nulls with coefficients sampled independently from $\text{Unif}(-\delta, -\delta/2] \cup [\delta/2, \delta])$ with $\delta = 2$. To create groups, we hierarchically cluster the features using a single-linkage cutoff of c_{corr} .

4.6 Application to non-Gaussian designs

So far, we have focused on the case where $X \sim \mathcal{N}(0, \Sigma)$. Now, we consider the non-Gaussian case. To do this, we first review two general methods of constructing knockoffs for non-Gaussian features.

First, Candès et al. (2018) proposed an approximate second-order knockoff construction for the non-Gaussian case. In particular, if X has mean zero and $\text{Cov}(X) = \Sigma$, the authors considered picking an S -matrix and sampling

$$\tilde{X} | X \sim \mathcal{N}(X - X\Sigma^{-1}S, 2S - S\Sigma^{-1}S). \quad (16)$$

This guarantees that the first two moments of X and \tilde{X} match, and moreover that $\text{Cov}([X, \tilde{X}]) = G_S$ as in the Gaussian case. These are *not* valid knockoffs and do not guarantee FDR control, but they may be fairly robust in practice. Furthermore, second-order knockoffs only require knowledge of the first two moments of \tilde{X} , as opposed to the joint density function. Candès et al. (2018) suggested setting $S = S_{\text{SDP}}$ in equation (16). We will demonstrate below that S_{MVR} and S_{ME} are more powerful alternatives.

Second, the Metropolized knockoff sampler introduced in Bates et al. (2020) allows one to sample exact, valid knockoffs for arbitrary distributions of X under the assumption that the unnormalized density Φ of X is known. Given an ordering of the features X_1, \dots, X_p , the key idea is to sample \tilde{X}_j by taking a step along a time-reversible Markov chain starting from X_j , such that $\mathcal{L}(X_j | X_{-j}, \tilde{X}_{1:(j-1)})$ is the stationary distribution of the chain. To accomplish this, Bates et al. (2020) employ a Metropolis–Hastings style proposal and acceptance scheme to ensure the exact validity of the knockoffs. The authors suggested using Gaussian *covariance-guided proposals*, where the Metropolis–Hastings proposals are sampled as second-order knockoffs according to equation (16). In the following simulations, we will generate second-order proposals X^* using the MVR, ME, and SDP S -matrices and then compare their power.

We consider three types of non-Gaussian designs: a t -tailed Markov chain discussed in Bates et al. (2020) with and without correlated signals, a “block-equicorrelated” design where each block is independently t -distributed with an equicorrelated covariance matrix, and a Gibbs measure on a $d \times d$ grid. In all cases, the features differ substantially from the Gaussian case: for example, in the first two settings, we let our t distributions have $\nu = 3$ degrees of freedom. See Appendix F.7 for more details on the precise design distributions as well as our knockoff generation mechanism for the “discrete grid” model.

In Figure 10, we let $Y | X \sim \mathcal{N}(X\beta, 1)$ with similar β as before (see the captions for details). We test three types of feature statistics: lasso coefficient differences, ridge coefficient differences, and debiased lasso (dlasso) coefficient differences (Javanmard and Montanari, 2014).

We make two observations about these plots. First, the only difference between the second-order and Metropolized knockoff procedures is that the Metropolized sampling procedure takes the second-order knockoffs as proposals X^* and sometimes rejects these knockoffs, setting $\tilde{X}_j = X_j$ in the case of a rejection. These rejections *increase* the marginal correlations between X_j and \tilde{X}_j , but outside of the Gibbs grid model, the SDP-guided Metropolized knockoffs have more power than their second-order counterparts. We interpret this as evidence that the reconstruction effect can still occur for non-Gaussian designs, and the Metropolis rejections (inadvertently) correct for this effect by increasing the marginal feature-knockoff correlations. Second, we observe that the MRC second-order knockoffs are substantially more powerful than their SDP counterparts. For the Metropolized knockoff sampler, the MVR and ME proposals increase the power of debiased lasso statistics while decreasing the power of the ridge statistics. The lasso power is fairly similar between all three methods throughout. We present the corresponding FDR plot in Appendix F.7.

Lastly, a few other works have introduced methods to sample approximate knockoffs when the distribution of X is unknown, using tools from the machine learning literature Romano et al. (2018); Jordon et al. (2019). These constructions generally use pairwise dependency measures such as the MAC as part of their optimization criteria, but we expect it to be straightforward to “plug in” the MRC objective criteria in place of the MAC. For example, one could repeatedly resample $X_j | X_{-j}, \tilde{X}$ to estimate $\text{Var}(X_j | X_{-j}, \tilde{X})$, although we leave such details to future work.

5 Discussion

This paper identifies an important flaw in previous knockoff generation mechanisms, which reduces power by allowing regression test statistics like the lasso to reconstruct non-null features using the other features and knockoffs. To solve this problem, we introduced minimum reconstructability knockoffs, which substantially increase the power of knockoffs for correlated designs. However, our work leaves several questions open for future research.

One immediate question is how to construct exact MRC knockoffs for non-Gaussian designs. Our technique in Section 4.6 minimizes the reconstructability of the *proposal* knockoffs X^* in the Metropolized knockoff

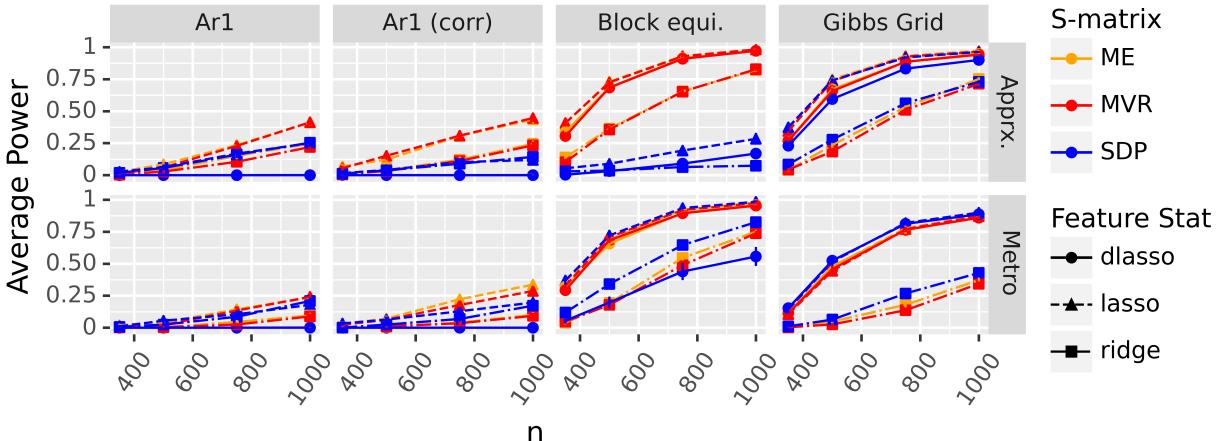


Figure 10: Power for non-Gaussian designs with a linear response. The “Apprx” panel refers to second-order knockoffs. The horizontal facets correspond to the designs described in Appendix F.7. We let $p = 500$ except for the Gibbs Grid, where $p = 625$. We let $Y | X \sim \mathcal{N}(X\beta, 1)$ with sparsity 10% and non-null coefficients drawn from $\text{Unif}([-\delta, -\delta/2] \cup [\delta/2, \delta])$ where $\delta = 0.3$ for the AR1 designs and $\delta = 0.4$ otherwise.

sampling framework, but the actual knockoffs \tilde{X} may have different properties due to the complex acceptance scheme of the sampler. Additionally, Huang and Janson (2020) introduced the idea of *conditional knockoffs*, which enable FDR control when the distribution of X is only known up to a parametric model. Constructing conditional MRC knockoffs may be practically useful, but it is not obvious even how to define conditional MRC knockoffs, with the exception of the Gaussian case, where conditional MRC knockoffs can be defined analogously to fixed- X MRC knockoffs (see Section 4.4).

Interestingly, for discrete designs with finite support, computing the distribution of ME knockoffs corresponds to a maximum entropy problem with linear constraints (see Appendix E for details). Such problems are convex and well studied (Persson and Clarke, 1986; Boyd and Vandenberghe, 2004), but the number of optimization variables and constraints grow exponentially with p , making the problem intractable. We discuss three potential ways around this in Appendix E, including using conditional independence properties of X , restricting the class of feasible knockoff distributions, or finding approximate solutions, but further study is needed.

Another interesting future direction would be to better understand the notion of reconstructability. For example, we observed in Section 2 that the reconstruction effect gets worse when feature statistics can reconstruct non-null features using sparse subsets of the other features and knockoffs. This can occur when G_S is particularly low rank, but it can also occur when the *eigenvectors* of G_S corresponding to small eigenvalues are (approximately) sparse. It may therefore be fruitful to incorporate the structure of the eigenvectors of G_S into knockoff generation mechanisms. Alternatively, we have advocated minimizing two specific measures of reconstructability, but we have not thoroughly investigated other possibilities. Further analysis on this front may turn out to further improve power.

Acknowledgements

The authors would like to thank Chenguang Dai, Buyu Lin, Jun Liu, Wenshuo Wang, and Xin Xing for valuable discussions and suggestions. The authors are also grateful to the anonymous referees for helpful comments. L. J. was partially supported by the William F. Milton Fund.

References

- Anderson, T. (2009). *An Introduction to Multivariate Statistical Analysis*. Wiley india pvt. limited. edition.
- Arratia, R. and Gordon, L. (1989). Tutorial on large deviations for the binomial distribution. *Bulletin of Mathematical Biology*, 51(1):125–131.
- Askari, A., Rebjock, Q., d’Aspremont, A., and Ghaoui, L. E. (2020). Fanok: Knockoffs in linear time. *arXiv preprint arXiv:2006.08790*.
- Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085.
- Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *Ann. Statist.*, 47(5):2504–2537.
- Barber, R. F., Candès, E. J., and Samworth, R. J. (2020). Robust inference with knockoffs. *Ann. Statist.*, 48(3):1409–1431.
- Bates, S., Candès, E., Janson, L., and Wang, W. (2020). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 0(0):1–15.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, USA.
- 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 (Statistical Methodology)*, 80(3):551–577.
- Chen, J., Hou, A., and Hou, T. Y. (2019). A prototype knockoff filter for group selection with FDR control. *Information and Inference: A Journal of the IMA*, 9(2):271–288.
- Dai, C., Lin, B., Xing, X., and Liu, J. S. (2020). False discovery rate control via data splitting. *arXiv preprint arXiv:2002.08542*.
- Dai, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. volume 48 of *Proceedings of Machine Learning Research*, pages 1851–1859, New York, New York, USA. PMLR.
- Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*.
- Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2020). Rank: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115(529):362–379. PMID: 32742045.
- Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
- Gimenez, J. R., Ghorbani, A., and Zou, J. Y. (2019). Knockoffs for the mass: New feature importance statistics with false discovery guarantees. In Chaudhuri, K. and Sugiyama, M., editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 2125–2133. PMLR.
- Gimenez, J. R. and Zou, J. (2019). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89, pages 2184–2192. PMLR.
- Huang, D. and Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *Ann. Statist.*, 48(5):3021–3042.
- Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31(1):253–258.

- Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 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*.
- Ke, Z. T., Liu, J. S., and Ma, Y. (2020). Power of fdr control methods: The impact of ranking algorithm, tampered design, and symmetric statistic. *arXiv preprint: arXiv:2010.08132*.
- Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88(2):365–411.
- Li, J. and Maathuis, M. H. (2019). Ggm knockoff filter: False discovery rate control for gaussian graphical models.
- Liu, J. and Rigollet, P. (2019). Power analysis of knockoff filters for correlated designs. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 15446–15455. Curran Associates, Inc.
- Lu, Y. Y., Fan, Y., Lv, J., and Noble, W. S. (2018). Deeppink: reproducible feature selection in deep neural networks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8690–8700.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020). Deep double descent: Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Oikonomou, K. N. and Grünwald, P. D. (2016). Explicit bounds for entropy concentration under linear constraints. *IEEE Transactions on Information Theory*, 62(3):1206–1230.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Persson, H. and Clarke, R. M. (1986). Algorithm 12: Solving the entropy maximization problem with equality and inequality constraints. *Environment and Planning A: Economy and Space*, 18(12):1665–1676.
- Pipeleers, G. and Vandenberghe, L. (2011). Generalized KYP lemma with real data. *IEEE Trans. Autom. Control*, 56(12):2942–2946.
- Romano, Y., Sesia, M., and Candès, E. J. (2018). Deep knockoffs.
- Sesia, M., Bates, S., Candès, E., Marchini, J., and Sabatti, C. (2020). Controlling the false discovery rate in gwas with population structure. *bioRxiv*.
- Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2019). Multi-resolution localization of causal variants across the genome. *bioRxiv*.
- Sesia, M., Sabatti, C., and Candès, E. J. (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18.
- Vandenberghe, L., Boyd, S., and Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.*, 19(2):499–533.
- Xing, X., Zhao, Z., and Liu, J. S. (2019). Controlling false discovery rate using gaussian mirrors.

A Proofs for Section 2

A.1 Proofs for the general reconstruction effect

In this section, we prove some general facts about the reconstruction effect, which we will apply to the equicorrelated case in the following section. To start, we prove Lemma 2.1, which shows that when X is Gaussian, minimizing the MAC will often cause G_{SDP} to be low rank.

Lemma 2.1. *Suppose $X \sim \mathcal{N}(0, \Sigma)$ and $\lambda_{\min}(\Sigma) \leq 0.5$. Then $\text{rank}(G_{\text{SDP}}) < 2p$. Furthermore, if Σ is block-diagonal with b blocks, each with an eigenvalue below 0.5, then $\text{rank}(G_{\text{SDP}}) \leq 2p - b$.*

Proof. The eigenvalues of G_{SDP} are the eigenvalues of S_{SDP} and $2\Sigma - S_{\text{SDP}}$. The case where $\lambda_{\min}(\Sigma) = 0.5$ is trivial, as we can set $S_{\text{SDP}} = I_p$, which implies $2\Sigma - S_{\text{SDP}}$ will have at least one eigenvalue that equals 0.

When $\lambda_{\min}(\Sigma) < 0.5$, assume for sake of contradiction that G_{SDP} is full rank, and thus $\lambda_{\min}(2\Sigma - S_{\text{SDP}}) = \gamma > 0$. This implies $2\Sigma - S_{\text{SDP}} - \gamma I_p \succcurlyeq 0$ as well. Represent $S_{\text{SDP}} = \text{diag}(s)$ and denote $s^* = \min(s + \gamma, 1)$, where the minimum is taken element-wise over the vector.

By the previous argument, $\text{diag}(s^*)$ is a feasible S -matrix with lower mean absolute correlation than S_{SDP} . Therefore, S_{SDP} cannot be the solution to the SDP. This is a contradiction and completes the proof of the non-block-diagonal statement. The block-diagonal statement follows because when Σ is block-diagonal, the solution to the SDP is the SDP solution to each of the blocks (Barber and Candès, 2015). \square

Next we prove Theorem 2.3, which we very slightly restate to make it easier to apply in our later proofs. We use \odot to denote elementwise multiplication.

Theorem 2.3. *Suppose we can represent $Y = f(g(X_J), X_{-J}, U)$ for some set $J \subset [p]$, functions f and g , and independent noise $U \sim \text{Unif}(0, 1)$. Equivalently, this means $Y \perp\!\!\!\perp X_J \mid g(X_J), X_{-J}$. Suppose a function g^* exists such that $g(X_J) = g^*(\tilde{X}_J)$ holds almost surely. If $Y^* = f(g^*(X_J), X_{-J}, U)$, then*

$$([X, \tilde{X}], Y) \stackrel{d}{=} ([X, \tilde{X}]_{\text{swap}(J)}, Y^*) \quad \text{and} \quad ([X, \tilde{X}], Y^*) \stackrel{d}{=} ([X, \tilde{X}]_{\text{swap}(J)}, Y).$$

In particular, this implies

$$w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \stackrel{d}{=} -\mathbf{1}_J \odot (w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}^*)), \quad (17)$$

which ensures $\mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})_j > 0) + \mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}^*)_j > 0) \leq 1$.

Proof. Note that since $[X_J, \tilde{X}_J] \stackrel{d}{=} [\tilde{X}_J, X_J]$, $g(X_J) = g^*(\tilde{X}_J)$ also implies $g(\tilde{X}_J) = g^*(X_J)$. Thus, we can rewrite

$$Y^* = f(g^*(X_J), X_{-J}, U) = f(g(\tilde{X}_J), X_{-J}, U).$$

This means that $Y^* \mid [X, \tilde{X}]_{\text{swap}(J)} \stackrel{d}{=} Y \mid [X, \tilde{X}]$. Since marginally $[X, \tilde{X}]_{\text{swap}(J)} \stackrel{d}{=} [X, \tilde{X}]$, this implies

$$([X, \tilde{X}], Y) \stackrel{d}{=} ([X, \tilde{X}]_{\text{swap}(J)}, Y^*).$$

The former equality holds for each i.i.d. row (X, Y) and therefore holds for the matrix versions $\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}$. Since w is a function of $[\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}$ and must obey the knockoff antisymmetry property, we obtain that

$$w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \stackrel{d}{=} w([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(J)}, \mathbf{y}^*) = -\mathbf{1}_J \odot w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}^*).$$

For each $j \in J$, this proves $\mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})_j > 0) + \mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}^*)_j > 0) \leq 1$. \square

A.2 Simple proofs for block-equicorrelated Gaussian designs

In this section, we will prove Lemma 2.2 and apply Theorem 2.3 to the equicorrelated case. These will be important tools for our later proof of Theorem 2.4.

Lemma 2.2. *In the equicorrelated case when $\rho \geq 0.5$, let \tilde{X} be generated according to the SDP procedure. Then G_{SDP} has rank $p + 1$, and $X_j + \tilde{X}_j = X_k + \tilde{X}_k$ for all $1 \leq j, k \leq p$.*

Proof. To find the solution for the SDP, we simply verify that $S_{\text{SDP}} = (2 - 2\rho) \cdot I_p$ by checking the KKT conditions in Lemma B.1. Then, we have

$$[X, \tilde{X}] \sim \mathcal{N}(0, G_{\text{SDP}}) \text{ for } G_{\text{SDP}} = \begin{bmatrix} \Sigma & \Sigma - (2 - 2\rho)I_p \\ \Sigma - (2 - 2\rho)I_p & \Sigma \end{bmatrix} \in \mathbb{R}^{2p \times 2p}.$$

First, we note that the eigenvalues of G_{SDP} are those of $(2 - 2\rho)I_p$ and $2\Sigma - (2 - 2\rho)I_p$. Note, however, that $\Sigma = (1 - \rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p^\top$, so $2\Sigma - (2 - 2\rho)I_p = 2\rho\mathbf{1}_p\mathbf{1}_p^\top$, which has a rank of 1. Since $(2 - 2\rho)I_p$ has full rank, this implies that G_{SDP} has rank $p + 1$.

To show that $X_j + \tilde{X}_j$ remains constant over all $1 \leq j \leq p$, denote the j th column of G_{SDP} as G_j . Simple arithmetic shows that $G_j + G_{j+p} = 2\rho \cdot \mathbf{1}_{2p}$. This implies that if $\mu \in \mathbb{R}^{2p}$ is the vector of all zeros except $\mu_1, \mu_{p+1} = 1$ and $\mu_j, \mu_{j+p} = -1$, then $\mu^\top G_{\text{SDP}} = 0 \in \mathbb{R}^{2p}$. If $\sqrt{G_{\text{SDP}}}$ is a symmetric square root of G_{SDP} , this implies $\mu^\top \sqrt{G_{\text{SDP}}} = 0$.

Now, let $\epsilon \sim \mathcal{N}(0, I_{2p})$. We may represent $[X, \tilde{X}] \sim \sqrt{G_{\text{SDP}}} \epsilon$. Using this representation, we conclude

$$X_1 + \tilde{X}_1 - X_j - \tilde{X}_j = \mu^\top [X, \tilde{X}] \sim \mu^\top \sqrt{G_{\text{SDP}}} \epsilon = 0.$$

□

Corollary A.1. *Suppose $\Sigma = \text{blockdiag}(\Sigma_1, \dots, \Sigma_\ell)$ and Σ_k is equicorrelated with correlation at least 0.5, for $1 \leq k \leq \ell$. Let D_k be the set of indices corresponding to block k . Then for all $j_1, j_2 \in D_k$, $X_{j_1} + \tilde{X}_{j_1} = X_{j_2} + \tilde{X}_{j_2}$ if \tilde{X} are SDP knockoffs.*

Proof. This follows directly from Lemma 2.2 since the solution to the SDP for a block-diagonal matrix is simply the diagonal matrix composed of the solutions to the blocks $\Sigma_1, \dots, \Sigma_\ell$. □

Corollary A.1 allows us to apply Theorem 2.3 to the block-equicorrelated case. We define some notation before doing this.

Definition A.1 (Negation Notation). *For coefficients $\beta \in \mathbb{R}^p$ and $J \subset [p]$, define $\beta_{\text{neg}(J)} = -\mathbf{1}_J \odot \beta$. For the response variable $Y = f(X, U)$, define $Y_{\text{neg}(J)} = f(-\mathbf{1}_J \odot X, U)$, where \odot denotes elementwise multiplication.*

Note that in the follow Corollary, $Y_{\text{neg}(J)}$ corresponds to Y^* in the Theorem 2.3.

Corollary A.2. *Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{blockdiag}(\Sigma_1, \dots, \Sigma_\ell)$, where each Σ_k is equicorrelated with correlation $\rho_k \geq 0.5$. Let $Y = f(X\beta, U)$ and suppose we generate knockoffs \tilde{X} according to the SDP procedure. Let $J \subset [p]$ such that J is contained among a single equicorrelated block and $\sum_{j \in J} \beta_j = 0$. Then*

$$w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \stackrel{d}{=} -\mathbf{1}_J \odot (w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}_{\text{neg}(J)})),$$

which implies $\mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})_j > 0) + \mathbb{P}(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}_{\text{neg}(J)})_j > 0) \leq 1$.

Proof. Without loss of generality assume $J = \{1, \dots, |J|\}$. Corollary A.1 tells us that for $j \in J$, $X_j = X_1 + \tilde{X}_1 - \tilde{X}_j$. This implies that $X_J \beta_J = \sum_{j \in J} \beta_j (X_1 + \tilde{X}_1) - \tilde{X}_J \beta_J = -\tilde{X}_J \beta_J$, where the last equality follows because $\sum_{j \in J} \beta_j = 0$. This satisfies the assumptions of Theorem 2.3, since $Y \perp\!\!\!\perp X_J \mid X_J \beta_J, X_{-J}$ and we can reconstruct $X_J \beta_J$ using \tilde{X}_J . Thus, if we set $Y_{\text{neg}(J)} = f(-\mathbf{1}_J \odot X\beta, U) = f(X\beta_{\text{neg}(J)}, U)$, Theorem 2.3 yields the result. □

A.3 Proof of Theorem 2.4

The proof for Theorem 2.4 is presented below. Our first task is to strengthen the result of Corollary A.2 by applying the permutation invariance assumption and considering the case where $\sum_{j \in J} \beta_j \approx 0$. Then, we will prove our main technical theorem, Theorem A.2. After that, we will prove Theorem 2.4. Throughout, we will defer overly technical details to Appendix B. In general, the value of universal constants c_0, c_1, C_0, C_1, C_2 may change by a constant from line to line.

The following proposition strengthens Corollary A.2.

Proposition A.1. *Let $X \sim \mathcal{N}(0, \Sigma)$ for block-equicorrelated Σ with within-block correlations at least 0.5. Let $Y = f(X\beta + \zeta_0, U)$ for $\zeta_0 \sim \mathcal{N}(0, \sigma_0^2)$ and $U \sim \text{Unif}(0, 1)$, with U, ζ_0, X jointly independent. Suppose we generate knockoffs \tilde{X} according to the SDP procedure and $W = w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$ is a permutation invariant feature statistic. For any j_1, j_2 in the same equicorrelated block, assume $|\beta_{j_1} + \beta_{j_2}| < d \in \mathbb{R}$. Then if $J = \{j_1, j_2\}$, for a universal constant c_0 depending only on σ_0^2 ,*

$$d_{\text{TV}}([W_{j_2}, W_{j_1}, W_{-J}], [-W_{j_1}, -W_{j_2}, W_{-J}]) \leq c_0 \sqrt{nd}.$$

In particular, this implies

$$d_{\text{TV}}([\text{sign}(\text{sorted}(W_J)), W_{-J}], [-\text{sign}(\text{sorted}(W_J)), W_{-J}]) \leq c_0 \sqrt{nd},$$

where $\text{sorted}(W_J)$ refer to W_J sorted in descending order of absolute value and $\text{sign}(\text{sorted}(W_J))$ are the signs of the sorted W_J .

Proof. To begin with, we analyze the case where $d = 0$ and $\beta_{j_1} = -\beta_{j_2}$. Let $\sigma : [p] \rightarrow [p]$ be the permutation which swaps j_1 with j_2 but leaves all other indices constant. Then we observe $\beta_{\text{neg}(J)} = \sigma(\beta)$, so

$$Y_{\text{neg}(J)} = f(X\sigma(\beta) + \zeta_0, U) = f(\sigma^{-1}(X)\beta + \zeta_0, U).$$

This implies that $Y_{\text{neg}(J)} \mid [X, \tilde{X}] \stackrel{d}{=} Y \mid [\sigma(X), \sigma(\tilde{X})]$ (note that $\sigma = \sigma^{-1}$). Since X and therefore \tilde{X} are exchangeable within a block, this implies

$$(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}_{\text{neg}(J)}) \stackrel{d}{=} (\sigma(\mathbf{X}), \sigma(\tilde{\mathbf{X}}), \mathbf{y})$$

which implies

$$w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}_{\text{neg}(J)}) \stackrel{d}{=} w([\sigma(\mathbf{X}), \sigma(\tilde{\mathbf{X}})], \mathbf{y}) \stackrel{d}{=} \sigma(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})),$$

where the right-hand side follows from the permutation invariance of W . We now apply Corollary A.2 to the term on the left-hand side to obtain

$$-\mathbf{1}_J \odot w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \stackrel{d}{=} w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}_{\text{neg}(J)}) \stackrel{d}{=} \sigma(w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})). \quad (18)$$

Now, we extend this analysis to the case where $|\beta_{j_1} + \beta_{j_2}| = d > 0$ using a total variation argument. Define the vector $\beta' \in \mathbb{R}^p$ where

$$\beta'_k = \begin{cases} \beta_k & k \neq j_1 \\ -\beta_{j_2} & k = j_1 \end{cases}$$

so in particular $\beta'_{j_1} = -\beta'_{j_2}$. Then let $Y' = f(X\beta' + \zeta_0, U)$ and let $\zeta \in \mathbb{R}^n$ denote the concatenation of all ζ_0 . Since $[\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}\beta + \zeta]$ are jointly multivariate Gaussian, it is simple to prove that

$$d_{\text{TV}}([\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}\beta + \zeta], [\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}\beta' + \zeta]) \leq c_0 \sqrt{nd}$$

for some c_0 depending only on σ_0^2 , which we show in Lemma B.6. Since \mathbf{y} is a function of independent noise and $[\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}\beta + \zeta]$, and respectively for \mathbf{y}' , this bound immediately implies

$$d_{\text{TV}}([\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}], [\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}']) \leq c_0 \sqrt{nd}. \quad (19)$$

Abbreviate $w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$ as W and $w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}')$ as W' . Applying the triangle inequality, we obtain that

$$\begin{aligned} d_{\text{TV}}([W_{j_1}, W_{j_2}, W_{-J}], [-W_{j_2}, -W_{j_1}, W_{-J}]) &= d_{\text{TV}}(W, -\mathbf{1}_J \odot \sigma(W)) \\ &\leq d_{\text{TV}}(W, W') + d_{\text{TV}}(W', -\mathbf{1}_J \odot \sigma(W')) + d_{\text{TV}}(-\mathbf{1}_J \odot \sigma(W'), -\mathbf{1}_J \odot \sigma(W)) \\ &\leq c_0 \sqrt{nd} + 0 + c_0 \sqrt{nd} \end{aligned}$$

where the last line follows from applying (18) to the middle term and (19) to the other two terms. The proposition now follows if we reset c_0 to be twice its original value. \square

The following corollary indicates that the conclusion of Proposition A.1 also holds conditionally with high probability. Since it is a straightforward property of joint total variation, we defer its proof to Appendix B.2.

Corollary A.3. *In the same setting as above, let α_1, α_2 be constants such that $\alpha_1 \cdot \alpha_2 > c_0 \sqrt{nd}$ and let A be some event in the sigma-algebra generated by $[\text{sorted}(W_J)], W_{-J}]$. Then if $\mathbb{P}(A) > \alpha_2$, we have that*

$$d_{\text{TV}}([\text{sorted}(W_J), W_{-J}], [-\text{sorted}(W_J), W_{-J}] \mid A) < \alpha_1. \quad (20)$$

Below, we prove our main technical theorem, of which Theorem 2.4 is a corollary.

Theorem A.2. *Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{blockdiag}(\Sigma_1, \dots, \Sigma_\ell)$ where for all $1 \leq j \leq \ell$, Σ_j is equicorrelated with correlation $\rho_j \geq 0.5$. Suppose $Y = f(X\beta + \zeta_0, U)$ for $\zeta_0 \sim \mathcal{N}(0, \sigma_0^2)$ and $U \sim \text{Unif}(0, 1)$ with U, ζ_0, X jointly independent. Let \tilde{X} be generated according to the SDP procedure and let $W = w([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$ be a permutation invariant feature statistic. Suppose we use knockoffs to control the FDR at level $q \leq 0.1$.*

Fix $b, m > 0$. Suppose we can group the features into M disjoint pairs G_1, \dots, G_M such that if $G_j = \{j_1, j_2\}$, then X_{j_1}, X_{j_2} are in the same equicorrelated block and $|\beta_{j_1} + \beta_{j_2}| < \frac{b}{m}$. This leaves $M' = p - 2M$ “singleton” features which are not in any pair.

Let α_1, α_2 be constants such that $\alpha_1 \cdot \alpha_2 > \frac{c_0 \sqrt{nb}}{m}$. Assume that $\alpha_1 < 0.001$ and $M' \leq \frac{p}{50}$. Then there exist universal constants C_1, C_2, C_3 such that if τ is the number of discoveries made by the procedure,

$$\frac{\mathbb{E}_{w, \beta}[\tau]}{p} \leq \frac{C_1 M' + C_2}{p} + C_3 \alpha_2.$$

Proof. To start, relabel the features such that the “singleton” statistics are labelled $W_1, \dots, W_{M'}$. We will make the worst-case assumption that $W_1, \dots, W_{M'} = +\infty$. This assumption is permissible since it only increases the number of discoveries made by the procedure, as formalized in Lemma B.4. Our proof now proceeds in three steps.

Step 1: First, we reformulate the problem in terms of $\eta = \text{sign}(\text{sorted}(W)) \in \{-1, 1\}^p$, since the power of knockoffs depends only on the values of the random vector η . We introduce some notation for this purpose. For $\epsilon \in \{-1, 1\}^p$ and any $k \in \mathbb{N}$, let $V_k^+(\epsilon)$ be the number of positive 1’s at or before the k th element of ϵ . Formally,

$$V_k^+(\epsilon) = \#\{j \leq k : \epsilon_j = 1\}.$$

Note for $k > p$, $V_k^+(\epsilon) = V_p^+(\epsilon)$ since ϵ only has p coordinates. This means that $V_k^+(\epsilon)$ is uniformly bounded by p . We also observe that $V_k^+(\epsilon)$ is nondecreasing in k . Define

$$\psi(\epsilon) = \max_{1 \leq k \leq \infty} \left\{ k : \frac{k - V_k^+(\epsilon) + 1}{V_k^+(\epsilon)} \leq q \right\} = \max_{1 \leq k \leq \infty} \{k : k \leq (1+q)V_k^+(\epsilon) - 1\} \leq \lfloor (1+q)p - 1 \rfloor, \quad (21)$$

where by convention $\psi(\epsilon) = 0$ if the set in the middle is empty. Intuitively, we can think of $\psi(\eta)$ as a reformulation of the data-dependent threshold in (3), where the knockoffs procedure selects all features in $\text{sorted}(W)$ which appear before position $\psi(\eta)$ and have positive signs. This implies that $\tau \leq \psi(\eta)$, as we prove formally in Lemma B.3. Furthermore, note that we let ψ take values greater than p to ensure that the

number of discoveries τ is a deterministic function of $\psi(\eta)$, which will be important in Step 3—see Lemma B.3 for details. By the prior analysis, it suffices to show that

$$\mathbb{E}_{w,\beta}[\psi(\eta)] \leq C_1 M' + C_2 + C_3 \alpha_2 p.$$

We may decompose

$$\psi(\eta) = \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \cdot \mathbb{I}(\psi(\eta) = k), \quad (22)$$

where we begin the sum at $\lfloor(1+q)M'-1\rfloor$ because we assume the first M' coordinates of η correspond to the singleton features and are guaranteed to be 1's, so we must have that $\psi(\eta) \geq \lfloor(1+q)M'-1\rfloor$. If we define

$$\Lambda_k = \{\epsilon \in \{-1, 1\}^p : \psi(\epsilon) = k\} \cap \{\epsilon : \mathbb{P}_{\beta,W}(\eta = \epsilon) \neq 0\}, \quad (23)$$

then

$$\mathbb{E}_{w,\beta}[\psi(\eta)] = \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \cdot \mathbb{P}_{\beta,W}(\eta \in \Lambda_k). \quad (24)$$

Step 2: In this step, we will show that $\mathbb{P}(\eta = \epsilon)$ cannot be much larger than 2^{-M} for most $\epsilon \in \{-1, 1\}^p$. The 2^{-M} comes because for each group $j \in [M]$, the conditional distribution of $\text{sorted}(W_{G_j})$ given $(W_{-G_j}, |\text{sorted}(W_{G_j})|)$ is approximately symmetric with high probability (see Corollary A.3).

The main task of this section will be to introduce notation such that we can write the event $\eta = \epsilon$ in terms of the random variables $\text{sign}(\text{sorted}(W_{G_j}))$. Note for the rest of this proof and Lemma A.3 (stated later), we use i and t to denote indices of $\text{sorted}(W)$ and j to denote indices of the unsorted feature statistics W or the pairs G_j . In other words, i, t mean “post-sorting,” and j means “pre-sorting.”

This step is the most conceptually challenging, so we will divide it into four sub-steps. Our notation will explicitly account for the singleton statistics, but we advise the reader to focus on the non-singleton statistics.

Step 2a: First, we will define a random vector $R = (R_1, \dots, R_p) \in \{0, \dots, M\}^p$. Intuitively, for $1 \leq i \leq p$, R_i takes the (post-sorting) feature statistic $\text{sorted}(W)_i$ and tells us which (pre-sorting) pair G_j it came from. Formally, we define

$$R_i = \begin{cases} j & \text{sorted}(W)_i \text{ came from pair } G_j \\ 0 & \text{sorted}(W)_i \text{ corresponds to a singleton statistic.} \end{cases}$$

As an example, fix $p = 7$ and suppose the pairs are $G_1 = \{1, 2\}$, $G_2 = \{3, 4\}$, $G_3 = \{5, 6\}$, and the 7th feature statistic is a singleton statistic (which we assume equals $+\infty$). Then suppose we observe

$$W = (W_1, W_2, W_3, W_4, W_5, W_6, W_7) = (1.3, 5, 0, -1.5, -2, -2.3, +\infty). \quad (25)$$

By the definition of our sorting function, this implies that

$$\text{sorted}(W) = (+\infty, 5, -2.3, -2, -1.5, 1.3, 0). \quad (26)$$

The first sorted coordinate $\text{sorted}(W)_1 = +\infty$ corresponds to the unsorted feature statistic W_7 , which is a singleton statistic. Thus, $R_1 = 0$. The second sorted coordinate $\text{sorted}(W)_2 = 5$ corresponds to the unsorted feature statistic W_2 which belongs to pair G_1 , so we set $R_2 = 1$. Continuing on, we find that

$$R = (0, 1, 3, 3, 2, 1, 2). \quad (27)$$

In general, the first M' coordinates of R will equal zero, since the M' singleton statistics equal $+\infty$ and make up the first M' coordinates of $\text{sorted}(W)$. However, the important information contained by R is captured by the last $p - M'$ coordinates, which allow us to convert between $\text{sorted}(W)_i$ and the groupings W_{G_j} . Crucially, R does not contain any information about the within-pair rankings of feature statistics, and thus R is in the sigma-algebra generated by $\{\text{sorted}(W_{G_j})\}_{j=1}^M$. This is important, since it means that $\text{sorted}(W_{G_j})$ is

approximately symmetric conditional on R —recall that Corollary A.3 tells us $\text{sorted}(W_{G_j})$ is approximately symmetric when we condition on $|\text{sorted}(W_{G_j})|$ and W_{-G_j} , but not necessarily if we condition on $|W_{G_j}|$.

Step 2b: The random vector R tells us for each i that $\text{sorted}(W)_i$ corresponds to one of the two elements of $\text{sorted}(W_{G_{R_i}})$, since G_{R_i} denotes the pair appearing at position i in $\text{sorted}(W)$, and $W_{G_{R_i}}$ denotes the feature statistics belonging to that pair. However, we do not know *which* of the two elements in $\text{sorted}(W_{G_{R_i}}) \in \mathbb{R}^2$ correspond to $\text{sorted}(W)_i$. To fill this notational gap, we will introduce the random vector $N(R) = (N_1(R), \dots, N_p(R)) \in \{1, 2\}^p$ such that $\text{sorted}(W)_i = \text{sorted}(W_{G_{R_i}})_{N_i(R)}$.

To do this, we note that both $\text{sorted}(W)$ and $\text{sorted}(W_{G_{R_i}})$ sort feature statistics in descending order of absolute value. For this reason, if $G_j = \{j_1, j_2\}$, then W_{j_1} will appear before W_{j_2} in $\text{sorted}(W_{G_j})$ if and only if W_{j_1} appears before W_{j_2} in $\text{sorted}(W)$. Therefore, we can define $N_i(R)$ as follows for $1 \leq i \leq p$:

$$N_i(R) = \begin{cases} 0 & \text{sorted}(W)_i \text{ corresponds to a singleton statistic} \\ 1 & \text{sorted}(W)_i \text{ appears before the other member of its pair in } \text{sorted}(W) \\ 2 & \text{sorted}(W)_i \text{ appears after the other member of its pair in } \text{sorted}(W). \end{cases}$$

This guarantees that $\text{sorted}(W)_i = \text{sorted}(W_{G_{R_i}})_{N_i(R)}$ for each $M' + 1 \leq i \leq p$. More importantly, we have defined $N(R)$ such that it is a deterministic function of R , meaning that W_{G_j} should be approximately symmetric even conditional on $N(R)$.

As a concrete example, recall equations (25), (26), and (27). To find $N_3(R)$, we note that $R_3 = 3$, meaning that $\text{sorted}(W)_3$ comes from G_3 . Since no previous coordinate of R equals 3, we conclude $N_3(R) = 1$. On the other hand, to find $N_4(R)$, we note $R_4 = 3$ as well, meaning $\text{sorted}(W)_4$ also comes from G_3 . Since $\text{sorted}(W)_4$ is not the first coordinate of $\text{sorted}(W)$ to come from pair G_3 , we set $N_4(R) = 2$. Continuing, we get that

$$N(R) = (0, 1, 1, 2, 1, 2, 2). \quad (28)$$

As always, the 0 in $N(R)$ signals the presence of a (nonrandom) singleton statistic.

Step 2c: We are now ready to write the event $\eta = \epsilon$ in terms of $\{\text{sign}(\text{sorted}(W_{G_j}))\}_{j=1}^M$ and R . Let $\epsilon \in \{-1, 1\}^p$ such that $\epsilon_t = 1$ for $1 \leq t \leq M'$. Recall that our notation guarantees that $\text{sorted}(W)_i = \text{sorted}(W_{G_{R_i}})_{N_i(R)}$. Thus, the following events are equivalent:

$$\text{sign}(\text{sorted}(W)) = \epsilon \Leftrightarrow \left(\bigcap_{i=M'+1}^p \text{sign}(\text{sorted}(W_{G_{R_i}}))_{N_i(R)} = \epsilon_i \right).$$

This motivates the following decomposition by the chain-rule of conditional probability. Define probabilities

$$\begin{aligned} p(r) &= \mathbb{P}(R = r), \\ p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) &= \mathbb{P}\left(\text{sign}(\text{sorted}(W_{G_{R_i}}))_{N_i(r)} = \epsilon_i \mid \bigcap_{t=M'+1}^{i-1} \text{sign}(\text{sorted}(W_{G_{R_t}}))_{N_t(r)} = \epsilon_t, R = r\right), \end{aligned} \quad (29)$$

where we also adopt the convention that for all i , $p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) = 0$ unless $\epsilon_t = 1$ for $1 \leq t \leq M'$, since the first M' statistics in $\text{sorted}(W)$ correspond to the (positive) singleton statistics. Note also that in the previous statement, we define $i : t$ to be the empty set when $i > t$. Now, by the law of total probability and the chain-rule of conditional probability,

$$\mathbb{P}(\eta = \epsilon) = \sum_{r: p(r) > 0} p(r) \prod_{i=M'+1}^p p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r), \quad (30)$$

where again, the product begins at $M' + 1$ because we assume the first M' values correspond to the singleton statistics.

Intuitively, we should expect (30) to have a value which does not substantially exceed 2^{-M} . To see this, imagine iterating through $p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r)$ as we increment i from $M' + 1$ to p . For any i , if $\text{sorted}(W)_i$ corresponds to a group we have not seen before—meaning $N_i(r) = 1$ —then we should expect $p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) \approx 1/2$ since $\text{sorted}(W_{G_{r_i}})$ is approximately symmetric. Slightly more formally, this holds because when $N_i(r) = 1$, we are only conditioning on R and the signs of the other groups, both of which are contained in the sigma-algebra generated by $\{|\text{sorted}(W_{G_{r_i}})|, W_{-G_{r_i}}\}$, in accordance with Corollary A.3. On the other hand, when we consider an i such that $N_i(r) = 2$, we can only trivially upper bound $p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r)$ by 1, since we are conditioning on the sign of $\text{sorted}(W_{G_{r_i}})_1$, which violates the assumption of Corollary A.3.

To formalize this idea, we need one final piece of notation, which we define in Step 2d.

Step 2d: The vector R allows us to convert from sorted coordinates i into unsorted groups G_j . The random vector $C(R) \in [p]^M$ will do (approximately) the inverse. We define $C(R)$ elementwise, such that for $j \in [M]$, $C_j(R)$ equals the coordinate of the first element of pair G_j in $\text{sorted}(W)$. In the example from equations (25), (26), (27), we would have that

$$C(R) = (2, 5, 3).$$

For example, $C_1(R) = 2$ because pair $j = 1$ appears for the first time in $\text{sorted}(W)$ at coordinate $i = 2$. In particular, $C_j(R) = i$ guarantees that $\text{sorted}(W)_i$ corresponds to pair G_j and $N_i(R) = 1$. With this in mind, we can state our main Lemma for Step 2.

Lemma A.3. *Let α_1, α_2 be constants such that $\alpha_1 \alpha_2 > \frac{c_0 \sqrt{n} b}{m}$. For any $1 \leq j \leq M$ and for any $\Lambda \subset \{-1, 1\}^p$,*

$$\alpha_2 > \sum_{\epsilon \in \Lambda} \sum_{r: p(r) > 0} p(r) \mathbb{I}(p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{i \neq C_j(r), i \geq M'+1} p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r). \quad (31)$$

Proof. We defer the proof to Appendix B.4 for brevity, but note that this follows directly from Corollary A.3. See the preceding logic for intuition. \square

Intuitively, $C_j(r)$ singles out the first term in the conditional decomposition of $\mathbb{P}(\eta = \epsilon)$ corresponding to G_j , and (31) indicates that this term $p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) \approx 1/2$ with high probability. Slightly more formally, we could capture this idea by stating that

$$\mathbb{P}((\eta, R) \in \{(\epsilon, r) : p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1\})$$

is quite small.

Step 3: In this step, we combine Steps 1 and 2 to prove the final result. Combining equation (24) and equation (30), we obtain

$$\mathbb{E}_{w,\beta}[\psi(\eta)] = \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r: p(r) > 0} \sum_{\epsilon \in \Lambda_k} p(r) \prod_{i=M'+1}^p p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r).$$

For convenience of notation, we define

$$p(\epsilon | r) = \prod_{i=M'+1}^p p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) = \mathbb{P}(\eta = \epsilon | R = r) \quad (32)$$

and

$$n(\epsilon, r) = \sum_{j=1}^M \mathbb{I}(p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) \geq 0.5 + \alpha_1).$$

Intuitively, $n(\epsilon, r)$ counts the number of pairs G_j which violate the (unconditional) total variation bound from Proposition A.1 by more than α_1 in each conditional decomposition.

Note that Lemma A.3 guarantees that because the sets $\{\Lambda_k\}$ are disjoint, for each pair $1 \leq j \leq M$, we have

$$\sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{\epsilon \in \Lambda_k} \sum_{r:p(r)>0} p(r)p(\epsilon|r) \frac{\mathbb{I}(p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r) \geq 0.5 + \alpha_1)}{p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r)} \leq \alpha_2. \quad (33)$$

The definition of Λ_k explicitly guarantees the denominator in the above expression is nonzero. Moreover, since $p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r) \leq 1$, this implies

$$\sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{\epsilon \in \Lambda_k} \sum_{r:p(r)>0} p(r)p(\epsilon|r) \mathbb{I}(p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r) \geq 0.5 + \alpha_1) \leq \alpha_2,$$

and since this holds for all $1 \leq j \leq M$, we sum over j to obtain

$$\begin{aligned} M \cdot \alpha_2 &\geq \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{\epsilon \in \Lambda_k} \sum_{r:p(r)>0} p(r)p(\epsilon|r) \sum_{j=1}^M \mathbb{I}(p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r) \geq 0.5 + \alpha_1) \\ &= \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{\epsilon \in \Lambda_k} \sum_{r:p(r)>0} p(r)p(\epsilon|r)n(\epsilon, r) \end{aligned} \quad (34)$$

by the definition of $n(\epsilon, r)$. Using this notation, we want to bound

$$\mathbb{E}_{w,\beta}[\psi(\eta)] = \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} \sum_{\epsilon \in \Lambda_k} p(r)p(\epsilon|r). \quad (35)$$

Fix $\ell_0 \in \mathbb{N}$ such that $\ell_0 > 200$. The main idea will be to split the sum in (35) into the two parts defined below:

$$U_k(r) = \left\{ \epsilon \in \Lambda_k : n(\epsilon, r) \geq \frac{\min(k, p)}{\ell_0} \right\} \text{ and } U_k^c(r) = \left\{ \epsilon \in \Lambda_k : n(\epsilon, r) < \frac{\min(k, p)}{\ell_0} \right\}.$$

Intuitively, $U_k(r)$ denotes the set of ϵ such that in the conditional decomposition of $\mathbb{P}(\eta = \epsilon|R = r)$ as defined by (32), many of the conditional probabilities in the product deviate substantially from 0.5. However, Lemma A.3 tells us that the overall probability assigned to this set must be fairly low. In contrast, $U_k^c(r)$ contains a very large number of ϵ , but the probability that η equals any $\epsilon \in U_k^c(r)$ is approximately 2^{-M} .

Applying this decomposition, we have

$$\mathbb{E}_{w,\beta}[\psi(\eta)] = \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \left[\sum_{\epsilon \in U_k(r)} p(\epsilon|r) + \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \right]. \quad (36)$$

We will bound the two terms in (36) separately. First, we consider the set $U_k(r)$. In this set, $n(\epsilon, r)$ is fairly large, which intuitively indicates that many of the $\text{sign}(\text{sorted}(W_{G_j}))$ differ in conditional total variation from $-\text{sign}(\text{sorted}(W_{G_j}))$ conditional on the values of the other groups. We can bound this term fairly quickly using (34). In particular, we have guaranteed that $(1+q) \cdot \ell_0 \cdot n(\epsilon, r) \geq (1+q) \min(k, p) \geq k$ for $\epsilon \in U_k(r)$. Applying (34), this implies,

$$\begin{aligned} (1+q)\ell_0 M \alpha_2 &\geq (1+q)\ell_0 \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{r:p(r)>0} \sum_{\epsilon \in \Lambda_k} p(r)p(\epsilon|r)n(\epsilon, r) \\ &\geq \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} \sum_{r:p(r)>0} \sum_{\epsilon \in U_k(r)} k \cdot p(r)p(\epsilon|r). \end{aligned}$$

Together with (36), this implies

$$\mathbb{E}_{w,\beta}[\psi(\eta)] \leq (1+q)\ell_0 M \alpha_2 + \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r). \quad (37)$$

To bound the term including $U_k^c(r)$, we note $n(\epsilon, r)$ is fairly small for each $\epsilon \in U_k^c(r)$. This means that most of the conditional distributions of $\text{sign}(\text{sorted}(W_{G_j}))$ do not differ too much from those of $-\text{sign}(\text{sorted}(W_{-G_j}))$. With some effort, we will be able to compare this sum to a geometric series to show it is bounded by a linear function of M' .

The key intuition is as follows. For any group ranking $r : p(r) > 0$, the first k coordinates of $\text{sorted}(W)$ must contain elements from at least $(k - M')/2$ unique pairs. In particular, this means it contains at least $(k - M')/2$ feature statistics which appear “first” in their pair, i.e., have a larger absolute value than the other feature statistic in their pair. For any $\epsilon \in U_k^c(r)$, we know that in the conditional decomposition of $p(\epsilon|r)$, all but at most $\frac{\min(k,p)}{\ell_0}$ of the groups G_j are mostly symmetric conditional on $|\text{sorted}(W_{G_j})|$ and the values of W_{-G_j} . This means that in the conditional decomposition of $p(\epsilon|r)$, the first k terms in the product should yield a value less than $(0.5 + \alpha_1)^{(k-M')/2 - \frac{\min(k,p)}{\ell_0}}$. Then, using the definition of Λ_k , we will bound the number of distinct values that $\epsilon_{1:k}$ can take.

To formalize this intuition, for $k \leq p$, define $L_k^c(r)$ as the possible values $\epsilon_{1:k}$ can take for $\epsilon \in U_k^c(r)$, i.e.,

$$L_k^c(r) = \{\epsilon_{1:k} : \epsilon \in U_k^c(r)\} \subset \{-1, 1\}^k.$$

For $k > p$, we define $L_k^c(r) = U_k^c(r)$, since in this case ϵ has only $p < k$ coordinates. Observe by definition of $p(\epsilon|r)$, when $k \leq p$,

$$\begin{aligned} \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) &= \sum_{\epsilon \in U_k^c(r)} \mathbb{P}(\eta = \epsilon \mid R = r) \\ &\leq \sum_{\epsilon_{1:k} \in L_k^c(r)} \mathbb{P}(\eta_{1:k} = \epsilon_{1:k} \mid R = r) \\ &= \sum_{\epsilon_{1:k} \in L_k^c(r)} \prod_{i=M'+1}^k p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r). \end{aligned}$$

Of course, the same logic applies when $k > p$ as long as we are careful to remember that η and ϵ only have p coordinates. This implies that for any k ,

$$\sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \leq \sum_{\epsilon \in L_k^c(r)} \prod_{i=M'+1}^{\min(k,p)} p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) \quad (38)$$

where we are abusing notation slightly in the last equation, since any $\epsilon \in L_k^c(r)$ is an element of $\{-1, 1\}^{\min(k,p)}$. Now, consider the product in (38). When $k > M'$, there must be $\min(k,p) - M'$ terms corresponding to non-singleton feature statistics, and at least $(\min(k,p) - M')/2$ of these correspond to the first sign from one of the groups. However, by the definition of $U_k^c(r)$, at most $\min(k,p)/\ell_0$ of these terms can be greater than $0.5 + \alpha_1$. This implies

$$\sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \leq |L_k^c(r)| (0.5 + \alpha_1)^{(\min(k,p) - M')/2 - \min(k,p)/\ell_0}.$$

Note that for every $\epsilon \in U_k^c(r)$, however, we must have that $\psi(\epsilon) = k$ since $U_k^c(r) \subset \Lambda_k$. By the definition of ψ , this implies that $V_k^+(\epsilon) = \left\lceil \frac{k+1}{1+q} \right\rceil$. Briefly, this must hold because if $V_k^+(\epsilon)$ were any larger, then $\psi(\epsilon) \geq k+1$, but if $V_k^+(\epsilon)$ were any smaller, then $\frac{k-V_k^+(\epsilon)+1}{V_k^+(\epsilon)} \geq q$, contradicting the definition of ψ . We prove this formally

in Lemma B.3. Combining this constraint with the fact that the first M' coordinates of ϵ must equal one, this means that for $k > (1+q)M'$, we have

$$|L_k^c(r)| \leq \#\{\epsilon_{1:k} : \epsilon \in \{-1, 1\}^p \text{ and } \psi(\epsilon) = k\} \leq \binom{\min(k, p) - M'}{\lceil \frac{k+1}{1+q} \rceil - M'},$$

where the binomial coefficient results from the fact that there are $\min(k, p) - M'$ free coordinates of ϵ , of which $\lceil \frac{k+1}{1+q} \rceil - M'$ must equal one.

For an arbitrary constant $C_0 > (1+q)$, the previous logic yields the following bound:

$$\begin{aligned} \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) &\leq C_0 M' + \sum_{k=C_0 M'}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \\ &\leq C_0 M' + \sum_{k=C_0 M'}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \binom{\min(k, p) - M'}{\lceil \frac{k+1}{1+q} \rceil - M'} (0.5 + \alpha_1)^{\frac{\min(k, p) - M'}{2} - \frac{\min(k, p)}{\ell_0}} \\ &\leq C_0 M' + \sum_{k=C_0 M'}^{\lfloor(1+q)p-1\rfloor} k \cdot \binom{\min(k, p) - M'}{\lceil \frac{k+1}{1+q} \rceil - M'} \cdot (0.5 + \alpha_1)^{(\min(k, p) - M')/2 - \min(k, p)/\ell_0}. \end{aligned} \tag{39}$$

The main idea is now to approximate the binomial term as an exponential term that can be dominated by the exponential term. This allows us to simplify (39) and bound it by a linear function of M' .

Lemma A.4. *For sufficiently large C_0 and some universal constants C_1, C_2 ,*

$$\sum_{k=C_0 M'}^{\lfloor(1+q)p-1\rfloor} k \cdot \binom{\min(k, p) - M'}{\lceil \frac{k+1}{1+q} \rceil - M'} \cdot (0.5 + \alpha_1)^{(\min(k, p) - M')/2 - \min(k, p)/\ell_0} \leq C_1 M' + C_2.$$

Note we use the assumptions that $M' \leq \frac{p}{50}$, $\ell_0 > 200$, $\alpha_1 < 0.001$, and $q \leq 0.1$.

Proof. Although this argument is fairly simple, it is quite tedious, so we prove it in Appendix B.4. \square

Combining this with (37), we conclude that

$$\begin{aligned} \mathbb{E}_{w, \beta}[\psi(\eta)] &\leq (1+q)\ell_0 M \alpha_2 + \sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \\ &\leq (1+q)\ell_0 M \alpha_2 + C_1 M' + C_2 \\ &\leq (1+q)\ell_0 p \alpha_2 + C_1 M' + C_2, \end{aligned} \tag{40}$$

where the last line follows because M is the number of groups, so $M \leq p$. This implies that

$$\text{Power}(w, \beta) = \frac{\mathbb{E}_{w, \beta}[\tau]}{p} \leq \frac{C_1 M'}{p} + (1+q)\ell_0 \alpha_2 + \frac{C_2}{p}$$

as desired. \square

Now we can prove Theorem 2.4.

Corollary A.4 (Theorem 2.4). *Let X be a block-equicorrelated Gaussian design with ℓ blocks of equal size and within-block correlations of at least 0.5. Assume $Y | X$ follows a noisy single-index model and let \mathcal{W} be the class of all permutation invariant feature statistic functions. Suppose we aim to control the FDR at level $q \leq 0.1$ using SDP knockoffs.*

Sample $\beta^{(n)}$ uniformly from $\mathcal{C}_{p,b}$, the p -dimensional hypercube centered at 0 with fixed side-length b . Let n be the number of data points and suppose there exists some $\epsilon > 0$ such that $n = o\left(\left(\frac{p}{\ell}\right)^{2-2\epsilon}\right)$. Further assume $\ell = o\left(\frac{p}{\log^{1/\epsilon} p}\right)$. Then as $n, p \rightarrow \infty$, for any $\delta > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\beta^{(n)} \sim \mathcal{C}_{p,b}} \left(\sup_{w \in \mathcal{W}} \text{Power}(\beta^{(n)}, W) \geq \delta \right) = 0.$$

Proof. Note each coefficient $\beta_j^{(n)}$ is sampled independently and uniformly from $(-b, b)$. To construct pairs of approximately alternating signs, as in Theorem A.2, fix an integer $m \in \mathbb{N}$ and let D_k be the set of features in the k th equicorrelated block of $\Sigma^{(n)}$. Define buckets $\{B_{j,k}\}$ for $1 \leq j \leq m$, $1 \leq k \leq \ell$ as follows:

$$B_{j,k} = \left\{ i \in D_k : \frac{(j-1) \cdot b}{m} \leq |\beta_i^{(n)}| \leq \frac{j \cdot b}{m} \right\}. \quad (41)$$

Furthermore, define $B_{j,k}^+ = \left\{ i \in B_{j,k} : \beta_i^{(n)} > 0 \right\}$ and $B_{j,k}^- = \left\{ i \in B_{j,k} : \beta_i^{(n)} < 0 \right\}$. For each bucket $B_{j,k}$, we can create $\min(|B_{j,k}^+|, |B_{j,k}^-|)$ pairs of coefficients such that the sum of each pair is less than $\frac{b}{m}$. Using this bucketing, we have M' singleton features where

$$M' = p - 2 \sum_{k=1}^{\ell} \sum_{j=1}^m \min(|B_{j,k}^+|, |B_{j,k}^-|).$$

Now, we are ready to show that $\mathbb{P}_{\beta^{(n)} \sim \mathcal{C}_{p,b}}(\sup_{w \in \mathcal{W}} \text{Power}(\beta^{(n)}, W) > \delta)$ is bounded by the probability that M' is larger than $\delta_{n,m,p} \cdot p$, for $\delta_{n,m,p}$ defined in a moment. Note that for the rest of this proof, we use \mathbb{P} as a shorthand for $\mathbb{P}_{\beta^{(n)} \sim \mathcal{C}_{p,b}}$. Using the notation of Theorem A.2, if we set

$$\alpha_2 = \frac{2000c_0\sqrt{nb}}{m}$$

and we set $\alpha_1 = 0.0005 < 0.001$, then there are constants C_1, C_2, C_3 such that

$$\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \leq \frac{C_1 M' + C_2}{p} + C_3 \cdot \frac{b\sqrt{n}}{m}.$$

Note that we obtain the “sup” in the preceding equation because Theorem A.2 holds uniformly over all $w \in \mathcal{W}$. Now, the event $\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta$ implies that

$$\delta \leq \sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \leq \frac{C_1 M' + C_2}{p} + C_3 \cdot \frac{b\sqrt{n}}{m},$$

which implies

$$M' \geq \frac{p}{C_1} \left(\delta - C_3 \cdot \frac{b\sqrt{n}}{m} \right) - \frac{C_2}{C_1}.$$

Define $\delta_{p,m,n}$ to be this quantity divided by $2p$, i.e.,

$$\delta_{p,m,n} = \frac{1}{2C_1} \left(\delta - C_3 \cdot \frac{b\sqrt{n}}{m} - \frac{C_2}{p} \right).$$

For now, assume that $\delta_{p,m,n} > 0$: we will show that this holds for a carefully chosen m and sufficiently large p later. By the preceding argument, we have that

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta\right) \leq \mathbb{P}(M' \geq 2\delta_{p,m,n} p). \quad (42)$$

In Appendix B.6, we apply concentration inequalities to prove that for any $\delta > 0$ and for K such that $mK \leq \frac{p}{4\ell}$,

$$\mathbb{P}(M' \geq 2\delta p) \leq 2\ell m \exp(-\delta^2 K) + \ell m \exp\left(-\frac{p}{5\ell m}\right). \quad (43)$$

As a result,

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta\right) \leq 2\ell m \exp(-\delta^2_{m,n,p} K) + \ell m \exp\left(-\frac{p}{5\ell m}\right). \quad (44)$$

For sufficiently large p , we may choose m such that $\delta_{p,m,n} \geq \delta/(4C_1)$. This corresponds to choosing

$$m = \left\lceil \left(\frac{\delta}{2} - \frac{C_2}{p}\right)^{-1} C_3 b \sqrt{n} \right\rceil.$$

Note when $p > (4C_2)/\delta$, we have that $\delta/2 - C_2/p \geq \delta/4$, so

$$m \leq \frac{4C_3 b \sqrt{n}}{\delta}.$$

These two computations yield

$$\begin{aligned} \mathbb{P}\left(\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta\right) &\leq \frac{8bC_3 \cdot \ell \sqrt{n}}{\delta} \exp\left(-\frac{\delta^2}{16C_1^2} K\right) + \frac{4bC_3 \cdot \ell \sqrt{n}}{\delta} \exp\left(-\frac{\delta p}{4C_3 b \cdot \ell \sqrt{n}}\right) \\ &\leq O(\ell \sqrt{n}) \exp\left(-\frac{\delta^2}{16C_1^2} K\right) + O(\ell \sqrt{n}) \exp\left(-\frac{\delta p}{4C_3 b \cdot \ell \sqrt{n}}\right). \end{aligned}$$

At this point, we use the assumption that for some $0 < \epsilon < 1$, $n = o\left(\left(\frac{p}{\ell}\right)^{2-2\epsilon}\right)$ which implies $\sqrt{n} = o\left(\left(\frac{p}{\ell}\right)^{1-\epsilon}\right)$. Let $K = \left(\frac{p}{\ell}\right)^\epsilon$, so $Km = O\left(\sqrt{n} \left(\frac{p}{\ell}\right)^\epsilon\right) = o\left(\frac{p}{\ell}\right)$. This satisfies the constraint $Km < \frac{p}{4\ell}$ for sufficiently large p , which allows us to substitute $\left(\frac{p}{\ell}\right)^\epsilon$ for K and rewrite the constant terms to conclude

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta\right) \leq O(\ell \sqrt{n}) \exp\left(-c_0 \left(\frac{p}{\ell}\right)^\epsilon\right) + O(\ell \sqrt{n}) \exp\left(-c_1 \frac{p}{\ell \sqrt{n}}\right),$$

for positive c_0 and c_1 . We can use the condition on n to bound this by

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \text{Power}(w, \beta^{(n)}) \geq \delta\right) \leq o(\ell^\epsilon p^{1-\epsilon}) \exp\left(-c_0 \left(\frac{p}{\ell}\right)^\epsilon\right) + o(\ell^\epsilon p^{1-\epsilon}) \exp\left(-c_1 \left(\frac{p}{\ell}\right)^\epsilon\right). \quad (45)$$

To show that (45) vanishes, we use the assumption that $\ell = o\left(\frac{p}{\log^{1/\epsilon}(p)}\right)$, which implies $\ell = o\left(\frac{p}{\log^{1/\epsilon}(p^d)}\right)$ for any fixed $d > 0$. Now, to analyze the right-hand term, note that $\frac{p}{\ell} = \Omega\left(\log^{1/\epsilon}(p^d)\right)$, so $\exp\left(-c_1 \left(\frac{p}{\ell}\right)^\epsilon\right) = o(p^{-c_1 d})$. For large enough d , $p^{-c_1 d}$ dominates the $\ell^\epsilon p^{1-\epsilon}$ term, since $\ell^\epsilon p^{1-\epsilon} \leq p^2$. This means that the right-hand term in (45) vanishes. The left-hand term vanishes by the same logic if we replace c_1 with c_0 . Thus, the bound in (45) vanishes, proving the theorem. \square

A.4 Proof of Proposition 2.1

Below, we restate Proposition 2.1 and prove a slightly more specific result, from which the original statement follows.

Proposition 2.1. Suppose $X \sim \mathcal{N}(0, \Sigma)$ for equicorrelated Σ with correlation $\rho \geq 0.5$, and let $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$ with $n > 2p$ and $\gamma < 2 - 2\rho$. Let $S = \gamma I_p$, and let $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ be OLS coefficients fit on $[\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}$. Then

$$\mathbb{E} \left(\sum_{j=1}^p (\hat{\beta}_j^{(\text{ext})} - \hat{\beta}_{j+p}^{(\text{ext})} - \beta_j)^2 \right) \propto \frac{2p}{\gamma} = \frac{2p}{1-v},$$

where $v = 1 - \gamma$ is the MAC. Since $\gamma, v \in (0, 1)$, this is increasing in v .

Proof. Note that because $\Sigma = \rho \mathbf{1}_p \mathbf{1}_p^\top + (1 - \rho) I_p$, we can represent

$$G_S = \begin{bmatrix} \rho \mathbf{1}_p \mathbf{1}_p^\top + (1 - \rho) I_p & \rho \mathbf{1}_p \mathbf{1}_p^\top + (1 - \rho - \gamma) I_p \\ \rho \mathbf{1}_p \mathbf{1}_p^\top + (1 - \rho - \gamma) I_p & \rho \mathbf{1}_p \mathbf{1}_p^\top + (1 - \rho) I_p \end{bmatrix} = \rho \mathbf{1}_{2p} \mathbf{1}_{2p}^\top + A,$$

where

$$A = \begin{bmatrix} (1 - \rho) I_p & (1 - \gamma - \rho) I_p \\ (1 - \gamma - \rho) I_p & (1 - \rho) I_p \end{bmatrix}.$$

It is straightforward to check that if we set $c_1 = \frac{1}{2} \left(\frac{1}{2-2\rho-\gamma} + \frac{1}{\gamma} \right)$ and $c_2 = \frac{1}{2} \left(\frac{1}{2-2\rho-\gamma} - \frac{1}{\gamma} \right)$, then $A^{-1} = \begin{bmatrix} c_1 I_p & c_2 I_p \\ c_2 I_p & c_1 I_p \end{bmatrix}$. Using the Sherman–Morrison inversion formula, we set

$$G_S^{-1} = A^{-1} - \frac{\rho \cdot A^{-1} \mathbf{1}_{2p} \mathbf{1}_{2p}^\top A^{-1}}{1 + \rho \cdot \mathbf{1}_{2p} A^{-1} \mathbf{1}_{2p}^\top} = A^{-1} - c_3 \mathbf{1}_{2p} \mathbf{1}_{2p}^\top \text{ for some } c_3 \in \mathbb{R}. \quad (46)$$

Now we are ready to analyze the estimation error in the theorem statement. We concatenate β and $\mathbf{0} \in \mathbb{R}^p$ to obtain $\beta^{(\text{ext})} = (\beta, \mathbf{0})$. Since $\hat{\beta}^{(\text{ext})}$ are OLS coefficients, we see that conditional on $[\mathbf{X}, \tilde{\mathbf{X}}]$,

$$\hat{\beta}^{(\text{ext})} - \beta^{(\text{ext})} | [\mathbf{X}, \tilde{\mathbf{X}}] \sim \mathcal{N} \left(0, \sigma^2 ([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}])^{-1} \right).$$

Since $[X, \tilde{X}]$ is Gaussian, we know that $\mathbb{E} \left(([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}])^{-1} \right) \propto G_S^{-1}$ and therefore for all $1 \leq j \leq p$,

$$\mathbb{E}[(\hat{\beta}_j^{(\text{ext})} - \hat{\beta}_{j+p}^{(\text{ext})} - \beta_j)^2] \propto (G_S^{-1})_{j,j} + (G_S^{-1})_{j+p,j+p} - 2(G_S^{-1})_{j,j+p}.$$

Combining this with equation (46), we note the c_3 constants cancel to yield that

$$\mathbb{E} \left((\hat{\beta}_j^{(\text{ext})} - \hat{\beta}_{j+p}^{(\text{ext})} - \beta_j)^2 \right) \propto 2c_1 - 2c_2 = \frac{2}{\gamma}.$$

Summing over $1 \leq j \leq p$ proves the theorem. \square

A.5 Discussion of assumptions on q and the coefficient distribution

Recall from Section 2.4 and Proposition A.1 that for equicorrelated designs with $\rho \geq 0.5$, if $\beta_j + \beta_k \approx 0$, a feature statistic like the lasso will be (approximately) equally likely to set $W_j, W_k \approx \beta_j, \beta_k$ or alternatively $W_j, W_k \approx -\beta_k, -\beta_j$. This observation indicates that we probably do not need the number of positive and negative signs in β to balance exactly, nor do we need $q \leq 0.1$. To see why, for simplicity, consider the case where β has coefficients with a constant absolute value with p^+ positive coefficients, p^- negative coefficients, and no null features. Assume without loss of generality that $p^+ > p^-$.

In this setting, the lasso will have p^- opportunities to accidentally flip the sign of a pair of coefficients with alternating signs with probability $\frac{1}{2}$. Heuristically, this means that we should expect $\frac{p^-}{p^+ + p^-}$ of the signs of the non-null W to be negative, even though all of the feature statistics should have roughly the same absolute value (see Proposition A.1). In the knockoffs procedure, however, we can only reject any features at all if there is some set of features with high absolute values such that a proportion greater than $\frac{1}{1+q}$ of

Equicorrelated Design with Varied FDR Control

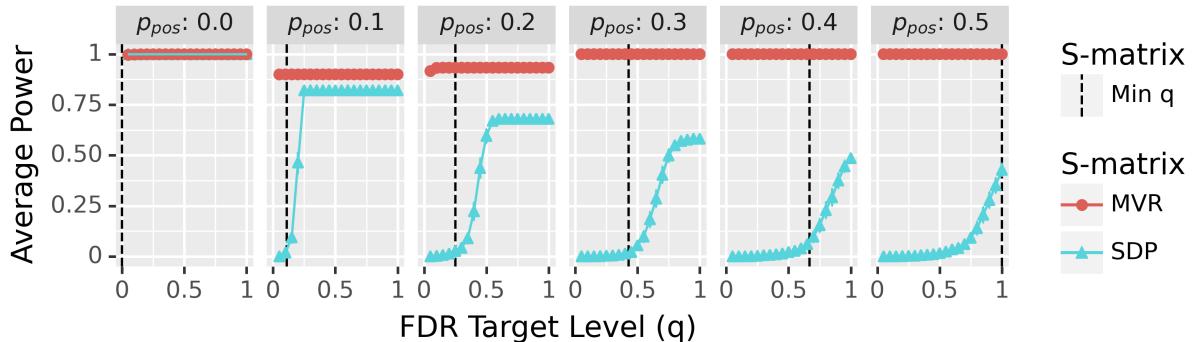


Figure 11: This figure empirically validates the informal conjecture from Appendix A.5. We let the design X be correlated with correlation $\rho = 0.6$, $Y \sim \mathcal{N}(X\beta, 1)$, with $n = 2000$ and $p = 500$. β has 250 non-null values with coefficient magnitudes sampled from $\text{Unif}(2.5, 5)$ and signs sampled from $\text{Bern}(p_{\text{pos}})$. We use lasso coefficient differences as our feature statistics. The dotted black lines are the conjectured minimum q -values.

them have positive signs, where q is the targeted level of FDR control. Heuristically, this suggests that the lasso will only have power if $p_{\text{pos}} = \frac{p^+}{p^+ + p^-} > \frac{1}{1+q}$.

We confirm this informal conjecture in Figure 11, and we see that knockoffs have nearly zero power when $p_{\text{pos}} \leq \frac{1}{1+q}$ even though the non-null coefficients do not have a constant absolute value. The exception is that when $p_{\text{pos}} = 0.5$ and $q \geq 0.75$, where knockoffs have up to 50% power. This behavior is not unexpected, however, since our analysis is asymptotic and p and n may need to be large for our conjecture to hold when q is close to 1. For example, when $q = 0.9$, the procedure will select *every* feature with a positive feature statistic as long as $\frac{1}{1+q} \approx 52.6\%$ of the feature statistics are positive. This will happen fairly frequently unless n and p are quite large, even under the global null, where the signs of the feature statistics are independent and perfectly symmetric. Despite this, note that in all cases, the power of SDP knockoffs when $p_{\text{pos}} = 0.5$ is much less than q , meaning that SDP knockoffs still always have trivial power.

B Technical proofs for Theorem 2.4

In this section, we prove some of the technical lemmas used in Appendix A.

B.1 Verifying the SDP solution for equicorrelated Gaussian designs

Lemma B.1. Suppose $X \sim \mathcal{N}(0, \Sigma)$ where Σ is an equicorrelated correlation matrix with correlation $\rho \geq 0.5$. Then $S_{\text{SDP}} = (2 - 2\rho) \cdot I_p$.

Proof. Recall that

$$S_{\text{SDP}} = \text{diag}(s_{\text{SDP}}),$$

where $s_{\text{SDP}} \in \mathbb{R}^p$ is the solution to the semidefinite program:

$$s_{\text{SDP}} = \arg \min_{s \in \mathbb{R}^p} \|1 - s\|_1 \text{ s.t. } s_j \geq 0 \text{ for each } j \in [p] \text{ and } \text{diag}(s) \preceq 2\Sigma.$$

We can rewrite this in a standardized dual form, as below:

$$s = \arg \max \sum_{j=1}^p s_j \text{ s.t. } \begin{bmatrix} 2\Sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_p \end{bmatrix} - \sum_{j=1}^p s_j A_j \succeq 0,$$

where $A_j \in \mathbb{R}^{3p \times 3p}$ is zero everywhere except $A_{j,j} = A_{2p+j,2p+j} = 1$ and $A_{p+j,p+j} = -1$. This means that

$$s = \arg \max \sum_{j=1}^p s_j \text{ s.t. } \begin{bmatrix} 2\Sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_p \end{bmatrix} - \begin{bmatrix} \text{diag}(s) & 0 & 0 \\ 0 & -\text{diag}(s) & 0 \\ 0 & 0 & \text{diag}(s) \end{bmatrix} \succeq 0.$$

Since a block matrix is positive semidefinite if and only if its blocks are, this constraint guarantees that $2\Sigma - \text{diag}(s) \succeq 0$ and also that $0 \preceq \text{diag}(s) \preceq I_p$.

This is the dual form of the corresponding primal SDP below:

$$\begin{aligned} \text{minimize} \quad & \begin{bmatrix} 2\Sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_p \end{bmatrix} \bullet K \\ \text{s.t.} \quad & A_j \bullet K = 1 \quad \text{for all } 1 \leq j \leq p \\ & K \succeq 0, \end{aligned}$$

where $K \in \mathbb{R}^{3p \times 3p}$ and for any matrix D , $D \bullet K \equiv \sum_{ij} D_{ij} K_{ij} = \text{tr}(DK)$. At this point, for $\rho \geq 0.5$, we define

$$\begin{aligned} s &= (2 - 2\rho) \cdot \mathbf{1}_p, \\ M &= -\frac{1}{p-1} \cdot \mathbf{1}_p \mathbf{1}_p^\top + \frac{p}{p-1} I_p \in \mathbb{R}^{p \times p}, \\ K &= \begin{bmatrix} M & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \end{aligned}$$

which guarantees that K is positive semi-definite. Note s is a feasible solution to the dual problem because (i) each $s_j = 2 - 2\rho \leq 1$ when $\rho \geq 0.5$ and (ii) $\lambda_{\min}(\Sigma) = 1 - \rho$ (see Lemma 2.2), so $0 \preceq \text{diag}(s) \preceq 2\Sigma$. Similarly, $A_j \bullet K = 1$ by construction of K and definition of A_j for each j , so K is a feasible solution as well. Finally, we note that the duality gap is zero, as

$$\sum_{j=1}^p s_j = (2 - 2\rho) \cdot p$$

and similarly

$$\begin{bmatrix} 2\Sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_p \end{bmatrix} \bullet K = 2 \sum_{1 \leq i, j \leq p} \Sigma_{ij} M_{ij} = 2p - \frac{1}{p-1} \cdot p \cdot (p-1) \cdot (2\rho) = (2 - 2\rho) \cdot p,$$

where the $2p$ term comes from the diagonals of M and the second term comes from the $p \cdot (p-1)$ off-diagonal elements of M . Since the dual gap is zero, the optimal solution for s is $(2 - 2\rho) \cdot \mathbf{1}_p$, which proves that $S_{\text{SDP}} = (2 - 2\rho) \cdot I_p$. \square

B.2 Proof of Corollary A.3

In Proposition A.1, we proved that for block-equicorrelated Gaussian designs with correlations greater than 0.5 and a single-index response model, if $J = \{j_1, j_2\} \subset [p]$ lie in the same equicorrelated block and $|\beta_{j_1} + \beta_{j_2}| < d$, then if we apply a permutation invariant w to SDP knockoffs,

$$d_{\text{TV}}([\text{sign}(\text{sorted}(W_J)), W_{-J}], [-\text{sign}(\text{sorted}(W_J)), W_{-J}]) \leq c_0 \sqrt{nd}.$$

Now, we prove Corollary A.3, which extends this result to apply conditionally.

Theorem B.2 (Corollary A.3). *In the same setting as Proposition A.1, let α_1, α_2 be constants such that $\alpha_1 \cdot \alpha_2 > c_0\sqrt{nd}$ and let A be some event in the sigma-algebra generated by $[\|\text{sorted}(W_J)\|, W_{-J}]$. Then if $\mathbb{P}(A) > \alpha_2$, we have that*

$$d_{\text{TV}}([\text{sorted}(W_J), W_{-J}], [-\text{sorted}(W_J), W_{-J}] \mid A) < \alpha_1. \quad (47)$$

Proof. Suppose for the sake of contradiction that equation (47) fails to hold and that $\mathbb{P}(A) > \alpha_2$. Then there must be some set B_1 such that if we define

$$I^+(W) = \{[\text{sorted}(W_J), W_{-J}] \in B_1\} \text{ and } I^-(W) = \{[-\text{sorted}(W_J), W_{-J}] \in B_1\},$$

then we have

$$\mathbb{P}(I^+(W) \mid A) - \mathbb{P}(I^-(W) \mid A) \geq \alpha_1.$$

By the definition of conditional probability, this means that

$$\mathbb{P}(I^+(W) \cap A) - \mathbb{P}(I^-(W) \cap A) \geq \alpha_1 \cdot \mathbb{P}(A) > \alpha_1 \cdot \alpha_2 > c_0\sqrt{nd}.$$

Since A is in the sigma-algebra generated by $[\|\text{sorted}(W_J)\|, W_{-J}]$, there is a measurable set B_2 such that

$$\{[\text{sorted}(W_J), W_{-J}] \in B_2\} = \{[-\text{sorted}(W_J), W_{-J}] \in B_2\} = A.$$

Applying this equivalence and the definitions of $I^+(W), I^-(W)$, we obtain a violation of the marginal total variation bound in Proposition A.1:

$$\mathbb{P}([\text{sorted}(W_J), W_{-J}] \in B_1 \cap B_2) - \mathbb{P}([- \text{sorted}(W_J), W_{-J}] \in B_1 \cap B_2) > c_0\sqrt{nd}.$$

□

B.3 Basic properties of the knockoff filter

In this section we prove two simple lemmas about ψ and the knockoff filter. As a quick reminder, recall that for any $\epsilon \in \{-1, 1\}^p$ and any $k \in \mathbb{N}$, we define $V_k^+(\epsilon)$ to count the number of ones in the first k coordinates of ϵ . More precisely,

$$V_k^+(\epsilon) = \#\{j \leq k : \epsilon_j = 1\}.$$

Furthermore, as in (21), we define $\psi(\epsilon)$ as follows:

$$\psi(\epsilon) = \max_{1 \leq k \leq \infty} \left\{ k : \frac{k - V_k^+(\epsilon) + 1}{V_k^+(\epsilon)} \leq q \right\}.$$

As discussed in the proof of Theorem 2.4, intuitively ψ is a reformulation of the knockoff data-dependent threshold T , such that the threshold depends only on the signs of the sorted feature statistics W . Indeed, the next lemma tells us that the number of discoveries made by the knockoffs procedure is a deterministic, increasing function of $\psi(\text{sign}(\text{sorted}(W)))$.

Lemma B.3. *Let W be feature statistics generated by any knockoff procedure in any setting, and let $\eta = \text{sign}(\text{sorted}(W))$ be the signs of the sorted feature statistics sorted by absolute value in descending order. Let τ be the number of discoveries of this procedure. Then when $\tau > 0$:*

$$\tau = \left\lceil \frac{\psi(\eta) + 1}{1 + q} \right\rceil \leq \psi(\eta). \quad (48)$$

Moreover, let $\epsilon \in \{-1, 1\}^p$ such that $\psi(\epsilon) = k > 0$. Then

$$V_k^+(\epsilon) = \left\lceil \frac{k + 1}{1 + q} \right\rceil. \quad (49)$$

Proof. First we prove (49). If $\psi(\epsilon) = k$, then

$$\frac{k - V_k^+(\epsilon) + 1}{V_k^+(\epsilon)} \leq q \implies V_k^+(\epsilon) \geq \frac{k+1}{1+q}$$

by the definition of ψ . As a result, to show (49), it suffices to show that $V_k^+(\epsilon) < \frac{k+2}{1+q}$. Suppose for the sake of contradiction that $V_k^+(\epsilon) \geq \frac{k+2}{1+q}$. This implies that

$$\frac{k - V_k^+(\epsilon) + 2}{V_k^+(\epsilon)} \leq q.$$

Since $V_{k+1}^+(\epsilon) \geq V_k^+(\epsilon)$, this implies that

$$\frac{k+1 - V_{k+1}^+(\epsilon) + 1}{V_{k+1}^+(\epsilon)} \leq \frac{k + V_k^+(\epsilon) - 2}{V_k^+(\epsilon)} \leq q,$$

where we are using the fact that $(k - V_k^+(\epsilon) + 2)/V_k^+(\epsilon)$ is decreasing in $V_k^+(\epsilon)$. However, by definition of ψ , this implies that $\psi(\epsilon) \geq k+1$. This is a contradiction, since we assumed $\psi(\epsilon) = k$. Thus (49) must hold.

Now we prove (48). In the knockoffs procedure, to control the FDR at level q , we define the data-dependent threshold

$$T = \min_{1 \leq i \leq p} \left\{ |W_i| : \frac{\#\{j : W_j \leq -|W_i|\} + 1}{\#\{j : W_j \geq |W_i|\}} \leq q \right\} \setminus \{0\}$$

and then we reject the null for features $\{j : W_j \geq T\}$. Without loss of generality assume W is sorted by absolute value, e.g. $|W_1| \geq |W_2| \geq \dots \geq |W_p|$, since neither $\psi(\eta)$ nor T depends on the initial order of the feature statistics. Then note that because $\eta = \text{sign}(W)$,

$$\{j : W_j \geq T\} = \{j \leq \psi(\eta) : \eta_j = 1\}.$$

This implies that $\tau = V_{\psi(\eta)}^+(\eta)$, since both procedures accept all of the positive feature statistics between 1 and $\psi(\eta)$. By (49), we conclude that

$$\tau = V_{\psi(\eta)}^+(\eta) = \left\lceil \frac{\psi(\eta) + 1}{1+q} \right\rceil \leq \psi(\eta).$$

□

Lemma B.4. Consider an arbitrary set of feature statistics $W = W_1, \dots, W_t, \dots, W_p$. Define $W^* = W_1^*, \dots, W_t^*$ such that for a fixed $t \in [p]$,

$$W_j^* = \begin{cases} +\infty & j \in \{1, \dots, t\} \\ W_j & \text{else} \end{cases}$$

Let $\tau(W)$ and $\tau(W^*)$ be the number of discoveries made according to the knockoff procedure when applied to W and W^* . Then $\tau(W) \leq \tau(W^*)$.

Proof. Let $\eta = \text{sign}(\text{sorted}(W))$ and $\eta^* = \text{sign}(\text{sorted}(W^*))$. In Lemma B.3, we proved that the number of discoveries is fully determined by and increasing in $\psi(\eta)$. Therefore it suffices to show $\psi(\eta) \leq \psi(\eta^*)$. By definition of ψ , we can show this by showing that for all k , $V_k^+(\eta^*) \geq V_k^+(\eta)$. To see this, note by definition of the sort function that since the absolute values of W_1^*, \dots, W_t^* are infinite and all of their signs are positive, we must have that $\eta^* = [\mathbf{1}_t, \text{sign}(\text{sorted}(W_{t+1:p}))]$. This implies that switching from W to W^* can only add positive signs to the first k elements of the path, for any k , since any feature statistic it swaps out is replaced with a positive sign. Therefore $V_k(\eta^*) \geq V_k^+(\eta)$ for all k . □

B.4 Computations from Theorem A.2

In this section, we prove Lemmas A.3 and A.4. Before doing so, it may be helpful to recall some context from Theorem A.2. In Theorem A.2, we grouped the coordinates of W into M' “singleton statistics” and M pairs (so $M' + 2M = p$). We denote the pairs G_1, \dots, G_M , and we constructed them such that if $j_1, j_2 \in G_j$, then $|\beta_{j_1} - \beta_{j_2}| < \frac{b}{m}$, where b is the maximum absolute value of any coefficient in the single-index model and $m > 0$ is an arbitrary natural number. We showed earlier in Corollary A.3 that $\text{sorted}(W_{G_j})$ has an approximately symmetric distribution, even conditional on $|\text{sorted}(W_{G_j})|$ and W_{-G_j} . The purpose of Lemma A.3 is to convert this result into a result that will help us show that $\mathbb{P}(\text{sign}(\text{sorted}(W)) = \epsilon)$ cannot stray above 2^{-M} for too many $\epsilon \in \{-1, 1\}^p$, where the 2^{-M} comes because there are M approximately symmetric pairs of feature statistics.

Lastly, for Lemma A.3, it will be necessary to recall the notation introduced in step 2 of Theorem 2.4. As a brief but incomplete refresher, we let $\eta = \text{sign}(\text{sorted}(W))$, and we defined R to be the random vector which takes a coordinate η_i of the sorted feature statistics and returns the index of the group G_j which η_i corresponds to. We used R to carefully decompose

$$\mathbb{P}(\eta = \epsilon) = \sum_{r:p(r)>0} p(r)p(\epsilon|r) = \sum_{r:p(r)>0} p(r) \prod_{i=M'+1}^p p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r),$$

where the product begins at $M' + 1$ because we make the worst-case assumption that the M' singleton statistics are equal to $+\infty$, meaning that the first M' coordinates of η will always equal 1. Crucially, we noted that the conditional probabilities $p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r)$ can be written purely in terms of the features $\{\text{sorted}(W_j)\}_{j=1}^M$ and R , allowing us to apply Corollary A.3. In Lemma A.3, we use the notation $C_j(r)$ to denote the first coordinate of η that corresponds to group G_j —for this reason, one can intuitively interpret the statement of Lemma A.3 to mean that $\text{sorted}(W_{G_j}) \mid R, \text{sorted}(W)_{1:C_j(r)-1}$ is symmetric with very high probability.

For more details, see step 2 of Theorem A.2.

Lemma A.3. *Let α_1, α_2 be constants such that $\alpha_1 \alpha_2 > \frac{c_0 \sqrt{n} b}{m}$. For any $1 \leq j \leq M$ and for any $\Lambda \subset \{-1, 1\}^p$,*

$$\alpha_2 > \sum_{\epsilon \in \Lambda} \sum_{r:p(r)>0} p(r) \mathbb{I}(p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{i \neq C_j(r), i \geq M'+1} p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r). \quad (50)$$

Proof. It suffices to prove (50) for $\Lambda = \{-1, 1\}^p$ since every element in the above sum is nonnegative. We will employ a proof by contradiction, so at this point we assume that (50) does not hold. The proof proceeds in two steps. First, we simplify the sum in (50). Second, we will use (50) to construct a violation of Corollary A.3. In particular, we will define a set A such that $\mathbb{P}(A) \geq \alpha_2/2$ but conditional on A , the total variation distance between $\text{sorted}(W_{G_j})$ and $-\text{sorted}(W_{G_j})$ is at least $2\alpha_1$.

Step 1: Simplification. Note that for any r , if we fix the first $C_j(r)$ coordinates of ϵ , we are summing over every possible value of the last $p - C_j(r)$ coordinates of ϵ . Since the last $p - C_j(r)$ coordinates of ϵ must take *some* value, all conditional probabilities depending on these last $p - C_j(r)$ coordinates should sum to one. The effect of this is that in the above product, we can drop all terms in the product where $i > C_j(r)$ and only sum over the first $C_j(r)$ coordinates of ϵ , as below.²

$$\begin{aligned} \alpha_2 &\leq \sum_{r:p(r)>0} \sum_{\epsilon \in \{-1, 1\}^{C_j(r)}} p(r) \mathbb{I}(p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{i \neq C_j(r), i \geq M'+1} p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r) \\ &= \sum_{r:p(r)>0} \sum_{\epsilon \in \{-1, 1\}^{C_j(r)}} p(r) \mathbb{I}(p(\epsilon_{C_j(r)} | \epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{M'+1 \leq i < C_j(r)} p(\epsilon_i | \epsilon_{(M'+1):(i-1)}, r). \end{aligned}$$

²Note that this logic does not apply to all p coordinates of ϵ since the indicator variable in (50) is not a conditional probability. However, it only depends on the first $C_j(r)$ coordinates of ϵ .

Note that in the latter equation, we slightly abuse notation and let ϵ denote a $C_j(r)$ -length vector instead of a p -length vector. Next, observe that there are only 2 possible values for $\epsilon_{C_j(r)} \in \{-1, 1\}$. Therefore by the pigeonhole principle, there exists some *fixed* $v \in \{-1, 1\}$ such that

$$\alpha_2/2 \leq \sum_{r:p(r)>0} \sum_{\epsilon \in \{-1, 1\}^{C_j(r)-1}} p(r) \mathbb{I}(p(v|\epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{M'+1 \leq i < C_j(r)} p(\epsilon_i|\epsilon_{(M'+1):(i-1)}, r). \quad (51)$$

Step 2: At this point, we will construct an event A such that $\mathbb{P}(A)$ equals the right-hand side of (51) and furthermore such that the distribution of $\text{sorted}(W_{G_j})$ is highly asymmetric conditional on A . To do this, recall that the definition of $C_j(r)$ guarantees that $\epsilon_{C_j(r)}$ corresponds to $\text{sorted}(W_{G_j})_1$. This plus the definition of $p(\epsilon_{C_j(r)}|\epsilon_{(M'+1):(C_j(r)-1)}, r)$, from (29), yields that

$$p(v|\epsilon_{(M'+1):(C_j(r)-1)}, r) = \mathbb{P}\left(\text{sign}(\text{sorted}(W_{G_j}))_1 = v \middle| \bigcap_{i=M'+1}^{C_j(r)-1} \text{sign}(\text{sorted}(W_{G_{r_i}}))_{N_i(r)} = \epsilon_i, R = r\right).$$

Define Λ_r to be the set of $\epsilon \in \{-1, 1\}^{C_j(r)-1}$ such that the above probability is greater than $0.5 + \alpha_1$. Formally,

$$\Lambda_r = \left\{ \epsilon \in \{-1, 1\}^{C_j(r)-1} : p(v|\epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1 \right\}.$$

Then define the event A as below:

$$A = \bigcup_{r:p(r)>0} \left(\{R = r\} \cap \left\{ \text{sign}(\text{sorted}(W))_{1:(C_j(r)-1)} \in \Lambda_r \right\} \right). \quad (52)$$

Note that A is in the sigma-algebra generated by $[\text{sorted}(W_{G_j}), W_{-G_j}]$ for two reasons. First, as noted in Theorem A.2, R is in the sigma-algebra generated by $\{\text{sorted}(W_{G_j})\}_{j=1}^M$. Second, v is fixed and $\text{sign}(\text{sorted}(W))_{1:C_j(r)-1}$ only depends on R and W_{-G_j} , by the definition of $C_j(r)$.

By the definition of A , we have that

$$\begin{aligned} \mathbb{P}(A) &= \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in \Lambda_r} \prod_{M'+1 \leq i < C_j(r)} p(\epsilon_i|\epsilon_{(M'+1):(i-1)}, r) \\ &= \sum_{r:p(r)>0} \sum_{\epsilon \in \{-1, 1\}^{C_j(r)-1}} p(r) \mathbb{I}(p(v|\epsilon_{(M'+1):(C_j(r)-1)}, r) > 0.5 + \alpha_1) \prod_{M'+1 \leq i < C_j(r)} p(\epsilon_i|\epsilon_{(M'+1):(i-1)}, r) \\ &\geq \alpha_2/2 \quad \text{by equation (51),} \end{aligned}$$

where the second step follows by definition of Λ_r , and in both steps, we slightly abuse notation so that ϵ is a $C_j(r) - 1$ dimensional vector. Thus, the law of total probability plus the definition of Λ_r implies that

$$\mathbb{P}(\text{sign}(\text{sorted}(W_{G_j}))_1 = v | A) > 0.5 + \alpha_1$$

which implies that

$$d_{\text{TV}}(\text{sorted}(W_{G_j}), -\text{sorted}(W_{G_j}) | A) > 2\alpha_1.$$

However, this contradicts Corollary A.3 since $2\alpha_1\alpha_2/2 > \frac{c_0\sqrt{nb}}{m}$. \square

Lemma A.4. *Under the conditions and notation of Theorem A.2, specifically regarding equation (39),*

$$\sum_{k=C_0 M'}^{\lfloor (1+q)p-1 \rfloor} k \cdot \left(\min(k, p) - M' \right) \cdot (0.5 + \alpha_1)^{(\min(k, p) - M')/2 - \min(k, p)/\ell_0} \leq C_1 M' + C_2$$

for sufficiently large C_0 and some universal constants C_1, C_2 . Note we use the assumptions that $M' \leq \frac{p}{50}$, $\ell_0 > 200$, $\alpha_1 < 0.001$, and $q \leq 0.1$.

Proof. We first change variables by resetting k to $k + C_0 M'$.

$$\begin{aligned} &= \sum_{k=0}^{\lfloor(1+q)p-1\rfloor-C_0M'} (k + C_0 M') \binom{\min(k + C_0 M', p) - M'}{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'} (0.5 + \alpha_1)^{(\min(k+C_0M',p)-M')/2 - \min(k+C_0M',p)/\ell_0} \\ &= C_0 M' \sum_{k=0}^{\lfloor(1+q)p-1\rfloor-C_0M'} \binom{\min(k + C_0 M', p) - M'}{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'} \cdot (0.5 + \alpha_1)^{(\min(k+C_0M',p)-M')/2 - \min(k+C_0M',p)/\ell_0} \quad (53) \end{aligned}$$

$$+ \sum_{k=0}^{\lfloor(1+q)p-1\rfloor-C_0M'} k \cdot \binom{\min(k + C_0 M', p) - M'}{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'} \cdot (0.5 + \alpha_1)^{(\min(k+C_0M',p)-M')/2 - \min(k+C_0M',p)/\ell_0} \quad (54)$$

We will show that (53) can be bounded by a linear function of M' , and that (54) is bounded by a constant which does not grow with M' or p . Our initial task will be to analyze the components of these sums, with the eventual goal of comparing them to a geometric series. For both the binomial term and the exponential term, we will break analysis into two cases: first, when $k + C_0 M' \leq p$, and second, when $k + C_0 M' > p$.

First, we will analyze the exponential term. In the first case where $k + C_0 M' \leq p$, we have that

$$\begin{aligned} (0.5 + \alpha_1)^{(\min(k+C_0M',p)-M')/2 - \min(k+C_0M',p)/\ell_0} &= (0.5 + \alpha_1)^{\left(\frac{1}{2} - \frac{1}{\ell_0}\right)(k+C_0M') - M'/2} \\ &\leq 2^{\left(\frac{1}{2} - \frac{1}{\ell_0}\right)\log_2(0.5+\alpha_1)(k+C_0M') + M'/2} \\ &\leq 2^{-0.49k + (-0.49C_0 + 0.5)M'}, \end{aligned}$$

where we obtain the 0.49 constant using the fact that $\alpha_1 < 0.001$ and $\ell_0 > 200$. We can pick C_0 large enough such that this is less than

$$\leq 2^{-0.49k - 0.48C_0M'}. \quad (55)$$

In the second case, when $k + C_0 M' > p$, we have that

$$(0.5 + \alpha_1)^{(\min(k+C_0M',p)-M')/2 - \min(k+C_0M',p)/\ell_0} \leq 2^{-0.49p - M'/2} \leq 2^{-0.48p} \quad (56)$$

where the last equation follows from the assumption that $M' \leq \frac{p}{50}$.

Second, we will analyze the binomial coefficient appearing in (53) and (54). We use the noncentral binomial bound below:

$$\frac{2^{nH(d/n)}}{\sqrt{n}} \leq \binom{n}{d} \leq 2^{nH(d/n)} \quad (57)$$

where $H(x) = -x \log_2(x) - (1-x) \log_2(1-x)$ is the binary entropy function. In the first case when $k + C_0 M' \leq p$,

$$\binom{\min(k + C_0 M', p) - M'}{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'} \leq 2^{H(a)(k+C_0M'-M')},$$

where we define

$$a = \frac{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'}{k + (C_0 - 1)M'} \geq \frac{\frac{k+(C_0-1)M'}{1+q} - \frac{qM'}{1+q}}{k + (C_0 - 1)M'} \geq \frac{1}{1+q} - \frac{qM'}{(k + (C_0 - 1)M')} \geq \frac{1}{1.1} - \frac{0.1}{C_0 - 1},$$

where in the last step on the right, we apply the fact that $q \leq 0.1$. Note that many of the bounds above are quite loose. The main point, however, is that we always have that $a > 0.5$ for $C_0 > 2$, and H is continuous and decreasing on $[0.5, 1]$. Note as C_0 gets larger, our lower bound on a approaches $1/(1.1)$. Thus, we can pick C_0 large enough to guarantee that $H(a) \leq H\left(\frac{1}{1.1} + \frac{0.1}{C_0-1}\right) < 0.44$, where the number 0.44 comes from the fact that $H(1/(1.1)) < 0.44$. All together, this implies that in the first case,

$$\binom{\min(k + C_0 M', p) - M'}{\lceil \frac{k+C_0M'+1}{1+q} \rceil - M'} < 2^{0.44k + 0.44C_0M'}. \quad (58)$$

In the second case, when $k + C_0 M' > p$, we will use the fact that $\binom{n}{d}$ is decreasing in d when $n/2 \leq d$. That assumption holds in this instance, as since $M' \leq p/50$,

$$\frac{\min(k + C_0 M', p) - M'}{2} \leq \frac{p}{2} \leq \left(\frac{1}{2} + \frac{1}{50} \right) p - M' \leq \left\lceil \frac{k + C_0 M' + 1}{1+q} \right\rceil - M'$$

where in the last equation, we use the fact that $q \leq 0.1$, so $1/2 + 1/50 \leq 1/1.1 \leq \frac{1}{1+q}$. As a result, in the second case,

$$\begin{aligned} \left(\frac{\min(k + C_0 M', p) - M'}{\lceil \frac{k + C_0 M' + 1}{1+q} \rceil - M'} \right) &= \left(\frac{p - M'}{\lceil \frac{k + C_0 M' + 1}{1+q} \rceil - M'} \right) \\ &\leq \left(\frac{p - M'}{\lceil \frac{p}{1+q} \rceil - M'} \right) \\ &\leq 2^{H\left(\frac{\lceil p/(1+q) \rceil - M'}{p - M'}\right)(p - M')}. \end{aligned} \quad (59)$$

Analyzing the $H(\cdot)$ term in (59), we use the assumptions that $M' \leq \frac{p}{50}$ and $q \leq 0.1$ to find that

$$\frac{\lceil p/(1+q) \rceil - M'}{p - M'} \geq \frac{p/(1.1) - \frac{p}{50}}{p - \frac{p}{50}} = \frac{\frac{1}{1.1} - \frac{1}{50}}{\frac{49}{50}} \geq 0.907.$$

Note that $H(0.907) < 0.45$, so $H\left(\frac{\lceil p/(1+q) \rceil - M'}{p - M'}\right) < 0.45$. This implies that when $p < k + C_0 M'$,

$$\left(\frac{\min(k + C_0 M', p) - M'}{\lceil \frac{k + C_0 M' + 1}{1+q} \rceil - M'} \right) < 2^{H\left(\frac{\lceil p/(1+q) \rceil - M'}{p - M'}\right)(p - M')} \leq 2^{0.45p}. \quad (60)$$

Now, we will work with the sum (53). As a reminder, we need to show that this sum can be bounded by a universal constant not depending on p or M' . We split this sum into two parts which correspond to the cases where $k + C_0 M' \leq p$ and vice versa:

$$\begin{aligned} &\sum_{k=0}^{\lfloor (1+q)p-1 \rfloor - C_0 M'} \left(\frac{\min(k + C_0 M', p) - M'}{\lceil \frac{k + C_0 M' + 1}{1+q} \rceil - M'} \right) \cdot (0.5 + \alpha_1)^{(\min(k + C_0 M', p) - M')/2 - \min(k + C_0 M', p)/\ell_0} \\ &\leq \sum_{k=0}^{p - C_0 M'} 2^{0.44k + 0.44C_0 M'} 2^{-0.49k - 0.48C_0 M'} + \sum_{k=p - C_0 M' + 1}^{\lfloor (1+q)p-1 \rfloor - C_0 M'} 2^{0.45p} 2^{-0.48p} \\ &\leq 2^{-0.04C_0 M'} \sum_{k=0}^{p - C_0 M'} 2^{-0.05k} + (1+q)p 2^{-0.03p}. \end{aligned} \quad (61)$$

The sum in the first term is bounded as $p \rightarrow \infty$, and all other terms are asymptotically decreasing in p and M' . This means that we can bound this by a constant which does not depend on p or M' .

Next, we will work with the sum (54) using almost exactly the same argument. In particular,

$$\begin{aligned} &\sum_{k=0}^{\lfloor (1+q)p-1 \rfloor - C_0 M'} k \cdot \left(\frac{\min(k + C_0 M', p) - M'}{\lceil \frac{k + C_0 M' + 1}{1+q} \rceil - M'} \right) \cdot (0.5 + \alpha_1)^{(\min(k + C_0 M', p) - M')/2 - \min(k + C_0 M', p)/\ell_0} \\ &\leq \sum_{k=0}^{p - C_0 M'} k \cdot 2^{0.44k + 0.44C_0 M'} 2^{-0.49k - 0.48C_0 M'} + \sum_{k=p - C_0 M' + 1}^{\lfloor (1+q)p-1 \rfloor - C_0 M'} k \cdot 2^{0.45p} 2^{-0.48p} \\ &\leq 2^{-0.04C_0 M'} \sum_{k=0}^{p - C_0 M'} k 2^{-0.05k} + (1+q)^2 p^2 2^{-0.03p}. \end{aligned} \quad (62)$$

Since the sum on the left converges and all other terms vanish as $M', p \rightarrow \infty$, this equation can be uniformly bounded as well. Combining this with (53) and (54), there are constants C_1 and C_2 such that

$$\sum_{k=\lfloor(1+q)M'-1\rfloor}^{\lfloor(1+q)p-1\rfloor} k \sum_{r:p(r)>0} p(r) \sum_{\epsilon \in U_k^c(r)} p(\epsilon|r) \leq C_1 M' + C_2 \quad (63)$$

holds uniformly under the conditions in the theorem. \square

B.5 Total variation bounds

In this section, we will make some simple applications of the following theorem, which is reproduced from Devroye et al. (2018):

Theorem B.5 (From Devroye et al. (2018)). *Let Z, Z' be p -dimensional multivariate normals with equal means but different covariance matrices E_Z and $E_{Z'}$. Then*

$$\frac{1}{100} \leq \frac{d_{\text{TV}}(Z, Z')}{\min\{1, \|E_{Z'} E_Z^{-1} - I_p\|_F\}} \leq \frac{3}{2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. On the other hand, if Z and Z' are univariate Gaussians with means μ_1, μ_2 and equal variances σ^2 , then

$$d_{\text{TV}}(Z, Z') \leq \frac{|\mu_1 - \mu_2|}{2\sigma}.$$

Lemma B.6. *Let $X \sim \mathcal{N}(0, \Sigma)$ and let $Y | X \sim \mathcal{N}(X\beta, \sigma_0^2)$. Let $\beta' \in \mathbb{R}^p$ such that β and β' differ in at most one coordinate. Suppose without loss of generality that this is the first coordinate and $|\beta_1 - \beta'_1| = d$. Then let $Y' | X \sim \mathcal{N}(X\beta', \sigma_0^2)$. If $\mathbf{X}, \mathbf{y}, \mathbf{y}'$ are the data matrix and response vectors for n independent data points, then there is a universal constant c_0 such that*

$$d_{\text{TV}}([\mathbf{X}, \mathbf{y}], [\mathbf{X}, \mathbf{y}']) \leq \frac{c_0 \sqrt{nd}}{\sigma_0}.$$

Proof. First, consider a single observation (X, Y, Y') , i.e., the case where $n = 1$. Then by Theorem B.5,

$$d_{\text{TV}}([X, Y], [X, Y' | X]) = d_{\text{TV}}(Y, Y' | X) \leq \frac{|X_1\beta_1 - X_1\beta'_1|}{2\sigma_0} = \frac{|X_1||\beta_1 - \beta'_1|}{2\sigma_0} = \frac{d}{2\sigma_0} \cdot |X_1|.$$

For any arbitrary random vectors R_1, R_2, R_3 , we know $d_{\text{TV}}(R_1, R_2) \leq \mathbb{E}[d_{\text{TV}}(R_1, R_2 | R_3)]$. This plus the prior equation yields that

$$d_{\text{TV}}([X, Y], [X, Y']) \leq \frac{d}{2\sigma_0} \mathbb{E}[|X_1|] = \frac{d}{\sigma_0 \sqrt{2\pi}},$$

since marginally $X_1 \sim \mathcal{N}(0, 1)$, so $|X_1| \sim \chi_1$. Now consider the case where $n \geq 1$. Note that the i.i.d. rows (X, Y) and (X, Y') are both $p+1$ dimensional Gaussians with mean 0. Let their covariance matrices be Γ_1 and Γ_2 . Applying Theorem B.5, this implies

$$\min\{1, \|\Gamma_1 \Gamma_2^{-1} - I_{p+1}\|_F\} \leq \frac{C_0 d}{\sigma_0},$$

where $C_0 \leq \frac{100}{\sqrt{2\pi}}$. Note this bound on C_0 is fairly loose, as pointed out by Devroye et al. (2018), who did not optimize the constants in Theorem B.5.

Now, we can think of (\mathbf{X}, \mathbf{y}) and $(\mathbf{X}, \mathbf{y}')$ as $n(p+1)$ dimensional multivariate Gaussians, with mean 0 and block-diagonal covariance matrices $\Gamma_1^{(n)}$ and $\Gamma_2^{(n)}$ (respectively), where

$$\Gamma_1^{(n)} = \begin{bmatrix} \Gamma_1 & 0 & \dots & 0 \\ 0 & \Gamma_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \Gamma_1 \end{bmatrix} \text{ and } \Gamma_2^{(n)} = \begin{bmatrix} \Gamma_2 & 0 & \dots & 0 \\ 0 & \Gamma_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \Gamma_2 \end{bmatrix}.$$

As a result, we see that

$$\begin{aligned} \|\Gamma_1^{(n)} \left(\Gamma_2^{(n)}\right)^{-1} - I_{n(p+1)}\|_F &= \left\| \begin{bmatrix} \Gamma_1 & 0 & \dots & 0 \\ 0 & \Gamma_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Gamma_1 \end{bmatrix} \begin{bmatrix} \Gamma_2^{-1} & 0 & \dots & 0 \\ 0 & \Gamma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Gamma_2^{-1} \end{bmatrix} - I_{(p+1)n} \right\|_F \\ &= \left\| \begin{bmatrix} \Gamma_1 \Gamma_2^{-1} & 0 & \dots & 0 \\ 0 & \Gamma_1 \Gamma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Gamma_1 \Gamma_2^{-1} \end{bmatrix} - I_{(p+1)n} \right\|_F = \sqrt{n} \|\Gamma_1 \Gamma_2^{-1} - I_{p+1}\|_F \leq \frac{C_0 \sqrt{nd}}{\sigma_0^2}. \end{aligned}$$

Here we are using the fact that the inverse of a block-diagonal matrix is simply the inverse of the diagonals and applying the definition of the Frobenius norm. Combined with the second theorem from Devroye et. al, this tells us that for n data points,

$$d_{\text{TV}}((\mathbf{X}, \mathbf{y}), (\mathbf{X}, \mathbf{y}')) \leq \frac{3C_0 \sqrt{nd}}{2\sigma_0},$$

which proves the lemma if we reset C_0 to be $3/2$ times its original value. \square

B.6 Lemmas about random coefficients

Let $\mathcal{C}_{b,p} = [-b, b]^p$ be the p -dimensional cube with side-length $2b$ centered at 0 and suppose β is drawn uniformly from $\mathcal{C}_{b,p}$. Note this is equivalent to letting $|\beta_i| \stackrel{i.i.d.}{\sim} \text{Unif}(0, b)$, $\text{sign}(\beta_i) \stackrel{i.i.d.}{\sim} \pm 1$ with equal probability, with $\text{sign}(\beta_i) \perp \!\!\! \perp |\beta_i|$. In this section, we analyze the distribution of the sizes of the buckets of coefficients defined in equation (41). As a quick reminder, we fix m and let D_k be the set of indices corresponding to block k . Then define

$$B_{j,k} = \left\{ i \in D_k : \frac{(j-1) \cdot b}{m} \leq |\beta_i| \leq \frac{j \cdot b}{m} \right\}.$$

Furthermore, we let $B_{j,k}^+ = \{i \in B_{j,k} : \beta_i > 0\}$ and $B_{j,k}^- = \{i \in B_{j,k} : \beta_i < 0\}$.

Lemma B.7. *Given the prior definitions, fix integers $m, K > 0$ such that $mK < \frac{p}{4\ell}$. Then for sufficiently large m ,*

$$\mathbb{P} \left(\min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}| < K \right) \leq \ell m \exp \left(-\frac{p}{5\ell m} \right).$$

Proof. Note that for any $j \in [m], k \in [\ell]$,

$$|B_{j,k}| \sim \text{Bin} \left(\frac{p}{\ell}, \frac{1}{m} \right).$$

Since $\frac{\ell K}{p} < \frac{1}{4m}$, we can use the Binomial bound in Theorem 1 of Arratia and Gordon (1989) which states that

$$\mathbb{P}(|B_{j,k}| \leq K) \leq \exp \left(-\frac{p}{\ell} D \left(\frac{\ell K}{p} \middle\| \frac{1}{m} \right) \right),$$

where $D(p_1 || p_2)$ denotes the relative entropy between two Bernoulli random variables with probability p_1 and p_2 . Define $x_1 = \ell K / p$ and $x_2 = 1/m$ and recall that we assume $x_1 < x_2/4$. When m is sufficiently large, x_2 will approach zero, ensuring that x_1 does as well, and finally that $(x_2 - x_1)/(1 - x_2)$ will too. As a result, we can apply the approximation $\log(1 + (x_2 - x_1)/(1 - x_2)) \geq c_0(x_2 - x_1)/(1 - x_2)$ for some $c_0 < 1$. Note

that by the prior logic, for any fixed c_0 , this approximation is valid for any sufficiently large m under the theorem conditions, regardless of the values of p or K . This approximation yields

$$\begin{aligned} D(x_1||x_2) &= x_1 \log\left(\frac{x_1}{x_2}\right) + (1-x_2) \log\left(\frac{1-x_1}{1-x_2}\right) \\ &= x_1 \log\left(\frac{x_1}{x_2}\right) + (1-x_2) \log\left(1 + \frac{x_2-x_1}{1-x_2}\right) \\ &\geq x_1 \log\left(\frac{x_1}{x_2}\right) + c_0(1-x_2) \frac{x_2-x_1}{1-x_2} \\ &= x_1 \log\left(\frac{x_1}{x_2}\right) + c_0 x_2 - c_0 x_1 \\ &\geq x_1 \log\left(\frac{x_1}{x_2}\right) + \frac{3c_0 x_2}{4} \end{aligned}$$

where the last step uses that $x_1 < \frac{x_2}{4}$. We will show that the final expression is greater than $x_2/5 = 1/(5m)$. To see this, note that for sufficiently large m , the above equation holds for c_0 large enough such that $3c_0/4 > (1/2 + 1/5)$. Then, we need only show that $x_2/2 > -x_1 \log(x_1/x_2)$. Note both quantities in the comparison are positive, so it suffices to show their ratio is greater than 1. Using the fact that $x_2/x_1 > 4$, we see

$$\frac{x_2/2}{-x_1 \log(x_1/x_2)} = \frac{1}{2} \cdot \frac{x_2/x_1}{\log(x_2/x_1)} > \frac{1}{2} \frac{4}{\log(4)} \geq 1,$$

where we use the fact that the function $\frac{x}{\log(x)}$ is increasing on (e, ∞) . This proves that $D(x_1||x_2) > 1/(5m)$. Applying this to the initial binomial bound, we obtain

$$\mathbb{P}(|B_{j,k}| \leq K) \leq \exp\left(-\frac{p}{5\ell m}\right).$$

Therefore by the union bound,

$$\mathbb{P}\left(\min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}| \leq K\right) = \mathbb{P}\left(\bigcup_{1 \leq j \leq m} \bigcup_{1 \leq k \leq \ell} |B_{j,k}| \leq K\right) \leq \ell m \exp\left(-\frac{p}{5\ell m}\right),$$

which completes the proof of the Lemma. \square

Lemma B.8. *In the previous setting, define*

$$M' = 2 \sum_{j=1}^m \sum_{k=1}^{\ell} \max(|B_{j,k}^+|, |B_{j,k}^-|) - p = p - 2 \sum_{j=1}^m \sum_{k=1}^{\ell} \min(|B_{j,k}^+|, |B_{j,k}^-|).$$

Fix integers $m, K > 0$ such that $mK < \frac{p}{4\ell}$. Then for sufficiently large m and all $\delta > 0$,

$$\mathbb{P}(M' > 2\delta p) \leq 2\ell m \exp(-\delta^2 K) + \ell m \exp\left(-\frac{p}{5\ell m}\right).$$

Proof. Let \mathcal{B} be the sigma-algebra generated by $\{|B_{j,k}|\}$. Observe for every j, k ,

$$\max(|B_{j,k}^+|, |B_{j,k}^-|) \leq \left(\frac{1}{2} + \delta\right) |B_{j,k}| + \frac{p}{\ell} \mathbb{I}\left(\max(|B_{j,k}^+|, |B_{j,k}^-|) > \left(\frac{1}{2} + \delta\right) |B_{j,k}|\right).$$

Summing over j and k , we find that

$$\sum_{j=1}^m \sum_{k=1}^{\ell} \max(|B_{j,k}^+|, |B_{j,k}^-|) \leq \left(\frac{1}{2} + \delta\right) \sum_{j=1}^m \sum_{k=1}^{\ell} |B_{j,k}| + \frac{p}{\ell} \sum_{j=1}^m \sum_{k=1}^{\ell} \mathbb{I}\left(\max(|B_{j,k}^+|, |B_{j,k}^-|) > (1/2 + \delta)|B_{j,k}|\right).$$

Plugging this in yields

$$\begin{aligned}
\mathbb{P}(M' \geq 2\delta p | \mathcal{B}) &= \mathbb{P}\left(\sum_{j=1}^m \sum_{k=1}^{\ell} \max(|B_{j,k}^+|, |B_{j,k}^-|) > \left(\delta + \frac{1}{2}\right)p | \mathcal{B}\right) \\
&\leq \mathbb{P}\left(\left(\delta + \frac{1}{2}\right)\sum_{j=1}^m \sum_{k=1}^{\ell} |B_{j,k}| + \frac{p}{\ell} \sum_{j=1}^m \sum_{k=1}^{\ell} \mathbb{I}\left(\max(|B_{j,k}^+|, |B_{j,k}^-|) > \left(\delta + \frac{1}{2}\right)|B_{j,k}|\right) > \left(\delta + \frac{1}{2}\right)p | \mathcal{B}\right) \\
&= \mathbb{P}\left(\sum_{j=1}^m \sum_{k=1}^{\ell} \mathbb{I}\left(\max(|B_{j,k}^+|, |B_{j,k}^-|) > \left(\delta + \frac{1}{2}\right)|B_{j,k}|\right) > 0 | \mathcal{B}\right), \tag{64}
\end{aligned}$$

where in the third line, we apply the fact that $\sum_{j,k} |B_{j,k}| = p$. At this point, we observe that

$$|B_{j,k}^+| \mid \mathcal{B} \sim \text{Bin}\left(|B_{j,k}|, \frac{1}{2}\right)$$

and $|B_{j,k}^-|$ has the same distribution conditional on \mathcal{B} . Therefore we apply a union bound plus Hoeffding's inequality for binomials to conclude:

$$\begin{aligned}
\mathbb{P}\left(\max(|B_{j,k}^+|, |B_{j,k}^-|) > \left(\delta + \frac{1}{2}\right)|B_{j,k}| \mid \mathcal{B}\right) &\leq \mathbb{P}\left(|B_{j,k}^+| > \left(\delta + \frac{1}{2}\right)|B_{j,k}| \mid \mathcal{B}\right) + \mathbb{P}\left(|B_{j,k}^-| > \left(\delta + \frac{1}{2}\right)|B_{j,k}| \mid \mathcal{B}\right) \\
&= 2\mathbb{P}\left(|B_{j,k}^+| > \left(\delta + \frac{1}{2}\right)|B_{j,k}| \mid \mathcal{B}\right) \\
&\leq 2\exp(-\delta^2|B_{j,k}|) \quad \text{by Hoeffding's inequality} \\
&\leq 2\exp(-\delta^2 \min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}|). \tag{65}
\end{aligned}$$

Applying a union bound to equations (64) and (65), we obtain the conditional bound

$$\mathbb{P}_{\beta}(M' \geq 2\delta p | \mathcal{B}) \leq 2\ell m \exp(-\delta^2 \min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}|).$$

This implies that unconditionally, for any K ,

$$\begin{aligned}
\mathbb{P}_{\beta}(M' \geq 2\delta p) &\leq \mathbb{P}_{\beta}\left(M' \geq 2\delta p \mid \min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}| \geq K\right) + \mathbb{P}_{\beta}\left(\min_{1 \leq j \leq m} \min_{1 \leq k \leq \ell} |B_{j,k}| < K\right) \\
&\leq 2\ell m \exp(-\delta^2 K) + \ell m \exp\left(-\frac{p}{5\ell m}\right),
\end{aligned}$$

where the last substitution combines the previous argument with Lemma B.7.

□

C Theory of MRC knockoffs

In this section, we prove Proposition 3.1 and Theorem 3.1. We also provide a few extra simulations demonstrating the effect of reconstruction on power, as explained in Section 3.

C.1 Estimation error

Proposition 3.1. Suppose $X \sim \mathcal{N}(0, \Sigma)$ for any Σ and $Y \mid X \sim \mathcal{N}(X\beta, \sigma^2)$. Let $\beta^{(\text{ext})} \in \mathbb{R}^{2p}$ be the concatenation of $\beta \in \mathbb{R}^p$ with p zeros. Suppose $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $[\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}$ and $n > 2p + 1$. Then

$$S_{\text{MVR}} = \arg \min_S \mathbb{E}[||\hat{\beta}^{(\text{ext})} - \beta^{(\text{ext})}||_2^2].$$

Proof. Fix S such that G_S is positive definite (otherwise we cannot fit OLS statistics). Since $\tilde{X} \perp\!\!\!\perp Y | X$,

$$Y | [X, \tilde{X}] \sim \mathcal{N}([X, \tilde{X}]\beta^{(\text{ext})}, \sigma^2).$$

Since $\hat{\beta}^{(\text{ext})}$ is the OLS statistic on $[\mathbf{X}, \tilde{\mathbf{X}}]$ and \mathbf{y} ,

$$\hat{\beta}^{(\text{ext})} | [\mathbf{X}, \tilde{\mathbf{X}}] \sim \mathcal{N}\left(\beta^{(\text{ext})}, \sigma^2 \left([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}]\right)^{-1}\right). \quad (66)$$

By construction, $[X, \tilde{X}] \sim \mathcal{N}(0, G_S)$, so $\left([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}]\right) \sim \mathbb{W}_{2p}(n, G_S)$ where \mathbb{W}_{2p} denotes the $2p$ -dimensional Wishart distribution. When $n > 2p$, $\left([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}]\right)^{-1} \sim \mathbb{W}_{2p}^{-1}(n, G_S)$ where \mathbb{W}_{2p}^{-1} is the inverse-Wishart distribution. The law of iterated expectation implies that for any $j \in [2p]$,

$$\mathbb{E}\left[\left(\hat{\beta}_j^{(\text{ext})} - \beta_j^{(\text{ext})}\right)^2\right] = \mathbb{E}\left[\sigma^2 \left(\left([\mathbf{X}, \tilde{\mathbf{X}}]^\top [\mathbf{X}, \tilde{\mathbf{X}}]\right)^{-1}\right)_{j,j}\right] = \frac{\sigma^2}{n - 2p - 1} (G_S^{-1})_{j,j}.$$

Summing over j , this yields

$$\mathbb{E}[|\hat{\beta}^{(\text{ext})} - \beta^{(\text{ext})}|_2^2] = \frac{\sigma^2}{n - 2p - 1} \text{Tr}(G_S^{-1}) \propto L_{\text{MVR}}(S). \quad (67)$$

Since S_{MVR} minimizes $L_{\text{MVR}}(S)$, it also minimizes $\mathbb{E}[|\hat{\beta}^{(\text{ext})} - \beta^{(\text{ext})}|_2^2]$. \square

C.2 Consistency

Theorem 3.1 tells us that when $X \sim \mathcal{N}(0, \Sigma)$ and $Y | X \sim \mathcal{N}(X\beta, \sigma^2)$, the power of OLS absolute coefficient difference statistics applied to MVR knockoffs converges to 1 in low-dimensional settings. To prove this, we first prove that such feature statistics W converge in squared ℓ_2 norm to the linear coefficients β .

Theorem C.1. Suppose $X \sim \mathcal{N}(0, \Sigma^{(n)})$, $Y | X \sim \mathcal{N}(X\beta^{(n)}, \sigma^2)$, and \tilde{X} is generated using S_{MVR} . Finally, suppose $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$, and let $W_j = |\hat{\beta}_j^{(\text{ext})}| - |\hat{\beta}_{j+p}^{(\text{ext})}|$ for $1 \leq j \leq p$.

Consider a sequence of covariance matrices $\Sigma^{(n)}$ such that the minimum eigenvalue of $\Sigma^{(n)} \in \mathbb{R}^{p \times p}$ is bounded uniformly above a fixed constant $\gamma \in \mathbb{R}^+$. Let $n, p \rightarrow \infty$ and assume that $p = o(n)$. Then for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{\beta^{(n)} \in \mathbb{R}^p} \mathbb{P}_{\beta^{(n)}} \left[\|W - |\beta^{(n)}|\|_2^2 > \epsilon \right] = 0.$$

Proof. Define $S^{(n)} = \gamma I_p$. For each $\Sigma^{(n)}$, let $G_{\text{MVR}}^{(n)}$ denote the G -matrix formed from $\Sigma^{(n)}$ and the MVR solution, and let $G_S^{(n)}$ be the G -matrix from $\Sigma^{(n)}$ and $S^{(n)}$. Standard Schur-complement analysis yields that

$$\text{Tr}((G_S^{(n)})^{-1}) = \sum_{j=1}^{2p} \frac{1}{\lambda_j(G_S^{(n)})} = \sum_{j=1}^p \frac{1}{\lambda_j(S^{(n)})} + \sum_{j=1}^p \frac{1}{\lambda_j(2\Sigma^{(n)} - S^{(n)})} \leq \frac{2p}{\gamma}.$$

The last step follows because by definition all of the eigenvalues of $S^{(n)}$ are γ , and since $\lambda_j(\Sigma^{(n)}) \geq \gamma$ for each j , we must have that $\lambda_j(2\Sigma^{(n)} - S^{(n)}) \geq \gamma$ as well. At this point, we note that by definition of the MVR loss, we have that

$$\text{Tr}\left(\left(G_{\text{MVR}}^{(n)}\right)^{-1}\right) \leq \text{Tr}\left(\left(G_S^{(n)}\right)^{-1}\right) \leq \frac{2p}{\gamma}.$$

Using equation (67) from Proposition 3.1, this implies that if we let $\beta^{(n,\text{ext})} \in \mathbb{R}^{2p}$ be the concatenation of $\beta^{(n)}$ with p zeros,

$$\mathbb{E}[|\hat{\beta}^{(\text{ext})} - \beta^{(n,\text{ext})}|_2^2] = \frac{\sigma^2}{n - 2p - 1} \text{Tr}\left(\left(G_{\text{MVR}}^{(n)}\right)^{-1}\right) \leq \frac{2p\sigma^2}{\gamma(n - 2p - 1)}.$$

This holds uniformly over $\beta^{(n)}$, implying

$$\lim_{n \rightarrow \infty} \sup_{\beta^{(n)} \in \mathbb{R}^p} \mathbb{E}[||\hat{\beta}^{(\text{ext})} - \beta^{(n, \text{ext})}||_2^2] \leq \lim_{n \rightarrow \infty} \frac{2p\sigma^2}{\gamma(n-2p-1)} = 0$$

where the last step follows because $p = o(n)$. Next, observe that

$$\begin{aligned} ||W - |\beta^{(n)}||_2^2 &= \sum_{j=1}^p \left(|\hat{\beta}_j^{(\text{ext})}| - |\hat{\beta}_{j+p}^{(\text{ext})}| - |\beta_j^{(n)}| \right)^2 \\ &= \sum_{j=1}^p \left(|\hat{\beta}_j^{(\text{ext})}| - |\beta_j^{(n)}| \right)^2 + \left(\hat{\beta}_{j+p}^{(\text{ext})} \right)^2 - 2 \sum_{j=1}^p \left(|\hat{\beta}_j^{(\text{ext})}| - |\beta_j^{(n)}| \right) \left| \hat{\beta}_{j+p}^{(\text{ext})} \right| \\ &\leq \sum_{j=1}^p (\hat{\beta}_j^{(\text{ext})} - \beta_j^{(n)})^2 + (\hat{\beta}_{j+p}^{(\text{ext})})^2 + 2 \sum_{j=1}^p \left| \hat{\beta}_j^{(\text{ext})} - \beta_j^{(n)} \right| \left| \hat{\beta}_{j+p}^{(\text{ext})} \right| \text{ by the reverse triangle inequality} \\ &\leq \left\| \hat{\beta}^{(\text{ext})} - \beta^{(n, \text{ext})} \right\|_2^2 + 2 \sqrt{\left\| \hat{\beta}_{1:p}^{(\text{ext})} - \beta^{(n)} \right\|_2^2 \cdot \left\| \hat{\beta}_{(p+1):2p}^{(\text{ext})} \right\|_2^2} \text{ by the Cauchy-Schwartz inequality} \\ &\leq 3 \left\| \hat{\beta}^{(\text{ext})} - \beta^{(n, \text{ext})} \right\|_2^2. \end{aligned}$$

This implies that $||W - |\beta^{(n)}||_2^2$ must converge in expectation to zero uniformly over $\beta^{(n)}$. This implies that it also converges in probability to zero uniformly over $\beta^{(n)}$, which proves the theorem. \square

Theorem 3.1. Suppose $X \sim \mathcal{N}(0, \Sigma^{(n)})$, $Y | X \sim \mathcal{N}(X\beta^{(n)}, \sigma^2)$, and \tilde{X} is generated using S_{MVR} . Suppose $\hat{\beta}^{(\text{ext})} \in \mathbb{R}^{2p}$ are OLS coefficients fit on $([X, \tilde{X}], y)$, and set $w([X, \tilde{X}], y) = |\hat{\beta}_{1:p}^{(\text{ext})}| - |\hat{\beta}_{(p+1):2p}^{(\text{ext})}|$.

Let $n, p \rightarrow \infty$ such that $p = o(n)$ and consider a sequence of covariance matrices $\Sigma^{(n)} \in \mathbb{R}^{p \times p}$ such that the minimum eigenvalue of $\Sigma^{(n)}$ is bounded above a fixed constant $\gamma \in \mathbb{R}^+$. Suppose we sample a sequence of random $\beta^{(n)}$ as follows. Let all but a uniformly drawn subset of $\lceil s_0 p \rceil$ entries of $\beta^{(n)}$ equal zero, for a fixed constant $s_0 \in (0, 1]$, and then sample the remaining (non-null) entries of $\beta^{(n)}$ from a $\lceil s_0 p \rceil$ -dimensional hypercube centered at 0 with any fixed side-length. Then

$$\text{Power}(w, \beta^{(n)}) \xrightarrow{P} 1.$$

Proof. Fix $\epsilon > 0$. We will show that $\lim_{n \rightarrow \infty} \mathbb{P}(\text{Power}(w, \beta^{(n)}) \geq 1 - \epsilon) = 1$. Note throughout this proof, we use $\mathbb{P}_{\beta^{(n)}}$ to denote a probability over the data for a fixed $\beta^{(n)}$, and we use \mathbb{P} when the probability is over a random $\beta^{(n)}$. Fix $\epsilon_0 > 0$ and define $B_{\epsilon_0}^{(n)} = \{j : |\beta_j^{(n)}| > \epsilon_0\}$, i.e., the set of all non-nulls with a coefficient values at least ϵ_0 away from zero. Then note that

$$||W - |\beta^{(n)}||_2^2 < \frac{\epsilon_0^2}{4} \implies \#\{j \in B_{\epsilon_0}^{(n)} : W_j \leq \frac{\epsilon_0}{2}\} = 0 \text{ and } \#\{j : W_j \leq -\frac{\epsilon_0}{2}\} = 0.$$

Intuitively, the former statement tells us that when W approximates $|\beta^{(n)}|$ sufficiently closely in ℓ_2 norm, all of the feature statistics corresponding to $B_{\epsilon_0}^{(n)}$ must be greater than $\epsilon_0/2$, and no feature statistics can be smaller than $-\epsilon_0/2$.

Note that by the definition of the knockoffs procedure, which ranks the W statistics by absolute values, the events above imply that the $|B_{\epsilon_0}^{(n)}|$ largest feature statistics are all positive. When $|B_{\epsilon_0}^{(n)}| \geq \lceil \frac{1}{1+q} \rceil$, this implies that the knockoffs procedure will reject at least $|B_{\epsilon_0}^{(n)}|$ non-nulls. Let τ be the number of non-nulls rejected by the procedure. Then this means that for sufficiently large $|B_{\epsilon_0}^{(n)}|$,

$$||W - |\beta^{(n)}||_2^2 < \frac{\epsilon_0^2}{4} \implies \tau \geq |B_{\epsilon_0}^{(n)}|.$$

This statement plus Theorem C.1 tells us that if $\mathcal{R}^{(n)} = \left\{ \beta^{(n)} \in \mathbb{R}^p : |B_{\epsilon_0}^{(n)}| \geq \left\lceil \frac{1}{1+q} \right\rceil \right\}$, then

$$\lim_{n \rightarrow \infty} \sup_{\beta^{(n)} \in \mathcal{R}^{(n)}} \mathbb{P}_{\beta^{(n)}} \left(\tau < |B_{\epsilon_0}^{(n)}| \right) = 0.$$

By the definition of power, this implies that for any $\epsilon_1 > 0$ and sufficiently large p , the following holds deterministically for any $\beta^{(n)}$ such that $|B_{\epsilon_0}^{(n)}| \geq \lceil \frac{1}{1+q} \rceil$:

$$\text{Power}(w, \beta^{(n)}) \geq (1 - \epsilon_1) \frac{|B_{\epsilon_0}^{(n)}|}{\lceil s_0 p \rceil}.$$

If j is one of the $\lceil s_0 p \rceil$ non-null coordinates, then $|\beta_j^{(n)}| \stackrel{i.i.d.}{\sim} \text{Unif}(0, b)$. This implies $\mathbb{P}(|\beta_j^{(n)}| \geq \epsilon_0) = 1 - \frac{\epsilon_0}{b}$. Since each coordinate is independent, the law of large numbers implies

$$\frac{|B_{\epsilon_0}^{(n)}|}{\lceil s_0 p \rceil} \xrightarrow{p} 1 - \frac{\epsilon_0}{b}.$$

Therefore, for any $\delta > 0$, there are sufficiently large n, p such that

$$\mathbb{P} \left(|B_{\epsilon_0}^{(n)}| > \left(1 - \frac{2\epsilon_0}{b} \right) \lceil s_0 p \rceil \right) \geq 1 - \delta,$$

and note that as p grows, this event guarantees that $\beta^{(n)} \in \mathcal{R}^{(n)}$. All of the prior analysis implies that for any $\epsilon_0, \epsilon_1, \delta > 0$, there exist sufficiently large n, p such that

$$\mathbb{P} \left(\text{Power}(w, \beta^{(n)}) > (1 - \epsilon_1) \left(1 - \frac{2\epsilon_0}{b} \right) \right) \geq 1 - \delta.$$

If we pick ϵ_0, ϵ_1 small enough such that $(1 - \epsilon_1) \left(1 - \frac{2\epsilon_0}{b} \right) \geq 1 - \epsilon$, this implies that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{Power}(w, \beta^{(n)}) \geq 1 - \epsilon) = 1.$$

□

C.3 The importance of harshly penalizing high levels of reconstructability

When sampling \tilde{X} , we will likely face trade-offs where to reduce the reconstructability of some feature X_j , we must increase the reconstructability of another feature X_k in order to maintain the pairwise exchangeability condition (1). To navigate these trade-offs, in Section 3.1, we suggested a general principle of harshly penalizing high levels of reconstructability to ensure no feature X_j is highly reconstructable from X_{-j}, \tilde{X} . In particular, we argued in Section 3.1 that if a non-null X_j is reconstructable using a knockoff \tilde{X}_k , then \tilde{X}_k may be assigned X_j 's variable importance, causing the feature statistic W_k to have a large magnitude but a negative sign.

As a concrete example, consider the case where $X \sim \mathcal{N}(0, \Sigma)$ for equicorrelated Σ with $\rho \geq 0.5$, as discussed in Section 2. Further suppose that $Y | X \sim \mathcal{N}(X\beta, 1)$ and we use SDP knockoffs, so $\text{Var}(X_j | X_{-j}, \tilde{X}) = 0$ for all j . Fix $j, k \in [p]$ and suppose that $\beta_k \approx 1$ and $\beta_j \approx -1$. Even though SDP knockoffs are asymptotically powerless in this setting, the intuition in Section 2.4 suggests that common feature statistics like lasso coefficients may correctly estimate the magnitude of W_k and W_j , such that $|W_k|, |W_j| \rightarrow 1$. However, with approximately 50% probability, the lasso will assign the feature importance of X_k to \tilde{X}_j and the feature importance of X_j to \tilde{X}_k , so $(W_k, W_j) \stackrel{d}{\approx} (-W_j, -W_k)$, as in Corollary A.3. This makes it difficult to discover *any* non-null features, because the feature statistics often have large magnitudes and negative signs.

Figure 2 in Section 3 demonstrates that for equicorrelated Gaussian random variables with a linear response, the reconstruction effect for SDP knockoffs causes many feature statistics to have large absolute values but

negative signs. In Figure 12, we demonstrate that the same effect occurs for non-exchangeable Gaussian features. In particular, we compute feature statistics for all of the data-generating processes described in Section 4.2, with the following specific parameters. In all cases we set $p = 100$ and $n = 190$ with $Y | X \sim \mathcal{N}(X\beta, 1)$ where β has 50 non-nulls with values ± 1 . We set $\rho = 0.6$ for the equicorrelated and block-equicorrelated designs, $a = 3$ for the AR1 data-generating processes, and the sparsity parameter equal to 0.2 for the ErdosRenyi designs.

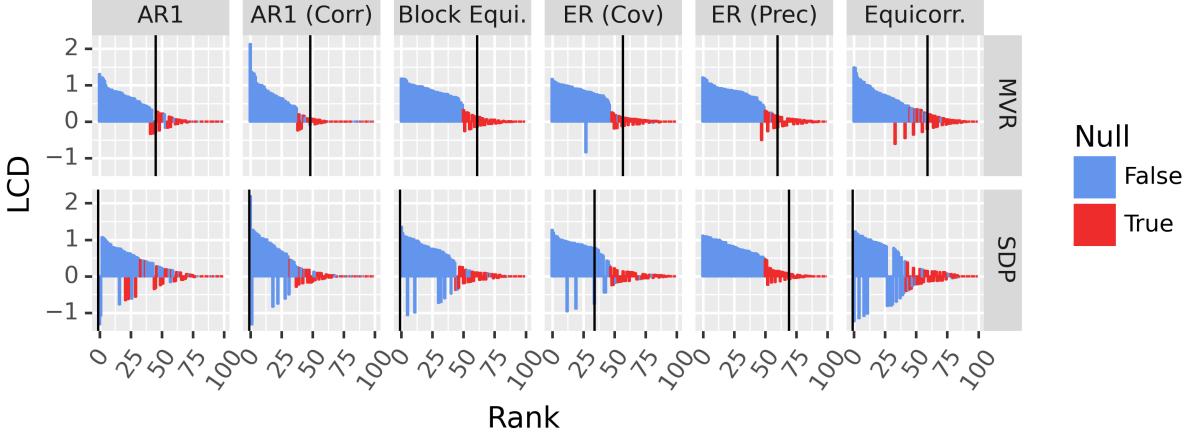


Figure 12: We plot LCD statistics sorted in descending order of absolute value for several Gaussian linear models. The horizontal facets correspond to the design distributions specified above. Note $Y | X \sim \mathcal{N}(X\beta, 1)$ where β has 50 non-nulls with values ± 1 , with equal probability. The black lines denote the data-dependent thresholds.

D Convexity and computation for MRC knockoffs

D.1 Proof of convexity

In this section, we prove that the MVR formulation in the Gaussian case can be reduced to a simple semidefinite program. Note Gimenez and Zou (2019) have previously shown a similar result for the ME formulation.

Lemma D.1. *Consider the MVR optimization problem $\min_S \text{Tr}(G_S^{-1})$ such that $0 \preceq S \preceq 2\Sigma$ with S diagonal. Let $R_1, R_2 \in \mathbb{R}^{p \times p}$ be slack variables and, as usual, S is a diagonal matrix. Then the MVR optimization problem is equivalent to the following SDP.*

$$\begin{aligned} & \text{minimize } \text{Tr}(R_1) + \text{Tr}(R_2) \\ & \text{subject to } B(R_1, R_2, S) \equiv \begin{bmatrix} R_1 & 0 & I_p & 0 \\ 0 & R_2 & 0 & I_p \\ I_p & 0 & S & 0 \\ 0 & I_p & 0 & 2\Sigma - S \end{bmatrix} \succcurlyeq 0. \end{aligned}$$

Proof. Recall that $L_{\text{MVR}}(S) \propto \text{Tr}(G_S^{-1}) = \text{Tr}((2\Sigma - S)^{-1}) + \text{Tr}(S^{-1})$. Schur complement analysis yields

$$B(R_1, R_2, S) \succcurlyeq 0 \text{ if and only if } \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} - \begin{bmatrix} S & 0 \\ 0 & 2\Sigma - S \end{bmatrix}^{-1} \succcurlyeq 0$$

which in turn implies that for any feasible solution,

$$\text{Tr}(R_1) + \text{Tr}(R_2) \geq \text{Tr}(S^{-1}) + \text{Tr}((2\Sigma - S)^{-1}).$$

Additionally, for a fixed S , one can always set $R_1 = S^{-1}$ and $R_2 = (2\Sigma - S)^{-1}$ and achieve $\text{Tr}(R_1) + \text{Tr}(R_2) = \text{Tr}(S^{-1}) + \text{Tr}((2\Sigma - S)^{-1})$. Therefore, minimizing $\text{Tr}(R_1) + \text{Tr}(R_2)$ is equivalent to minimizing $\text{Tr}(S^{-1}) + \text{Tr}(2\Sigma - S)^{-1}) \propto L_{\text{MVR}}(S)$. \square

Unfortunately, generic solvers for this problem may be as slow as $O(p^6)$ (Pipeleers and Vandenberghe, 2011). In the next two sections, we develop a much faster algorithm to compute S_{MVR} and S_{ME} .

D.2 Computing MVR knockoffs

In this section, we introduce Algorithm 1, which computes S_{MVR} in $O(n_{\text{iter}}p^3)$. This algorithm is inspired by Askari et al. (2020) and uses their overall strategy, although the technical details differ.

The key idea behind the algorithm is as follows. Fix $j \in [p]$ and a diagonal matrix $S \succcurlyeq 0$ such that $D \equiv 2\Sigma - S \succcurlyeq 0$. Furthermore, let M be the matrix of all zeros except $M_{j,j} = 1$. We first use Schur complements to decompose

$$L_{\text{MVR}}(S + \delta_j M) = \text{Tr}(2\Sigma - S - \delta_j M)^{-1} + \text{Tr}(S + \delta_j M)^{-1}. \quad (68)$$

We seek to find δ_j which minimizes this quantity. Since $S + \delta_j M$ is diagonal, $\text{Tr}(S + \delta_j M)^{-1} = 1/(S_{j,j} + \delta_j) + \sum_{k \neq j} \frac{1}{S_{k,k}}$, and only the term $1/(S_{j,j} + \delta_j)$ depends on δ_j . Then, we apply the Sherman–Morrison rank-one inversion formula to note

$$(D - \delta_j M)^{-1} = D^{-1} - \frac{-\delta_j D^{-1} M D^{-1}}{1 - \delta_j D_{j,j}^{-1}}, \quad (69)$$

where D is constant with respect to δ_j . Let $c_n = -\text{Tr}(D^{-1} M D^{-1})$ and let $c_d = D_{j,j}^{-1}$. Then, equations (68) and (69) imply that

$$L_{\text{MVR}}(S + \delta_j M) = \frac{1}{S_{j,j} + \delta_j} - \frac{\delta_j c_n}{1 - \delta_j c_d} + c \quad (70)$$

for some constant c . Taking the derivative of equation (70) with respect to δ_j yields the quadratic optimality condition (71), which can be solved in constant time:

$$(-c_n - c_d^2)\delta_j^2 + 2(-c_n S_{j,j} + c_d)\delta_j + (-c_n S_{j,j}^2 - 1) = 0. \quad (71)$$

Note equation (71) and the constraint from (70) that $-S_{j,j} < \delta_j < \frac{1}{c_d}$ yield a unique solution for δ_j . To compute c_n and c_d efficiently, we could initially compute D^{-1} and then update D^{-1} at each step using rank-1 updates, which has time complexity $O(p^2)$. Unfortunately, as Askari et al. (2020) observed, the rank-1 updates to D^{-1} are numerically unstable. Following their approach, we detail an alternative which is equally efficient but maintains a Cholesky decomposition of D instead of maintaining D^{-1} .

Indeed, given a Cholesky decomposition of $D = LL^\top$, we can compute c_n and c_d in $O(p^2)$. To see how, let e_j be the j th basis vector. Using the fact that $\text{Tr}(uv^\top) = u^\top v$ for any two vectors $u, v \in \mathbb{R}^p$, we can then represent

$$c_n = -(D^{-1}e_j)^\top(D^{-1}e_j) \text{ and } c_d = e_j^\top D^{-1}e_j. \quad (72)$$

To compute c_n , note if we let $v_n = D^{-1}e_j$, then $c_n = -\|v_n\|_2^2$. However, we can solve for v_n by solving the system $LL^\top v_n = e_j$. Since L is triangular, this can be solved in $O(p^2)$ using forward-backward substitution. Second, we observe that $e_j^\top D^{-1}e_j = e_j^\top(LL^\top)^{-1}e_j = e_j^\top(L^\top)^{-1}L^{-1}e_j = (L^{-1}e_j)^\top(L^{-1}e_j)$. Therefore, if we let $v_d = L^{-1}e_j$, then $c_d = \|v_d\|_2^2$. Since L is triangular, we can find v_d as the solution to $Lv_d = e_j$ in $O(p^2)$.

This motivates the following algorithm, which maintains a running copy of LL^\top . After updating the value of $S_{j,j}$, we perform a rank one update of L , which can be done in $O(p^2)$. The result is Algorithm 1, which is numerically stable and runs in $O(n_{\text{iter}}p^3)$.

Algorithm 1 Stable Coordinate Descent for MVR Knockoffs

- 1: Initialize $S = \lambda_{\min}(\Sigma) \cdot I_p$
 - 2: Solve for L such that $LL^\top = 2\Sigma - S$
 - 3: **for** $\ell = 1, 2, \dots, n_{\text{iter}}$ **do**
 - 4: **for** $j = 1, 2, \dots, p$ **do**
 - 5: Compute c_n and c_d using L as detailed in (72)
 - 6: Solve δ_j^* as the solution to the optimality condition (71)
 - 7: Set $S_{j,j} = S_{j,j} + \delta_j^*$
 - 8: Compute the rank-1 update to L for δ_j^*
-

D.3 Coordinate descent for ME knockoffs

In this section, we introduce a coordinate descent algorithm to efficiently compute the S -matrix for ME knockoffs. Note that this algorithm is similar to the barrier formulation for the SDP posed in Askari et al. (2020), with only minor changes to account for a different optimization function.

In particular, Askari et al. (2020) observed that if we let $D = 2\Sigma - S$, we can write

$$\log \det(D) = \log(D_{j,j} - D_{-,j}^\top D_{-,j}^{-1} D_{-,j}) + \log \det(D_{-,j}), \quad (73)$$

where only the term on the left depends on $S_{j,j}$. We can therefore write

$$\log \det(G_S) = \log \det(2\Sigma - S) + \log \det(S) = \log(S_{j,j}) + \log(2\Sigma_{j,j} - S_{j,j} - D_{-,j}^\top D_{-,j}^{-1} D_{-,j}) + c, \quad (74)$$

where c is a constant not depending on $S_{j,j}$. It also may be helpful to note that $D_{-,j}$ and $D_{-,j}^{-1}$ do not depend on $S_{j,j}$. Taking the derivative of equation (74) with respect to $S_{j,j}$, we find that we should update

$$S_{j,j} = \frac{2\Sigma_{j,j} - D_{-,j}^\top D_{-,j}^{-1} D_{-,j}}{2}. \quad (75)$$

To efficiently calculate $c_m = D_{-,j}^\top D_{-,j}^{-1} D_{-,j}$, Askari et al. (2020) observed that $c_m = \|v_m\|_2^2$ where v_m is the solution to the equation $Lv_m = u$ for $u \in \mathbb{R}^p$ where $u_i = 2\Sigma_{i,j}$ if $i \neq j$ and $u_j = 0$ otherwise. Since L is triangular and we can update the Cholesky decomposition $2\Sigma - S = LL^\top$ using rank-1 updates, computing c_m takes only $O(p^2)$. This yields Algorithm 2.

Algorithm 2 Stable Coordinate Descent for ME Knockoffs

- 1: Initialize $S = \lambda_{\min}(\Sigma) \cdot I_p$
 - 2: Solve for L such that $LL^\top = 2\Sigma - S$
 - 3: **for** $\ell = 1, 2, \dots, n_{\text{iter}}$ **do**
 - 4: **for** $j = 1, 2, \dots, p$ **do**
 - 5: Compute c_m using L
 - 6: Solve $S_{j,j}$ as the solution to the optimality condition (75)
 - 7: Compute the rank-1 update to L corresponding to the new value of $S_{j,j}^*$
-

D.4 Solutions for equicorrelated Gaussian designs

In this section, we prove that when Σ is equicorrelated, the same solution, $(1-\rho)I_p$, is asymptotically optimal for both L_{MVR} and L_{ME} . Our result is asymptotic in p , but Figure 13 demonstrates via simulations that $\|S_{\text{MVR}} - S_{\text{ME}}\|_\infty$ quickly converges to 0 for finite p . As we mentioned in Section 3.2, this means that we should expect ME knockoffs to achieve approximately the same optimality guarantees as MVR knockoffs for Gaussian equicorrelated features.

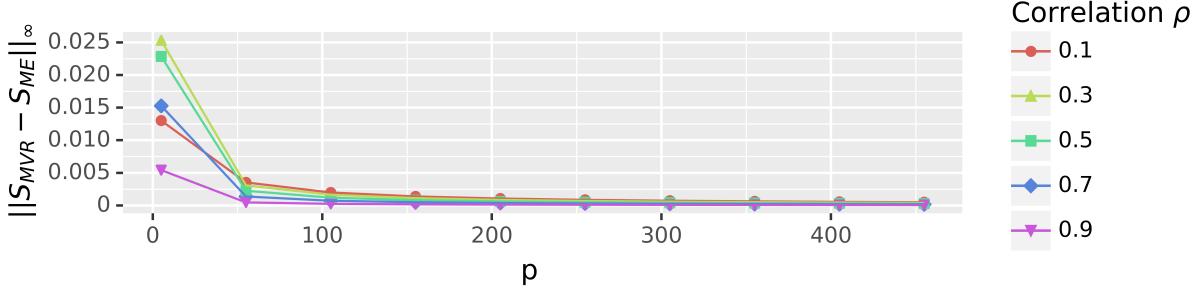


Figure 13: MVR and ME solutions for exchangeable features: we plot $\|S_{\text{MVR}} - S_{\text{ME}}\|_\infty$ for equicorrelated Σ . We vary the correlation ρ and the dimension p and observe that the norm converges towards zero quite quickly.

Theorem D.2. Let $X \sim \mathcal{N}(0, \Sigma)$ for equicorrelated Σ with correlation ρ . Then $S^* = (1 - \rho)I_p$ is asymptotically optimal for both the MVR and ME optimization problems. In particular,

$$\lim_{p \rightarrow \infty} L_{\text{MVR}}(S_{\text{MVR}}) - L_{\text{MVR}}(S^*) = \lim_{p \rightarrow \infty} L_{\text{ME}}(S_{\text{ME}}) - L_{\text{ME}}(S^*) = 0.$$

This result is nontrivial in the sense that the limit of each individual objective as $p \rightarrow \infty$ is ∞ .

Proof. We begin by showing that S^* is asymptotically optimal for the ME problem. Recall that the eigenvalues of G_{S^*} are those of S^* and $2\Sigma - S^*$. Since Σ is a rank-one update to the identity (see Lemma B.1), we can check that G_{S^*} has two eigenvalues: $1 - \rho$ with multiplicity $2p - 1$ and $\lambda_{\max}(G_{S^*}) = 2p\rho + 1 - \rho$ with multiplicity 1. This implies $\log \det(G_{S^*}^{-1}) = -(2p - 1) \log(1 - \rho) - \log(\lambda_{\max}(G_{S^*}))$.

The primal ME problem is to maximize $\log \det G_S$ subject to $G_S \succcurlyeq 0$, or equivalently to minimize $\log \det G_S^{-1}$ subject to the same constraint. The dual of this problem (see Vandenberghe et al. (1998)) is written below for a variable $R \in \mathbb{R}^{2p \times 2p}$:

$$\text{maximize } \mathcal{O}_{\text{ME}}(R) \equiv \log \det(R) - \text{Tr}(G_0 R) + 2p \quad (76)$$

$$\text{such that } R \succcurlyeq 0 \quad (77)$$

$$R_{j,j+p} = R_{j+p,j} = 0 \quad \forall j \in [p],$$

where G_0 is the G_S matrix which corresponds to setting $S = 0 \cdot I_p$. To show S^* is asymptotically optimal for the ME problem, it suffices to show that we can find a sequence of R^* such that the dual gap between R^* and S^* vanishes, i.e., $\lim_{p \rightarrow \infty} \log \det G_{S^*}^{-1} - \mathcal{O}_{\text{ME}}(R^*) = 0$. In particular, let $c_d = \frac{1}{1-\rho}$ and $c_r = \frac{1}{2p-2} \cdot \left(\frac{1}{\lambda_{\max}(G_{S^*})} - \frac{1}{1-\rho} \right)$. Then set

$$D = c_r \mathbf{1}\mathbf{1}^\top + (c_d - c_r)I_p \text{ and } R^* = \begin{bmatrix} D & D - c_d I_p \\ D - c_d I_p & D \end{bmatrix}$$

Intuitively, R^* is the $2p \times 2p$ matrix which takes the value c_d on its diagonal, 0 on the diagonals of its off-diagonal blocks, and c_r everywhere else. We first observe that

$$\text{Tr}(G_0 R^*) = \sum_{j=1}^{2p} \sum_{k=1}^{2p} R_{jk}^* (G_0)_{jk} = 2p \cdot c_d + (4p^2 - 4p)\rho c_r.$$

The first term in the last equality corresponds to the $2p$ diagonal elements of G_0 (which equal 1) and those of R^* (which equal c_d). The second term corresponds to the other $4p^2 - 4p$ nonzero elements of R^* , which equal c_r , and the corresponding elements of G_0 , which equal ρ . Simplifying yields that

$$\text{Tr}(G_0 R^*) = \frac{2p}{1-\rho} + 2p\rho \left(\frac{1}{2p\rho + 1 - \rho} - \frac{1}{1 - \rho} \right) = 2p - \frac{2p\rho}{2p\rho + 1 - \rho}.$$

This implies that $\lim_{p \rightarrow \infty} -\text{Tr}(G_0 R^*) + 2p = 1$. As a result, it suffices to show that $\lim_{p \rightarrow \infty} \log \det(G_{S^*}^{-1}) - \log \det(R^*) = -1$. Note the eigenvalues of R^* are those of $2D - c_d I_p$ and those of $c_d I_p$. Applying rank one theory, this yields eigenvalues of $c_d = \frac{1}{1-\rho}$ with multiplicity p , $(2p-2)c_r + \frac{1}{1-\rho} = \lambda_{\max}(G_{S^*})$ with multiplicity 1, and $c_d - 2c_r$ with multiplicity $p-1$. Note that the first $p+1$ of these eigenvalues agree with $p+1$ of the eigenvalues of G_{S^*} , as previously calculated. This yields

$$\begin{aligned}\log \det(G_{S^*}^{-1}) - \log \det(R^*) &= (p-1) \left[\log \left(\frac{1}{1-\rho} \right) - \log \left(\frac{1}{1-\rho} - \frac{1}{p-1} \left(\frac{1}{2p\rho+1-\rho} - \frac{1}{1-\rho} \right) \right) \right] \\ &= (p-1) \left[\log \left(\frac{1}{1-\rho} \right) - \log \left(\frac{1}{1-\rho} \left(1 + \frac{1}{p-1} - \frac{1-\rho}{(p-1)(2p\rho+1-\rho)} \right) \right) \right] \\ &= -(p-1) \log \left(1 + \frac{1}{p-1} - \frac{1-\rho}{(p-1)(2p\rho+1-\rho)} \right).\end{aligned}\quad (78)$$

To show (78) converges to -1 , it suffices to show the following Lemma, where we take $x = p-1$.

Lemma D.3. *Let $a, b \in \mathbb{R}$ for $a \neq 0$. Then*

$$\lim_{x \rightarrow \infty} x \log \left(1 + \frac{1}{x} + \frac{1}{ax^2 + bx} \right) = 1$$

Proof. By L'Hopital's rule,

$$\begin{aligned}\lim_{x \rightarrow \infty} \frac{\log(1 + 1/x + 1/(ax^2 + bx))}{1/x} &= \lim_{x \rightarrow \infty} \frac{-x^2}{1 + 1/x + 1/(ax^2 + bx)} \cdot \left(-\frac{1}{x^2} - \frac{2ax + b}{(ax^2 + bx)^2} \right) \\ &= \lim_{x \rightarrow \infty} \frac{1}{1 + 1/x + 1/(ax^2 + bx)} \cdot \left(1 + \frac{x^2(2ax + b)}{(ax^2 + bx)^2} \right)\end{aligned}\quad (79)$$

and (79) converges to 1, as the left term in the product converges to 1, and $\frac{x^2(2ax+b)}{(ax^2+bx)^2} \rightarrow 0$ since the numerator has degree 3 but the denominator has degree 4. \square

Next, we show the same asymptotic optimality of S^* for the MVR problem. The formulation of the MVR problem in Lemma D.1 is a standard SDP, which admits the following dual. Let $R_1, R_{21}, R_{22}, R_{23}, R_{24}, R_{41}, R_{42}, R_{44} \in \mathbb{R}^{p \times p}$. Then the dual is

$$\begin{aligned}&\text{maximize } \mathcal{O}_{\text{MVR}}(R) \equiv -2\text{Tr}(R_{21}) - 2\text{Tr}(R_{24}) - 2\text{Tr}(\Sigma R_{44}) \\ &\text{such that } (R_{41})_{j,j} = (R_{44})_{j,j} \forall j \in [p]\end{aligned}\quad (80)$$

$$R \equiv \begin{bmatrix} I_p & R_1 & R_{21} & R_{22} \\ R_1 & I_p & R_{23} & R_{24} \\ R_{21} & R_{23} & R_{41} & R_{42} \\ R_{22} & R_{24} & R_{42} & R_{44} \end{bmatrix} \succcurlyeq 0.$$

As before, we will find a sequence of dual feasible R^* such that $\lim_{p \rightarrow \infty} L_{\text{MVR}}(S^*) - \mathcal{O}_{\text{MVR}}(R^*) = 0$. In particular, we will pick R^* of the form

$$R^* = \begin{bmatrix} I_p & 0 & bI_p & 0 \\ 0 & I_p & 0 & aI_p - \frac{a}{p}\mathbf{1}\mathbf{1}^\top \\ bI_p & 0 & \left(c - \frac{c}{p}\right)I_p & 0 \\ 0 & aI_p - \frac{a}{p}\mathbf{1}\mathbf{1}^\top & 0 & cI_p - \frac{c}{p}\mathbf{1}\mathbf{1}^\top \end{bmatrix}\quad (81)$$

for the constants a, b, c , defined below. Note we define a, b, c in terms of each other and in terms of the function $f(p) = (\sqrt{2p(2p-1)} - \sqrt{2p(2p-3)+2})^2$ for convenience of notation. For now, it is best to ignore $f(p)$ —intuitively, $f(p)$ scales the constants a, b, c by a number that converges to 1 from above, which will help ensure that R^* is feasible. Now, let

$$c = f(p) \left(\frac{2p-1}{2p} \right) \left(\frac{1}{1-\rho} \right)^2 \left(\frac{p}{p-1} \right),$$

$$b = -\sqrt{c - \frac{c}{p}},$$

$$a = -\frac{p}{p-1} \left(\frac{2p-1}{p(1-\rho)} + b \right).$$

For convenience, we will write $R^* = \begin{bmatrix} I_{2p} & R_2^* \\ R_2^* & R_4^* \end{bmatrix}$ where R_2^*, R_4^* can also be represented as block matrices as in (80). Note this construction guarantees that the diagonals of R_{41}^* and R_{44}^* match. Therefore, we only need to show that $R^* \succcurlyeq 0$ to show it is dual feasible. Using the block formulation of R^* , it suffices to show that (i) $R_4^* \succcurlyeq 0$ and (ii) $R_4^* - R_2^*(R_2^*)^\top \succcurlyeq 0$. To do this, we will repeatedly use the fact that for a matrix of the form $D = d_1 I_p - d_2 \mathbf{1}_p \mathbf{1}_p^\top$ for $d_1, d_2 > 0$, $D \succcurlyeq 0$ if $d_2 \leq d_1/p$. (This is a simple consequence of Sherman–Morrison rank one theory.)

To show (i), note $R_4^* = \begin{bmatrix} R_{41}^* & 0 \\ 0 & R_{44}^* \end{bmatrix}$, so we need only show that $R_{41}^* \succcurlyeq 0$ and $R_{44}^* \succcurlyeq 0$. $R_{41}^* = \left(c - \frac{c}{p}\right) I_p \succcurlyeq 0$ follows because $c - \frac{c}{p} > 0$. Furthermore, $R_{44}^* = cI_p - \frac{c}{p} \mathbf{1} \mathbf{1}^\top \succcurlyeq 0$ by the previous consequence of Sherman–Morrison theory. To show (ii), note

$$R_4^* - R_2^*(R_2^*)^\top = \begin{bmatrix} \left(c - \frac{c}{p} - b^2\right) I_p & 0 \\ 0 & cI_p - \frac{c}{p} \mathbf{1} \mathbf{1}^\top - \left(aI_p - \frac{a}{p} \mathbf{1} \mathbf{1}^\top\right) \left(aI_p - \frac{a}{p} \mathbf{1} \mathbf{1}^\top\right)^\top \end{bmatrix}.$$

The top-left block is positive-semidefinite since $c - \frac{c}{p} - b^2 \geq 0$ holds with equality by the definition of b . Simplifying the bottom-right block, we want to show

$$(c - a^2)I_p + \left(-\frac{c}{p} + \frac{2a^2}{p} - p \cdot \frac{a^2}{p^2}\right) \mathbf{1} \mathbf{1}^\top = (c - a^2)I_p - \left(\frac{c - a^2}{p}\right) \mathbf{1} \mathbf{1}^\top \succcurlyeq 0,$$

which holds with equality based on the Sherman–Morrison theory. Lastly, we need to show that the coefficient on I_p is nonnegative, i.e., $c - a^2 \geq 0$. To see this, we plug in the definitions:

$$\begin{aligned} c - a^2 &= f(p) \left(\frac{2p-1}{2p}\right) \left(\frac{1}{1-\rho}\right)^2 \left(\frac{p}{p-1}\right) - \left(\frac{p}{p-1}\right)^2 \left(\frac{2p-1}{p(1-\rho)} - \sqrt{f(p) \left(\frac{2p-1}{2p}\right) \left(\frac{1}{1-\rho}\right)^2}\right)^2 \\ &= \left(\frac{1}{1-\rho}\right)^2 \left(\frac{p}{p-1}\right) \left[f(p) \left(\frac{2p-1}{2p}\right) - \left(\frac{p}{p-1}\right) \left(\frac{2p-1}{p} - \sqrt{f(p) \left(\frac{2p-1}{2p}\right)}\right)^2\right] \\ &= \left(\frac{1}{1-\rho}\right)^2 \left(\frac{p}{p-1}\right) \left[-\left(\frac{f(p)}{p-1}\right) \left(\frac{2p-1}{2p}\right) - \left(\frac{p}{p-1}\right) \left(\frac{2p-1}{p}\right)^2 + 2 \left(\frac{p}{p-1}\right) \left(\frac{2p-1}{p}\right) \sqrt{f(p) \left(\frac{2p-1}{2p}\right)}\right], \end{aligned}$$

where in the last line, we expand the squared term in the brackets and then use the fact that $f(p) \left(\frac{2p-1}{2p}\right) - \left(\frac{p}{p-1}\right) \left(\sqrt{f(p) \left(\frac{2p-1}{2p}\right)}\right)^2 = -\left(\frac{f(p)}{p-1}\right) \left(\frac{2p-1}{2p}\right)$. Continuing, we can factor $\frac{1}{p-1}$ and $\frac{2p-1}{p}$ out of the terms in the brackets:

$$\begin{aligned} c - a^2 &= \frac{2p-1}{(1-\rho)^2(p-1)^2} \left[-\frac{f(p)}{2} - (2p-1) + 2p \sqrt{f(p) \left(\frac{2p-1}{2p}\right)} \right] \\ &= \frac{2p-1}{(1-\rho)^2(p-1)^2} \left[-\frac{f(p)}{2} + \sqrt{f(p)} \sqrt{(2p)(2p-1)} - (2p-1) \right]. \end{aligned}$$

At this point, we will show that the term in the brackets equals zero. This holds because that expression is a quadratic function of $\sqrt{f(p)}$, with roots at

$$\frac{-\sqrt{2p(2p-1)} \pm \sqrt{2p(2p-1) - 4 \cdot -\frac{1}{2} \cdot -(2p-1)}}{2 \cdot -\frac{1}{2}} = \sqrt{2p(2p-1)} \pm \sqrt{2p(2p-3)+2}.$$

Since by definition, $\sqrt{f(p)} = \sqrt{2p(2p-1)} - \sqrt{2p(2p-3)+2}$, we have that $c - a^2 \geq 0$ with equality.

Now that we know R^* is dual feasible, we need to show the difference between $O_{\text{MVR}}(R^*)$ and $L_{\text{MVR}}(S^*)$ vanishes asymptotically. It may be helpful to note at this point that by construction, $c - \frac{c}{p} = f(p) \cdot \frac{2p-1}{2p(1-\rho)^2}$. Now, we can easily check that

$$\text{Tr}(\Sigma R_{44}^*) = \sum_{j,k \in [p]} \Sigma_{j,k}(R_{44}^*)_{j,k} = p \left(c - \frac{c}{p} \right) - (p^2 - p)\rho \frac{c}{p} = f(p) \left[\frac{2p-1}{2(1-\rho)^2} - \rho \left(\frac{2p-1}{2(1-\rho)^2} \right) \right] = f(p) \cdot \frac{2p-1}{2(1-\rho)},$$

where as usual, in the second equality, the first term corresponds to the p diagonal elements of Σ and R_{44}^* , and the second term corresponds to the $p^2 - p$ off-diagonal elements. Finally, we note that by the definition of a , we have that $a - \frac{a}{p} + b = -\frac{2p-1}{p(1-\rho)}$. This yields that

$$\text{Tr}(R_{21}^*) + \text{Tr}(R_{24}^*) = p \left(b + a - \frac{a}{p} \right) = -\frac{2p-1}{1-\rho}.$$

Therefore

$$O_{\text{MVR}}(R^*) = -2\text{Tr}(R_{21}^*) - 2\text{Tr}(R_{24}^*) - 2\text{Tr}(\Sigma R_{44}^*) = 2 \cdot \frac{2p-1}{1-\rho} - f(p) \cdot \frac{2p-1}{1-\rho} = \frac{2p-1}{1-\rho} + (1-f(p)) \frac{2p-1}{1-\rho}.$$

Intuitively, since $f(p)$ converges to 1 extremely quickly, $O_{\text{MVR}}(R^*) \approx \frac{2p-1}{1-\rho}$. Recall from previous computations of the eigenvalues of G_{S^*} that $L_{\text{MVR}}(S^*) = \frac{2p-1}{1-\rho} + \frac{1}{2p\rho+1-\rho}$. As a result,

$$L_{\text{MVR}}(S^*) - O_{\text{MVR}}(S^*) = (1-f(p)) \frac{2p-1}{1-\rho} + \frac{1}{2p\rho+1-\rho},$$

where the last term clearly vanishes as $p \rightarrow \infty$. Therefore, to show $L_{\text{MVR}}(S^*) - O_{\text{MVR}}(S^*)$ vanishes, it suffices to show that $(1-f(p)) \cdot p$ vanishes. In particular, we can apply the definition of $f(p)$ and do some algebra to find that

$$\begin{aligned} (1-f(p))p &= p - p \left(\sqrt{2p(2p-1)} - \sqrt{2p(2p-3)+2} \right)^2 \\ &= p - p \left((2p)(2p-1) - 2\sqrt{(2p)(2p-1)(2p(2p-3)+2)} + 2p(2p-3) + 2 \right) \\ &= p - p \left(8p^2 - 8p + 2 - 2\sqrt{4p(p-1)(2p-1)^2} \right) \\ &= p - p \left(8p^2 - 8p + 2 - 4(2p-1)\sqrt{p(p-1)} \right) \\ &= p \left(1 - 8p^2 + 8p - 2 + 8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)} \right) \\ &= -p \left(8p^2 - 8p + 1 - 8p\sqrt{p(p-1)} + 4\sqrt{p(p-1)} \right). \end{aligned}$$

At this point, we multiply and divide by $8p^2 - 8p + 1 + 8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)}$ to rationalize the expression:

$$\begin{aligned} &= \frac{-p \left((8p^2 - 8p + 1)^2 - (8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)})^2 \right)}{8p^2 - 8p + 1 + 8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)}} \\ &= \frac{-p}{8p^2 - 8p + 1 + 8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)}}. \end{aligned}$$

where in the second step, we use the fact that $(8p^2 - 8p + 1)^2 - (8p\sqrt{p(p-1)} - 4\sqrt{p(p-1)})^2 = 1$. Since the denominator of this fraction has higher degree than the numerator, this proves that $\lim_{p \rightarrow \infty} (1-f(p))p = 0$. By the previous analysis, this proves that the dual gap vanishes asymptotically, so S^* is asymptotically optimal for both the MVR and ME losses. \square

D.5 Speedups for structured covariance matrices

While our coordinate descent algorithms are substantially faster than generic semidefinite program solvers, they are still prohibitively expensive when p is very large. One way out of this is to follow the approach of Askari et al. (2020) and approximate Σ using a rank- k factor model, in which case both algorithms will run in $O(n_{\text{iter}}pk^2)$ (see Askari et al. (2020) for details). Alternatively, following the approach of Candès et al. (2018), we note below that the MVR and ME optimization problems can be efficiently parallelized when Σ is approximated as a block-diagonal matrix.

Lemma D.4. *Suppose $\Sigma = \text{blockdiag}(\Sigma_1, \dots, \Sigma_\ell)$. Let S_j^* be the result of the Gaussian MVR optimization problem for Σ_j . Then $S_{\text{MVR}} = \text{blockdiag}(S_1^*, \dots, S_\ell^*)$. The same holds for the ME problem.*

Proof. Showing this result requires showing that the constraints and the objective function in both problems are separable across the blocks. To deal with the constraints, observe that $S = \text{blockdiag}(S_1, \dots, S_\ell) \succcurlyeq 0$ if and only if $S_j \succcurlyeq 0$ for all $1 \leq j \leq \ell$, and similarly $2\Sigma - S \succcurlyeq 0$ if and only if $2\Sigma_j - S_j \succcurlyeq 0$ for all $1 \leq j \leq \ell$. For the MVR loss, note that the inverse of a block-diagonal matrix is the block-diagonal matrix of its blocks' inverses. Therefore

$$L_{\text{MVR}}(S) \propto \text{Tr}((2\Sigma - S)^{-1}) + \text{Tr}(S^{-1}) = \sum_{j=1}^{\ell} \text{Tr}(2\Sigma_j - S_j)^{-1} + \text{Tr}(S_j^{-1}).$$

To see this result for ME knockoffs, note that

$$\log \det(G_S) = \sum_{j=1}^{\ell} \log \det \left(\begin{bmatrix} \Sigma_j & \Sigma_j - S_j \\ \Sigma_j - S_j & \Sigma_j \end{bmatrix} \right).$$

For both problems, the constraints and objective function are separable, which completes the proof. \square

This motivates the AMVR (resp. AME) construction:

Step 1. Approximate Σ as a block-diagonal matrix Σ_{approx} and find S_{approx} as the solution to the MVR (resp. ME) problem for Σ_{approx} .

Step 2. Run a grid search over $[0, 1]$ to find $\gamma = \arg \max_{\gamma \in [0, 1]} L_{\text{MVR}}(\gamma \cdot S_{\text{approx}})$ s.t. $2\Sigma \succcurlyeq \gamma \cdot S_{\text{approx}}$. For AME, replace L_{MVR} with L_{ME} .

Finally, we return $S^* = \gamma \cdot \hat{S}$.

D.6 Runtime Simulations

The state-of-the-art algorithms to compute SDP knockoffs have the same or slower computational complexities than MVR and ME knockoffs (Askari et al., 2020). However, such analysis hides constant factors that may have substantial effects in practice. In this section, we plot the average computation time for MVR, ME, and SDP Gaussian knockoffs, including settings where we use block-diagonal approximations and factor approximations of Σ to speed up computation, as in Section D.5. To compute SDP knockoffs, we use two algorithms: one which directly solves an SDP (Candès et al., 2018) and one which uses coordinate descent (Askari et al., 2020).

Figure 14 shows that the runtimes for each of these algorithms are very similar for the “AR1” covariance matrix from Section 4, with the exception of the coordinate descent SDP method, which takes longer to converge in the “no approximation” setting. Other than this, for each approximation strategy, all average runtimes are roughly within a factor of two of each other, and many runtimes are indistinguishable. Of course, these results should be taken with a grain of salt, since the precise runtime of each algorithm depends on implementation-specific details. Despite this, Figure 14 shows that generating MVR and ME knockoffs should not take substantially longer than generating SDP knockoffs.

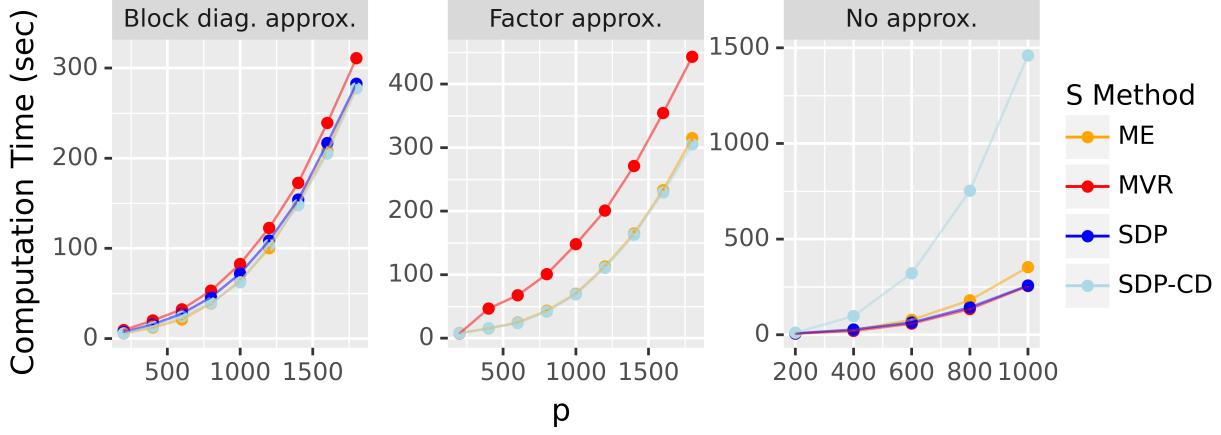


Figure 14: Runtimes for MVR, ME, and SDP Gaussian knockoffs. We generate Σ as in the “AR1” setting from Section 4, with p varied between 200 and 2000. In the left and middle panels, we apply factor and block-diagonal approximations of Σ with 25 factors and a maximum block size of 100, respectively. In the right panel, we generate exact MVR, ME, and SDP knockoffs for p between 200 and 1000. The “SDP” and “SDP-CD” curves correspond to computing SDP knockoffs by directly solving a SDP or using coordinate descent, respectively. Note that the “SDP” method cannot take advantage of factor approximations, so it does not appear in the middle panel.

E ME knockoffs for discrete features

In this section, we describe a formulation for exact ME knockoffs for discrete features, as discussed in Sections 3 and 5. Although this formulation is too computationally expensive to be practical, it may provide a starting point for future work, and furthermore it allows us to prove Lemma 3.2. For simplicity, we will consider the binary case where $[X, \tilde{X}] \in \{0, 1\}^{2p}$, but this analysis generalizes naturally as long as each feature has a finite support.

To begin with, let $v_k = \mathbb{P}([X, \tilde{X}] = \text{binary}(k))$ where $\text{binary}(k) \in \{0, 1\}^{2p}$ is the binary representation of $k \in \{0, \dots, 2^{2p} - 1\}$. To define a valid joint knockoff distribution for $[X, \tilde{X}]$, we need the variables v_k to preserve the marginal distribution of X and to satisfy the pairwise exchangeability condition (1). First, we show how we can ensure that constraint on the marginal distribution of X using linear constraints. For each configuration $\epsilon \in \{0, 1\}^p$, this corresponds to ensuring

$$\sum_{k: \text{binary}(k)_{1:p} = \epsilon} v_k = \mathbb{P}(X = \epsilon). \quad (82)$$

There are 2^p values for ϵ and therefore 2^p such linear constraints. Next, to define notation, we let $\text{binary}(k)_{\text{swap}(j)}$ be equal to $\text{binary}(k)$ except with the digits in locations j and $j + p$ swapped. The pairwise exchangeability condition (1) requires that for any $k, k' \in \{0, \dots, 2^{2p} - 1\}$,

$$v_k = v_{k'} \text{ if } \exists j \in [p] \text{ s.t. } \text{binary}(k)_{\text{swap}(j)} = \text{binary}(k'). \quad (83)$$

Constraint (83) can be viewed as the linear constraint $v_k - v_{k'} = 0$. Alternatively, it may be more efficient to consolidate the variables v_k and $v_{k'}$ into a single variable,³ but to ease notation, we will not do so.

Let $v = (v_1, \dots, v_{2^{2p}})$ be the vector of optimization variables. Additionally, let $b \in \mathbb{R}^{2^p}$ be the vector of probabilities of $X = \text{binary}(k)$ for $k \in \{0, \dots, 2^p - 1\}$. Let m be the number of distinct pairwise exchangeability constraints. Then for a suitable $A \in \mathbb{R}^{2^{2p} \times 2^p}$ corresponding to (82) and a matrix $B \in \mathbb{R}^{2^{2p} \times m}$

³It can be shown that doing this for each pairwise exchangeability constraint leaves 3^p optimization variables overall.

corresponding to the pairwise exchangeability constraints, we can solve

$$\begin{aligned} & \text{maximize} \quad - \sum_{k=0}^{2^p} v_k \log(v_k) \\ & \text{s.t. } Av = b \\ & \quad Bv = 0 \\ & \quad 0 \leq v_k \leq 1 \quad \forall k \in \{1, \dots, 2^p\} \end{aligned} \tag{84}$$

to find the distribution for $[X, \tilde{X}]$ which has the maximal total entropy while maintaining the marginal distribution of X and the pairwise exchangeability condition. Note that in (84) and throughout, we use the convention that $0 \cdot \log(0) = \lim_{x \rightarrow 0^+} x \log(x) = 0$.

This optimization problem is convex and has been well-studied, especially since A is sparse (Persson and Clarke, 1986; Boyd and Vandenberghe, 2004). Unfortunately, the number of variables and constraints grows exponentially in p , making this formulation intractable. One way to resolve this may be to use special structure, such as conditional independence properties of the distribution of X , to simplify the constraints. Furthermore, one might settle for minimizing mutual information over a slightly restricted class of knockoff distributions in order to reduce the number of optimization variables. A last approach might be to carefully analyze concentration of measure results, which indicate that for some constrained maximum entropy problems, almost all feasible solutions approximately maximize the entropy objective (Oikonomou and Grünwald, 2016). We leave more concrete analysis to future work.

This formulation allows us to prove Lemma 3.2. We begin with a technical lemma regarding the finite case. In general, we use the convention that $0 \cdot \log(0) = \lim_{x \rightarrow 0} x \log(x) = 0$.

Lemma E.1. *Let X have finite support \mathcal{X} . Suppose there exists a valid distribution $[X, \tilde{X}]$ with support $A \subset \mathcal{X} \times \mathcal{X}$. Then when \tilde{X} are ME knockoffs, the support of $[X, \tilde{X}]$ contains A .*

Proof. As discussed above, we can represent the ME problem as follows. Let $p(x, \tilde{x})$ represent the probability mass function of $[X, \tilde{X}]$. Then we want to solve

$$\begin{aligned} & \text{maximize } H(p) \equiv - \sum_{x, \tilde{x} \in \mathcal{X} \times \mathcal{X}} p(x, \tilde{x}) \log(p(x, \tilde{x})) \\ & \text{s.t. } \sum_{\tilde{x} \in \mathcal{X}} p(x, \tilde{x}) = p(x) \quad \forall x \in \mathcal{X} \end{aligned} \tag{85}$$

$$p([x, \tilde{x}]_{\text{swap}(j)}) = p([x, \tilde{x}]) \quad \forall j \in [p], \forall [x, \tilde{x}] \in \mathcal{X} \times \mathcal{X}. \tag{86}$$

Suppose a feasible solution p_0 has support A . Suppose for the sake of contradiction that the optimal solution p^* has support A^* where $A \not\subset A^*$. Note that for all $\alpha \in [0, 1]$, $\alpha p^* + (1 - \alpha)p_0$ is a feasible solution to the problem as well, as it obeys (85) and (86). We use this property twice. First, define $p' = 0.5p_0 + 0.5p^*$, which is a feasible solution that has support $A \cup A^*$. We will use p' to construct a contradiction. Second, since all convex combinations $\alpha p^* + (1 - \alpha)p'$ are feasible as well, we note

$$\frac{\partial}{\partial \alpha} H(\alpha p^* + (1 - \alpha)p') = - \sum_{(x, \tilde{x}) \in A \cup A^*} (p^*(x, \tilde{x}) - p'(x, \tilde{x})) \log(\alpha p^*(x, \tilde{x}) + (1 - \alpha)p'(x, \tilde{x})) \tag{87}$$

$$= - \sum_{(x, \tilde{x}) \in A \setminus A^*} -p'(x, \tilde{x}) \log((1 - \alpha)p'(x, \tilde{x})) \tag{88}$$

$$- \sum_{(x, \tilde{x}) \in A^*} (p^*(x, \tilde{x}) - p'(x, \tilde{x})) \log(\alpha p^*(x, \tilde{x}) + (1 - \alpha)p'(x, \tilde{x})), \tag{89}$$

where (88) follows because we assume $p^*(x, \tilde{x}) = 0$ for all $(x, \tilde{x}) \notin A$.

We will show that $\lim_{\alpha \rightarrow 1} \frac{\partial}{\partial \alpha} H(\alpha p^* + (1 - \alpha)p') = -\infty$. To show this, first observe that the top term (88) approaches $-\infty$ as $\alpha \rightarrow 1$. This follows because for $(x, \tilde{x}) \in A^* \setminus A$, $p'(x, \tilde{x}) > 0$ by assumption, and for

each $(x, \tilde{x}) \in A \setminus A^*$, $\lim_{\alpha \rightarrow 1} \log((1 - \alpha)p'(x, \tilde{x})) = -\infty$. Second, note that the limit of the bottom term (89) as $\alpha \rightarrow 1$ is $-\sum_{(x, \tilde{x}) \in A^*} (p^*(x, \tilde{x}) - p'(x, \tilde{x})) \log(p^*(x, \tilde{x}))$, which is finite because it is a finite sum of finite elements. Thus, $\lim_{\alpha \rightarrow 1} \frac{\partial}{\partial \alpha} H(\alpha p^* + (1 - \alpha)p') = -\infty$.

This means that for all α sufficiently close to 1, $\frac{\partial}{\partial \alpha} H(\alpha p^* + (1 - \alpha)p')$ is negative. Since H is continuous, this implies that for some $\alpha^* \in [0, 1)$, $H(\alpha^* p^* + (1 - \alpha^*)p') > H(p^*)$. This is a contradiction, since we assumed p^* is the optimal solution. Therefore $A \subset A^*$. \square

Lemma 3.2. *Let X have finite support \mathcal{X} . For any $j \in [p]$ and any $x, \tilde{x} \in \mathcal{X} \times \mathcal{X}$, if there exists a valid distribution $[X, \tilde{X}]$ such that $\text{Var}(X_j | X_{-j} = x_{-j}, \tilde{X} = \tilde{x}) > 0$, then ME knockoffs \tilde{X} satisfy this property as well.*

Proof. Without loss of generality let $j = 1$. Suppose p_0 is a valid joint feature-knockoff distribution such that under p_0 , $\text{Var}(X_1 | X_{-1} = x_{-1}, \tilde{X} = \tilde{x}) > 0$. This implies that there exist at least two values $x_1^{(1)}, x_1^{(2)}$ such that $p_0((x_1^{(1)}, x_{-1}, \tilde{x}), p_0((x_1^{(2)}, x_{-1}, \tilde{x})) > 0$. By Lemma E.1, if p^* is the joint PMF of the features and ME knockoffs, then $p^*(x_1^{(1)}, x_{-1}, \tilde{x}), p^*(x_1^{(2)}, x_{-1}, \tilde{x}) > 0$ as well. This implies that for the ME knockoffs \tilde{X} , $\text{Var}(X_1 | X_{-1} = x_{-1}, \tilde{X} = \tilde{x}) > 0$. \square

F Further simulation results

F.1 Further simulations for equicorrelated designs

In this section, we present Figure 15, which demonstrates that S_{MVR} and S_{ME} are approximately optimal over all S -matrices of the form $\gamma \cdot 2\lambda_{\min}(\Sigma)I_p$ when X is equicorrelated. Note that the “maximum” S -matrix corresponding to $\gamma = 1$ is not always equal to S_{SDP} , because the SDP will never set any value of $\text{diag}(S)$ to greater than 1. Indeed, Lemma B.1 tells us S_{SDP} corresponds to $\gamma = \frac{1}{2-2\rho}$ when $\rho < 0.5$ and $\gamma = 1$ when $\rho \geq 0.5$.

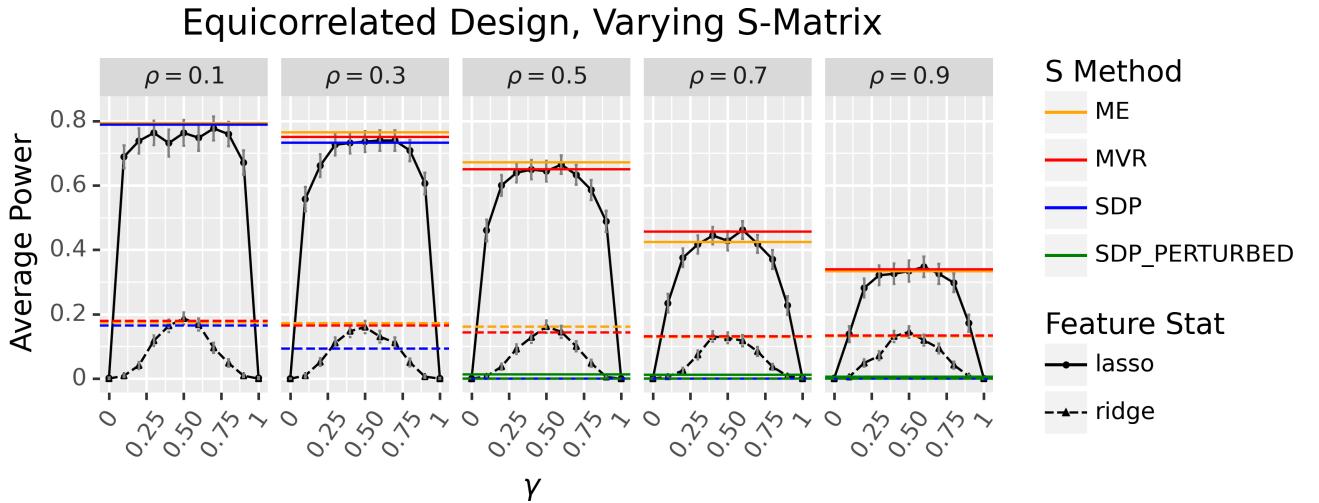


Figure 15: Empirical powers when X is an exchangeable Gaussian design and $Y | X \sim \mathcal{N}(X\beta, 1)$. The number of non-nulls is 50 with coefficients sampled independently from $\text{Unif}([-δ, -δ/2] \cup [\delta/2, δ])$. We let $δ = 2$ for $\rho = 0.9$ and $δ = 1$ otherwise. We use lasso and ridge coefficient differences and compare the performance of different S -matrices, as defined in equation (13). Here, $n = 250$ and $p = 500$. All horizontal lines have standard errors less than 1.8%.

F.2 Comparison to conditional independence knockoffs

In this section, we compare the MRC framework to the CI knockoffs introduced by Liu and Rigollet (2019) and further analyzed by Ke et al. (2020). We make two observations about the CI knockoff framework.

First, even when CI knockoffs are well-defined, the conditional independence condition $X_j \perp\!\!\!\perp \tilde{X}_j \mid X_{-j}$ does not tell us whether X_j is reconstructable from the joint information in X_{-j}, \tilde{X} . To illustrate this, suppose $X \sim \mathcal{N}(0, \Sigma)$ where Σ is block-equicorrelated with a block-size of $\ell = 2$ and correlation $\rho \geq 0.5$, which is one of the correlation structures analyzed by Ke et al. (2020). In this case, as ρ approaches 1, the conditional independence approach yields almost exactly the same answer as the MAC-minimizing approach. In particular, $S_{SDP} = (2 - 2\rho)I_p$, $S_{CI} = (1 - \rho^2)I_p$, and $S_{MVR} \approx S_{ME} \approx (1 - \rho)I_p$, so if we set $\rho = 0.9$ as an example, then $S_{SDP} = 0.2 \cdot I_p$, $S_{CI} = 0.19 \cdot I_p$, and $S_{MVR} \approx S_{ME} \approx 0.1 \cdot I_p$. As a result, CI knockoffs become less powerful than MRC knockoffs as ρ increases, as demonstrated by the left panel of Figure 16.

Second, as noted by Liu and Rigollet (2019), CI knockoffs are only well-defined for a fairly restrictive class of Gaussian designs. Recently, Ke et al. (2020) proposed the following extension to the case where $X \sim \mathcal{N}(0, \Sigma)$ for general Σ . They suggest computing $S_{CI} = (\text{diag}(\Sigma^{-1}))^{-1}$ naively and then performing a binary search to find the maximum $\gamma \in [0, 1]$ such that $\gamma \cdot S_{CI}$ satisfies $0 \leq \gamma \cdot S_{CI} \leq 2\Sigma$. Unfortunately, the γ produced by this binary search may be quite small, limiting the power of these generalized CI knockoffs. For example, consider an ErdosRenyi covariance matrix where 99% of the entries of Σ equal zero. In this example, as illustrated by Figure 17, the binary search sets most of the diagonal elements of $\gamma \cdot S_{CI}$ to approximately equal 0.1, whereas S_{MVR} chooses a variety of values ranging between 0 and 1. As shown in the right panel of Figure 16, this substantially increases the power of the MRC methods relative to CI knockoffs.

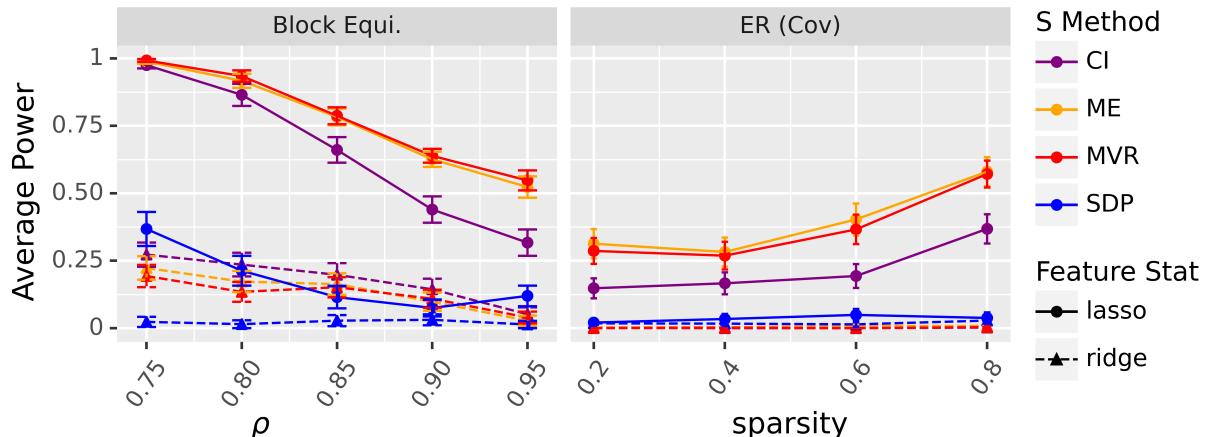


Figure 16: We let $X \sim \mathcal{N}(0, \Sigma)$ and $Y \mid X \sim \mathcal{N}(X\beta, 1)$. In both cases, $p = 500$, the number of non-nulls is 50. On the left panel, the non-nulls are sampled as independent symmetric random signs, and the covariance matrix is block-equicorrelated with block-size $\ell = 2$ and correlation ρ . We set $n = 333$ and we control the FDR at level $q = 0.05$. On the right panel, the non-nulls are sampled independently from $\text{Unif}([1, -1/2] \cup [1/2, 1])$ we use an ErdosRenyi covariance matrix (see Appendix F.4), and we vary the sparsity of the covariance matrix between 20% and 80%, with $n = 375$ and $q = 0.1$.

F.3 Examples from the literature

In this section, we discuss two examples in the literature where SDP knockoffs fail to have any power. First, Xing et al. (2019) ran simulations in the setting where X is Gaussian and equicorrelated, and $Y \mid X \sim \mathcal{N}(X\beta, 1)$. They let $n = 1000$, $p = 300$, and vary ρ between 0 and 0.8. The linear coefficients β have 60 non-nulls sampled independently from $\mathcal{N}(0, 20/\sqrt{n})$. They found that the performance of model-X knockoffs fell to 0 when $\rho \geq 0.5$, including when they estimated the covariance matrix using the data. They implemented a ‘‘model-X-fix’’ method, which involves projecting and slightly perturbing the design matrix. However, the

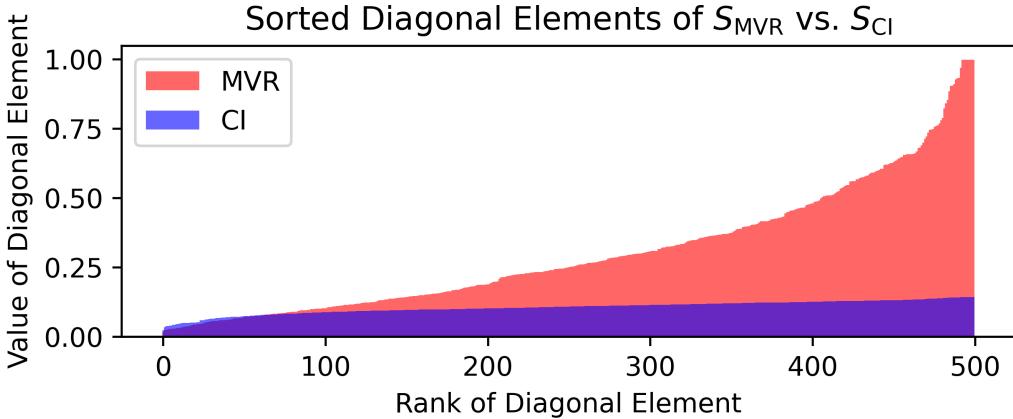


Figure 17: We let $p = 500$ and sample one ErdosRenyi covariance matrix (see Appendix F.4) which is 99% sparse and compute S_{MVR} and S_{CI} using the binary search method from Ke et al. (2020). We plot the sorted diagonal elements of both methods to illustrate that S_{CI} has very small diagonal elements compared to S_{MVR} , explaining the power loss in Figure 16.

power of the model-X-fix method still drops off substantially as ρ grows larger.

We rerun these experiments using MVR knockoffs instead of SDP knockoffs. The MVR knockoffs outperform both the SDP knockoffs as well as the “model-X-fix” method by large margins, as much as 60 and 25 percentage points, respectively. Although the Gaussian mirror method slightly outperforms knockoffs in this setting, the performance gap is fairly small (on the order of 2 – 5%, depending on the value of ρ). These results are depicted in the left panel Figure 18.

Second, Dai et al. (2020) ran simulations with a Gaussian equicorrelated design with a constant pairwise correlation of $\rho \in [0, 0.8]$. They set the number of data points $n = 500$ and vary the dimensionality $p \in [500, 2000]$. For the response, they let $Y | X \sim \mathcal{N}(X\beta, 1)$ where β has 50 non-nulls with a signal size of $\mathcal{N}(0, 10\sqrt{\log(p)/n})$. They use lasso coefficient differences as feature statistics. In the high-dimensional case where $p = 2000$, the SDP MX-knockoffs have almost zero power. In the low-dimensional case where $p = 500$, MX-knockoffs have very low but nonzero power.⁴ In both cases, MX-knockoffs have the lowest power compared to every competitor.

We rerun these experiments in exactly the same setting, except we generate MVR knockoffs instead of SDP knockoffs. As a sanity check, we note that the power of MVR knockoffs agrees with the original paper when $\rho = 0$, since in this simple case, $S_{\text{MVR}} = S_{\text{SDP}} = I_p$. On the other hand, when $\rho = 0.8$, we see that MVR knockoffs outperform every competitor by a wide margin—they are so powerful they do not even fit on the initial charts from the paper. These results are presented in the right panel of Figure 18 for the high-dimensional case where $\rho = 0.8$. We present the full results for various ρ and p in Figure 19.

F.4 Simulation details for sections 4.2, 4.3, 4.4

In this section, we provide additional details on the covariance matrices Σ in Sections 4.2 through 4.4 as well as the sparse nonlinear models in Figure 10. The covariance matrices are defined below. In all cases, we rescale each covariance matrix Σ such that $\Sigma_{jj} = 1$ for all $j \in [p]$ after construction.

- Equi. or Equicorrelated covariance matrices refer to Σ where $\Sigma_{ij} = \rho$ for $i \neq j$ and 1 otherwise. Throughout, unless specified otherwise, we will set $\rho = 0.5$.
- Block equi. or block equicorrelated covariance matrices refer to Σ where Σ is block-diagonal with p/ℓ

⁴This result may seem to contradict Theorem 2.4, which states that SDP knockoffs should have zero power asymptotically since the feature-knockoff covariance matrix should have rank $p + 1$. However, in practice, many statistical packages attempt to prevent *exact* low rank structure, which likely accounts for the low but nonzero power of SDP knockoffs.

Figure 4b, Xing et. al 2019

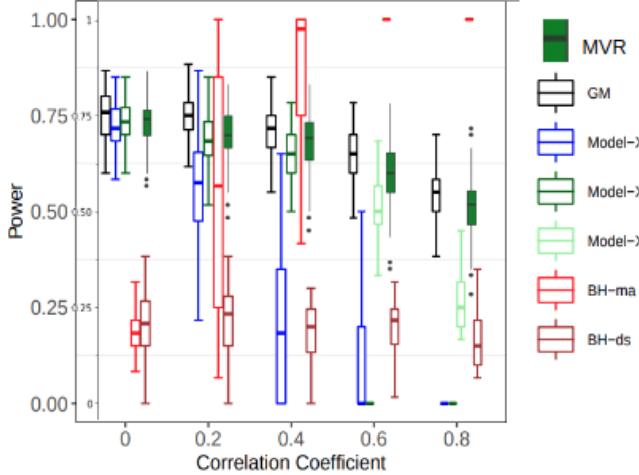


Figure 4, Dai et. al 2020

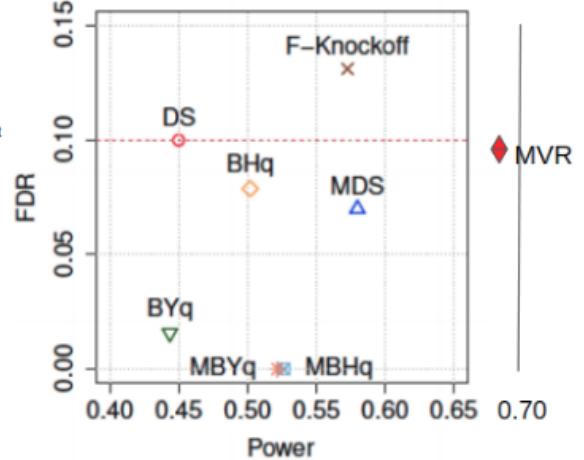


Figure 18: Left: The replication of Figure 4b from Xing et al. (2019) with the addition of MVR knockoffs in solid dark green. Right: The replication of Figure 4 from Dai et al. (2020) with the addition of MVR knockoffs as a solid red diamond. The standard errors for the MVR dot are less than 0.005 for power and 0.0025 for FDR.

$\ell \times \ell$ equicorrelated blocks with correlation ρ inside the blocks. We set $\ell = 5$ and $\rho = 0.5$ unless specified otherwise. In this setting, signals are clustered according to the blocks, meaning that all features in a block are either null or non-null. This simulation setting, including the clustering of the signals, follows Dai and Barber (2016).

- ER (ErdosRenyi) matrices are constructed as follows. We generate a random upper-triangular matrix V where for $i > j$,

$$V_{ij} \stackrel{i.i.d.}{\sim} \pm \delta_{ij} \cdot U_{ij}$$

where $\delta_{ij} \stackrel{i.i.d.}{\sim} \text{Bern}(0.2)$ and $U_{ij} \stackrel{i.i.d.}{\sim} \text{Unif}(0.1, 1)$. In the ER (Cov) setting, we set

$$\Sigma = (V + V^\top) + (0.1 + \lambda_{\min}(V^\top + V)) \cdot I_p$$

In the ER (Prec) setting, we set the precision matrix Σ^{-1} equal to the same quantity. In both cases, we rescale Σ to be a correlation matrix. This simulation set-up roughly follows Li and Maathuis (2019).

- Finally, we generate AR1 covariance matrices where we can represent $X_1 \sim \mathcal{N}(0, 1)$, and for $j \in \{2, \dots, p\}$,

$$X_j = \rho_j X_{j-1} + \sqrt{1 - \rho_j^2} Z_j$$

for $Z_j \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Throughout, we sample $\rho_j \stackrel{i.i.d.}{\sim} \text{Beta}(3, 1)$ to simulate a challenging setting where features are highly correlated. After sampling the correlations, if $\lambda_{\min}(\Sigma) < 0.001$, we add $(0.001 - \lambda_{\min}(\Sigma))I_p$ to increase the eigenvalues above a numerical tolerance and then rescale to a correlation matrix. Since non-nulls are often clustered together in genetic studies, in the AR1 (Corr) setting, we partially capture this idea by letting the non-nulls lie along a single continuous block of features.

Note in Section 4.3, we use the package sklearn for implementations of the graphical lasso and Ledoit–Wolf covariance estimation methods (Pedregosa et al., 2011).

For the nonlinear simulations in Figure 5, we run simulations involving the five following conditional means:

- cos: $\mu(X) = \cos(X)\beta$, where cos denotes the elementwise cosine operation.

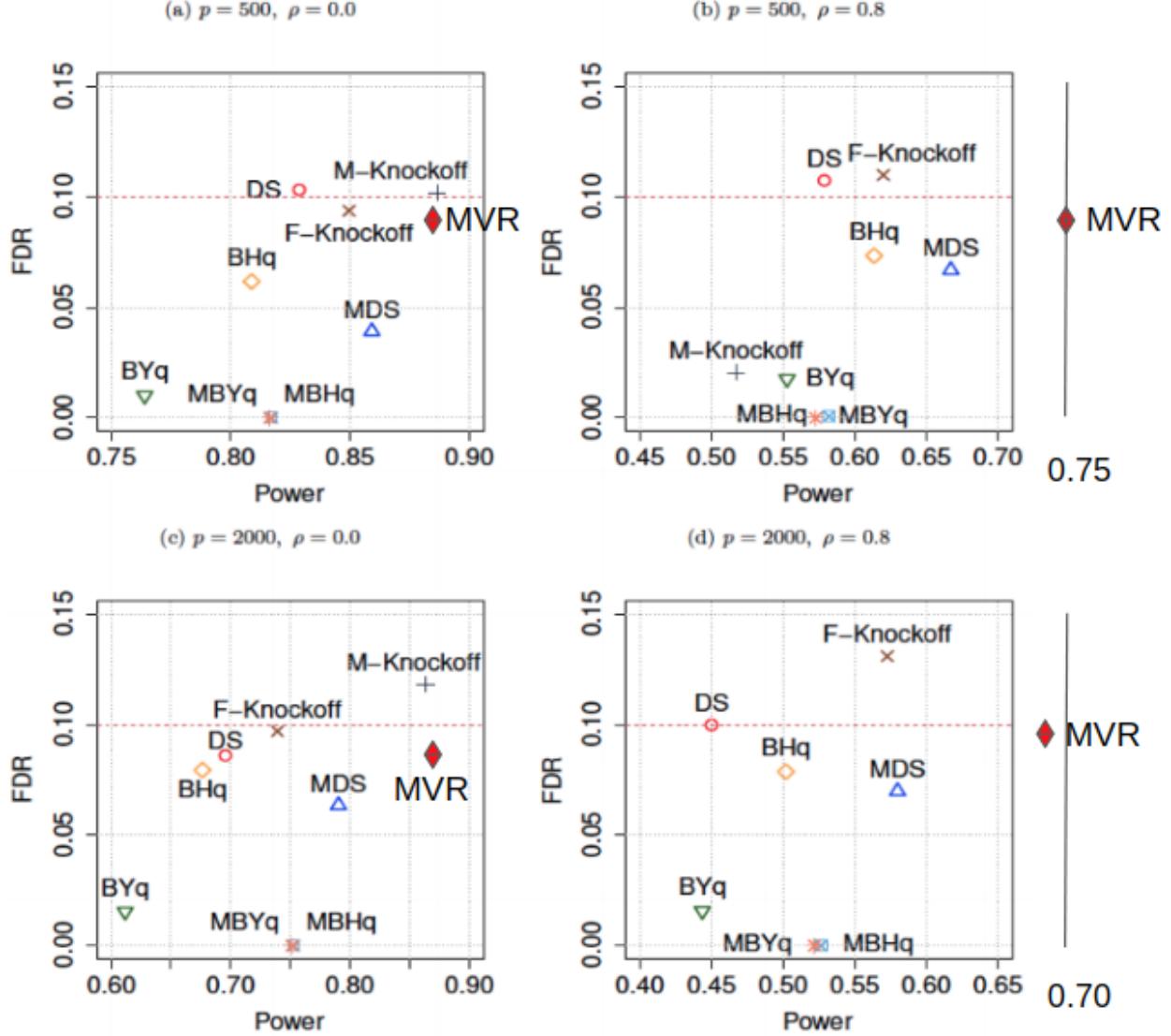


Figure 19: The replication of Figure 4 from Dai et al. (2020) with the addition of MVR knockoffs as solid red diamonds. All standard errors are less than 0.005 for power and 0.0025 for FDR.

- cubic: $\mu(X) = (X^3)\beta - X\beta$ where X^3 denotes the elementwise cubing operation.
- pairint: $\mu(X) = \sum_{i,j \in [p]} \beta_{i,j} X_i X_j$.
- quadratic: $\mu(X) = (X^2)\beta$.
- trunclinear: $\mu(X) = \sum_{j=1}^p \text{sign}(\beta_i) \mathbb{I}(X_j \beta_j > 0)$.

In all cases, we sample nonzero coefficients from ± 5 with equal probability. In all settings except the pairint setting, we choose 30 non-nulls uniformly at random. In the pairint setting, we select 30 non-null features uniformly at random and sequentially group them into disjoint pairs, from left to right. For example, if features 1, 13, 52, 61 were the first four non-nulls, then features 1 and 13 would have an interaction and features 52 and 61 would have an interaction. This means each feature may participate in at most one pairwise interaction.

F.5 Further experiments in the AR1 setting

In this section, we vary the correlation in the AR1 setting detailed in Appendix F.4. Previously, we sampled $\text{Cor}(X_j, X_{j+1}) \stackrel{i.i.d.}{\sim} \text{Beta}(a, b)$ for $a = 3$ and $b = 1$. In Figure 20, we present the results for when we vary $a \in \{0.5, 1, 2, 3\}$. As expected, we see that the MRC methods outperform the SDP by higher margins when the correlation is higher. We also present results when the correlation $\text{Cor}(X_j, X_{j+1}) = \rho$ for a constant $\rho \in (0, 1)$. Interestingly, in this case, MRC knockoffs do not outperform their SDP counterparts—as shown in Figure 21, MRC knockoffs seem to very slightly outperform SDP knockoffs when the non-nulls are clustered together, but they seem to slightly under-perform SDP knockoffs when the non-nulls are not clustered together. We conjecture that the key reconstructability condition (5) will hold approximately for some $J \subset [p]$ with high probability when we randomly sample Σ , but when Σ has special structure such as constant between-feature correlations, the condition (5) may not hold.

Of course, when sampling $\text{Cor}(X_j, X_{j+1}) \stackrel{i.i.d.}{\sim} \text{Beta}(3, 1)$, many correlations will be extremely close to 1. One might wonder what the effect of these extremely large correlations is. To analyze this, in Figure 22, we sample pairwise correlations $\text{Cor}(X_j, X_{j+1}) \stackrel{i.i.d.}{\sim} \min(m_c, \text{Beta}(3, 1))$, for $m_c \in \{0.7, 0.8, 0.9\}$. The results show that even when $m_c = 0.7$, lasso-based MRC knockoffs can outperform their SDP counterparts by as much as 20 percentage points. Thus, the advantage of MRC knockoffs persists even without extremely high correlations.

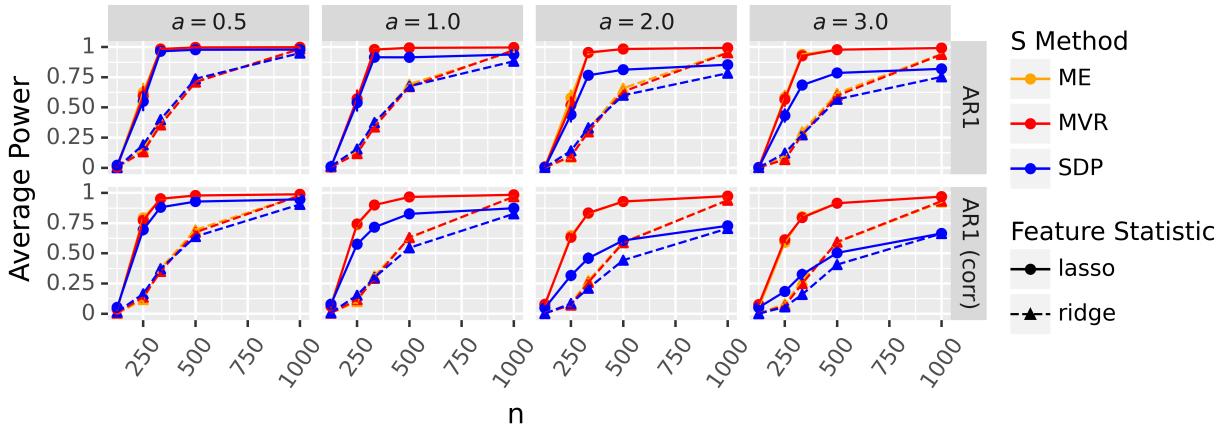


Figure 20: Power for Gaussian AR1 designs with varied correlations. We sample $X \sim \mathcal{N}(0, \Sigma)$ for the AR1 designs as defined in Appendix F.4 where $\text{Cor}(X_j, X_{j+1})$ is sampled independently from $\text{Beta}(a, 1)$ for $a \in \{0.5, 1, 2, 3\}$. We let $Y \sim \mathcal{N}(X\beta, 1)$ with $p = 500$ and 50 non-nulls, where the non-null coefficients are sampled independently from $\text{Unif}([-2, -1] \cup [1, 2])$. We use cross-validated lasso and ridge coefficient differences as feature statistics.

F.6 Further experiments for nonlinear responses

In this section, we present further results where $Y | X$ is nonlinear. In particular, we simulate the power of DeepPINK and LCD feature statistics.

For the DeepPINK statistics, in Figure 23, we again see that MRC knockoffs tend to outperform their SDP counterparts, although the DeepPINK statistic has low power for all methods for some conditional means. We also observe that occasionally, adding *more data* to the DeepPINK feature statistic seems to worsen performance. Although this is strange, this phenomenon has been observed fairly consistently for some deep models in the machine learning literature (Nakkiran et al., 2020). Note that our DeepPINK implementation is very similar to but not identical to that of Lu et al. (2018)—for example, we use a different set of hyperparameters (such as batchsize during training) than the original paper. Please see our code for more details.

For the LCD statistics in Figure 24, we often see large gains for MRC knockoffs over SDP knockoffs for the trunclinear and cubic responses. For the other conditional means, which are highly nonlinear, the LCD

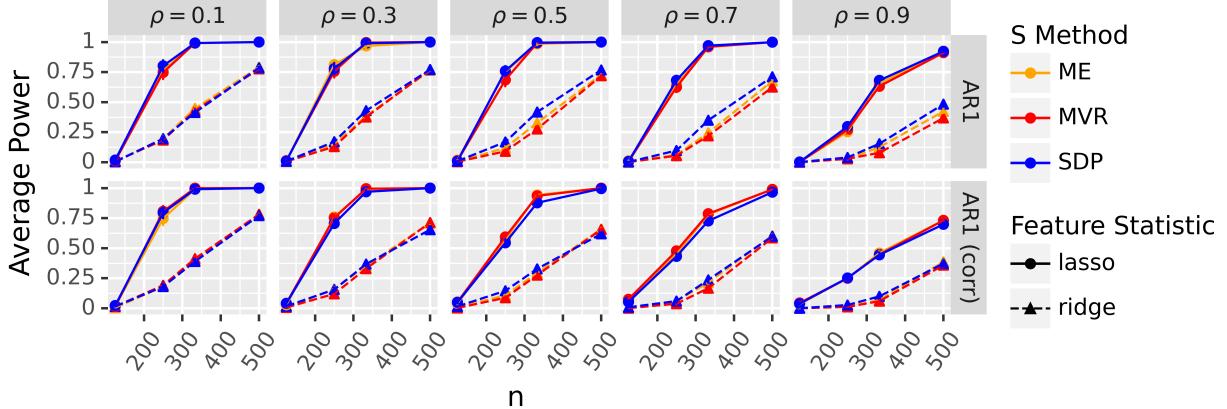


Figure 21: The same setting as Figure 20, except we set $\text{Cor}(X_j, X_{j+1}) = \rho$ for a constant ρ , and we sample the non-null coefficients independently from $\text{Unif}([-1, -0.5], [0.5, 1])$.

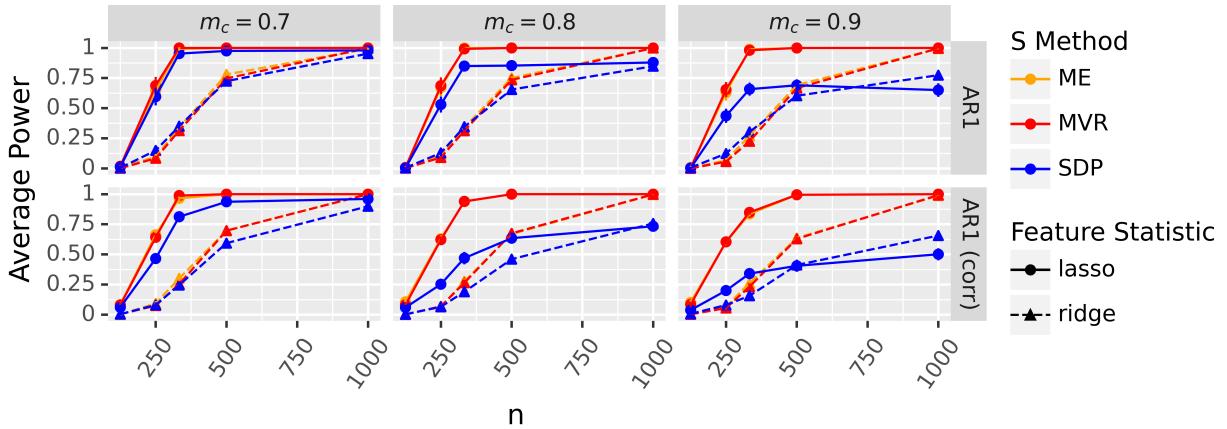


Figure 22: The same setting as Figure 20, except we sample $\text{Cor}(X_j, X_{j+1}) \stackrel{i.i.d.}{\sim} \min(m_c, \text{Beta}(3, 1))$ for a “maximum correlation” value m_c .

coefficient statistics unsurprisingly have zero power for all methods.

F.7 Details for non-Gaussian simulations in Section 4.6

In this section, we present precise details about the three design distributions we simulate in Section 4.6. We also include Figure 25, which is the corresponding FDR plot for Figure 10.

Block-Equicorrelated t : First, as an analogy to the block-equicorrelated Gaussian case, we partition X_1, \dots, X_p into blocks of size 5. We let $X_J \sim t_\nu(0, \frac{\nu-2}{\nu}\Sigma_J)$ where $\Sigma_J \in \mathbb{R}^{5 \times 5}$ is equicorrelated with correlation $\rho = 0.5$ and $\nu = 3$.

Heavy-Tailed Markov Chain: Second, we consider the t -tailed Markov chain discussed in Bates et al. (2020). In particular, assume $R_j \stackrel{i.i.d.}{\sim} t_\nu$, and let $X_1 = \sqrt{\frac{\nu-2}{\nu}}R_1$ and $X_{j+1} = \rho_j X_j + \sqrt{1-\rho_j^2}\sqrt{\frac{\nu-2}{\nu}}R_{j+1}$. In our simulations, we set $\nu = 3$. Similar to the AR1 setting in Section 4.2, we sample the values of $\rho_j \stackrel{i.i.d.}{\sim} \text{Beta}(3, 1)$. As in Appendix F.4, we refer to the default case with the name “AR1”, and we also refer to a case where the non-nulls are clustered together as the “AR1 (Corr)” case.

Gibbs Grid Model: Lastly, we consider a $d \times d$ discrete grid, where the density Φ of X factors over a set of

Various Gaussian Designs, DeepPINK Feature Statistic

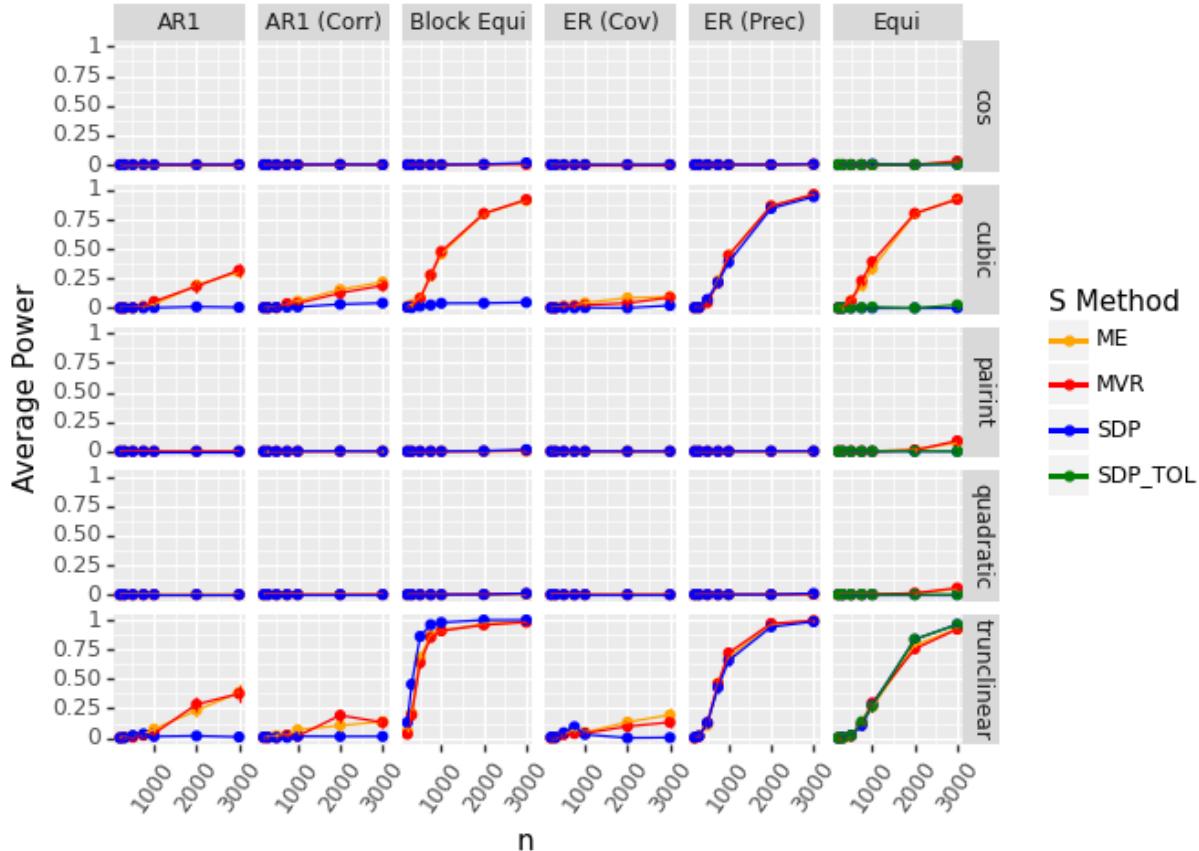


Figure 23: We let $X \sim \mathcal{N}(0, \Sigma)$ for various Σ and $Y | X \sim \mathcal{N}(\mu(X), 1)$. The precise definitions of the covariance matrices and conditional responses are presented in Appendix F.4. We let $p = 200$ with 30 non-nulls. We use DeepPINK feature statistics (see Lu et al. (2018)).

cliques \mathcal{C} which follow a grid structure similar to an Ising model (Ising, 1925). In particular,

$$\Phi(x) = \prod_{\{j_1, j_2\} \in \mathcal{C}} \psi_{j_1, j_2}(x_{j_1}, x_{j_2}) \text{ for } x \in \mathbb{R}^p.$$

We set $\psi_{\{j_1, j_2\}}(x_{j_1}, x_{j_2}) = \exp(\beta_{\{j_1, j_2\}}|x_{j_1} - x_{j_2}|)$ and we let $\beta_{j_1, j_2} = \pm 1$ with equal probability. We allow each feature X_j to take one of 20 evenly spaced values between -2.375 and 2.375 and we choose our covariance-guided proposals with probability proportional to the relevant Gaussian density on each of these values. We employ a Gibbs sampler to sample design matrices X . Furthermore, since this grid has a complicated conditional dependence structure, we use two additional tricks introduced in Bates et al. (2020) to increase the efficiency of the Metro sampler.

First, as outlined in Bates et al. (2020), efficient Metropolized knockoff (Metro) sampling requires that the estimated covariance $\hat{\Sigma}$ used to generate the proposals X^* satisfies certain conditional independence properties. Namely, if the distribution of X is represented as an undirected graphical model (UGM), for any features j_1 and j_2 which are not connected in the UGM, we must have $\hat{\Sigma}_{j_1, j_2}^{-1} = 0$. However, unlike in our experiments with t -distributions, this property does not even hold when using the true covariance matrix Σ . As a result, we use the glasso package from the R programming language to estimate $\hat{\Sigma}$ for the Gibbs grid data to ensure that $\hat{\Sigma}^{-1}$ has the necessary sparsity pattern for efficient sampling. Note that since $\hat{\Sigma}$ is a “compatible proposal” in that it allows for efficient knockoff generation and the probability of our discrete proposals depend only on $\hat{\Sigma}$, our discrete proposals will be compatible as well. We use the same estimated covariance matrix for each of our replications.

Various Gaussian Designs, Lasso Feature Statistic

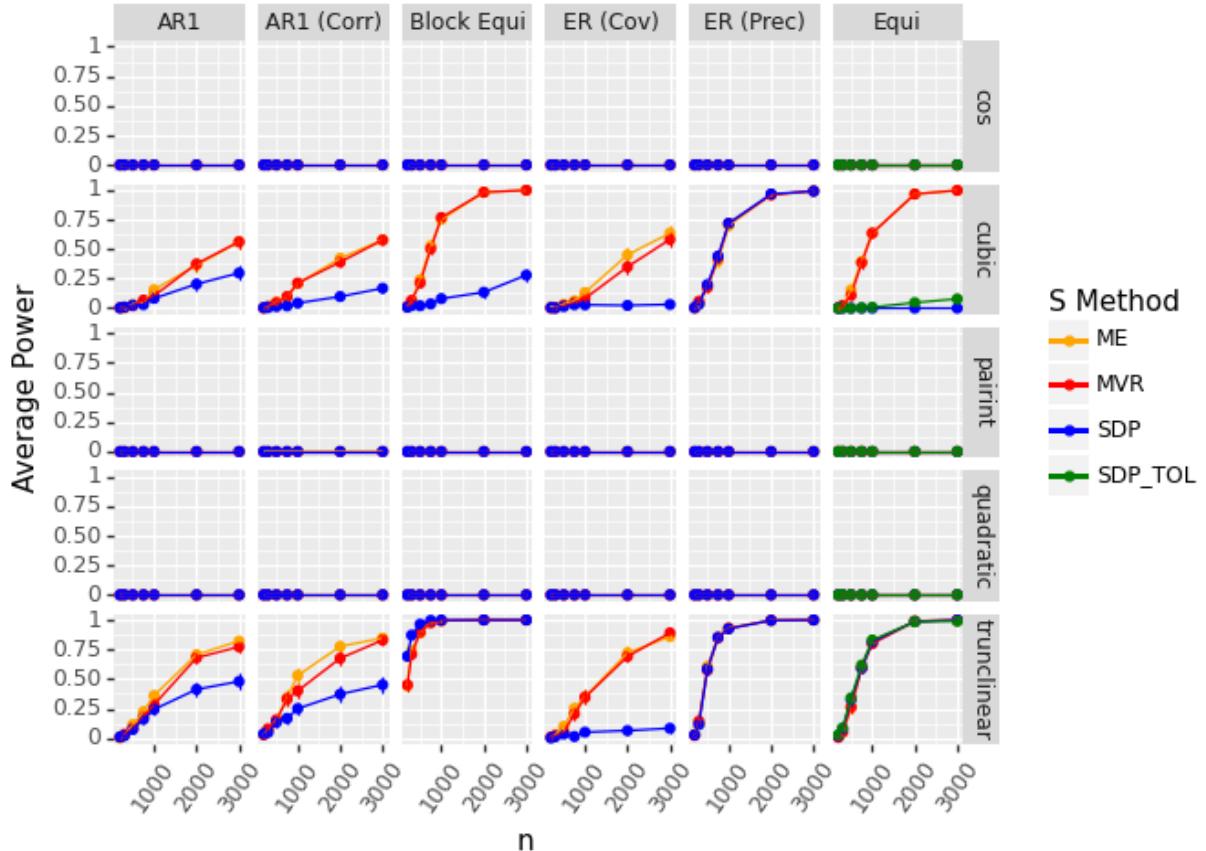


Figure 24: We let $X \sim \mathcal{N}(0, \Sigma)$ for various Σ and $Y | X \sim \mathcal{N}(\mu(X), 1)$. The precise definitions of the covariance matrices and conditional responses are presented in Appendix F.4. We let $p = 200$ with 30 non-nulls. We use lasso coefficient differences as our feature statistics.

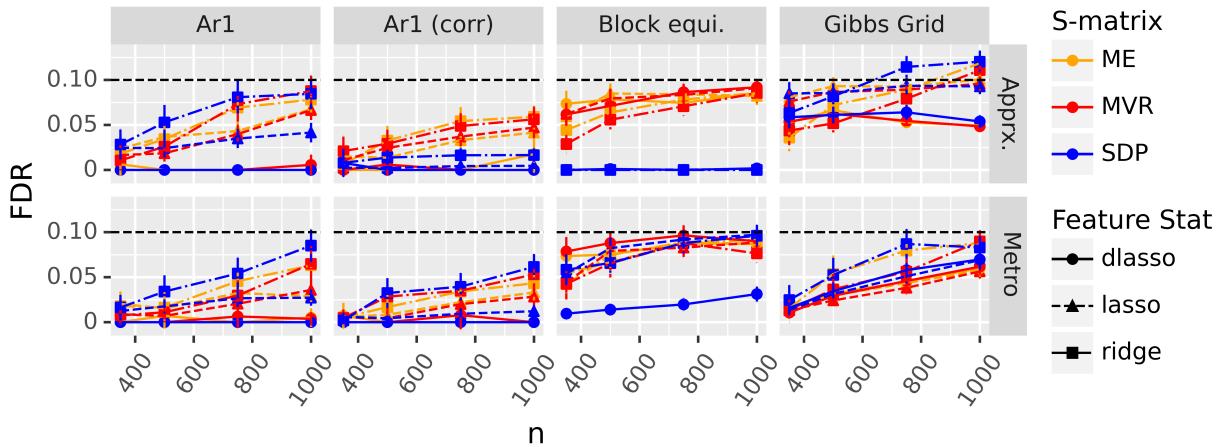


Figure 25: The corresponding FDR plot for Figure 10.

Second, the time complexity of sampling from the Gibbs grid model is prohibitively complex. In particular, for a grid of width $d_1 \times d_2$, the computational complexity of sampling knockoffs for the grid runs in $O(d_1 d_2 2^{\min(d_1, d_2)})$. Bates et al. (2020) suggested a “divide-and-conquer” approach to reduce complexity,

where we set the knockoffs for several rows of the grid to be deterministically equal to the features. Following this approach, we set every 5th row of knockoffs deterministically equal to the features. To increase power for these points, however, we repeat this for both rows and columns on two different translations of the dividing rows, leading to four different divide and conquer mechanisms, which we apply to one quarter of the rows of the data each.

Lastly, we will explain how we generate the proposals X^* for the discrete Gibbs grid model. Suppose $X \sim [0, \Sigma]$, and fix $j \in [p]$. Also fix the G_S matrix, which we will refer to as G for notational convenience. Recall that in the continuous case, when generating proposal X_j^* , we sample X_j^* from a univariate Gaussian with mean $\mu_j(X, X_{1:(j-1)}^*)$ and variance σ_j^2 , where

$$\mu_j(X, X_{1:(j-1)}^*) = G_{1:(p+j-1), p+j}^\top G_{1:(p+j-1), 1:(p+j-1)}^{-1} [X, X_{1:(j-1)}^*]$$

and

$$\sigma_j^2 = G_{p+j, p+j} - G_{1:(p+j-1), p+j}^\top G_{1:(p+j-1), 1:(p+j-1)}^{-1} G_{1:(p+j-1), p+j}.$$

In the discrete case, if X_j can take one of K values v_1, \dots, v_K , we sample X_j^* such that $\mathbb{P}(X_j^* = v_k | X, X_{1:j}^*) \propto \varphi(v_k; \mu_j(X, X_{1:(j-1)}^*), \sigma_j^2)$ where $\varphi(\cdot; \mu, \sigma^2)$ is the PDF of a univariate Gaussian with mean μ and variance σ^2 .

G Overview of knockpy

In this section, we briefly list some of the features of `knockpy`. See <https://github.com/amspector100/knockpy> for more detailed tutorials and documentation.

S-Matrix computation for Gaussian knockoffs: `knockpy` offers efficient algorithms to compute a wide variety of S -matrices, including a fast SDP solver and implementations of Algorithm 1 and Algorithm 2, which respectively compute S_{MVR} and S_{ME} .

Metropolized knockoff sampler: `knockpy` includes a fully general Metropolized knockoff sampler which can generate exactly valid covariance-guided Metropolized knockoffs for any unnormalized density Φ . The Metropolized sampler can also take further advantage of conditional independence structure of X to speed up queries to Φ , yielding an $O(p)$ improvement over the only previous implementation we know of (which did not use covariance-guided proposals).

Feature statistics: `knockpy` offers a whole suite of built-in feature statistics, including cross-validated lasso, ridge, and group-lasso coefficients, lasso-path statistics, the DeepPINK statistic (Lu et al., 2018) and random forest statistics with swap and swap integral importances (Gimenez et al., 2019).

Modularity for development: `knockpy` is built to be modular, such that researchers and analysts can easily layer functionalities on top of it. For example, in under three lines of code, `knockpy` can wrap any python class capable of predicting Y given X and generate feature importances via the swap and swap integral importance procedures from (Gimenez et al., 2019). Each of the major classes of `knockpy` offer explicit ways for users to mix and match `knockpy`'s features with their own code.

Miscellaneous features: `knockpy` also supports a host of other features from the knockoffs literature, including fixed-X knockoffs (Barber and Candès, 2015), some early support for group knockoffs (Dai and Barber, 2016) and knockoff “recycling” (Barber and Candès, 2019).