Asymptotically Optimal Knockoff Statistics via the Masked Likelihood Ratio

Asher Spector[1] and William Fithian[2]

[1]Department of Statistics, Stanford University
[2]Department of Statistics, UC Berkeley

Abstract

This paper introduces a class of asymptotically optimal knockoff statistics called masked likelihood ratio statistics. Our contribution is threefold. First, we derive the optimal *estimand* for knockoff feature statistics, called the oracle masked likelihood ratio (MLR). Second, we show how to estimate the oracle MLR, and we call the resulting procedure the masked likelihood ratio statistic. Our main theoretical result is that MLR statistics are asymptotically average-case optimal, i.e., they maximize the expected number of discoveries made by knockoffs when averaging over a user-specified prior on any unknown parameters. Our main assumption is a "local dependence" condition which depends only on simple quantities that can be calculated from the data, and our theory places no further restrictions on the dimensionality of the problem or the unknown relationship between the response and covariates. Third, we develop concrete instantiations of MLR statistics which are efficient and powerful in practical settings, with a particular focus on linear models, generalized additive models, and binary generalized linear models. We show in simulations and three real data applications that MLR statistics outperform state-of-the-art feature statistics, often by wide margins. We implement several classes of MLR statistics in the open-source python package `knockpy`; our implementation is often (although not always) more efficient than computing a cross-validated lasso.

1 Introduction

Given a design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$ and a response vector $\mathbf{y} \in \mathbb{R}^n$, the task of *controlled feature selection* is, informally, to discover features which influence $\mathbf{y}$ while controlling the false discovery rate (FDR). Knockoffs (Barber and Candès, 2015; Candès et al., 2018) is a powerful method for performing this statistical task. Informally, knockoffs are fake variables $\widetilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$ which act as negative controls for the features $\mathbf{X}$. Remarkably, employing knockoff variables allows analysts to use nearly any machine learning model or test statistic, often known as a "feature statistic," to discover important features while controlling the FDR in finite samples. As a result, knockoffs has become quite popular in a wide variety of settings, including analysis of genetic studies, financial data, clinical trials, and more (Sesia et al., 2018, 2019; Challet et al., 2021; Sechidis et al., 2021).

The flexibility of knockoffs has inspired the development of a wide variety of feature statistics, based on penalized regression coefficients, sparse Bayesian models, random forests, neural networks, and

more (see, e.g., Barber and Candès (2015); Candès et al. (2018); Gimenez et al. (2019); Lu et al. (2018)). These feature statistics not only reflect different modeling assumptions, but more fundamentally, they estimate different quantities, including coefficient sizes, Bayesian posterior inclusion probabilities, and various other measures of variable importance. Yet to our knowledge, there has been relatively little theoretical comparison of these methods, in large part because analyzing the power of knockoffs can be very technically challenging (see Section 1.4 for further discussion). Our work aims to fill this gap: in particular, we consider the question of designing provably optimal knockoff statistics.

1.1 Contribution and overview of results

This paper develops a class of feature statistics called *masked likelihood ratio (MLR)* statistics which are asymptotically optimal, computationally efficient, and powerful in a variety of practical settings. In particular, our contribution is threefold.[1]

1. Conceptual contribution: selecting the estimand. We first argue that a "masked likelihood ratio" is the right *estimand* for knockoff feature statistics. To do this, we reformulate knockoffs as a guessing game on *masked data* $D = (\mathbf{y}, \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p)$. In particular, to discover feature j, an analyst who observes D must recover the true feature $\mathbf{X}_j$ from the unordered pair $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$. This is a binary decision problem, and we show that the statistic which asymptotically maximizes the number of discoveries is the Neyman-Pearson test statistic for testing $H_0 : \mathbf{X}_j = \widetilde{\mathbf{x}}_j$ against $H_a : \mathbf{X}_j = \mathbf{x}_j$, where $\mathbf{x}_j$ (resp. $\widetilde{\mathbf{x}}_j$) is the observed value of the jth feature (resp. knockoff) vector. We refer to this as the *oracle masked likelihood ratio*, shown below:

$$\mathrm{MLR}_j^{\mathrm{oracle}} := \log\left(\frac{L_\theta(\mathbf{X}_j = \mathbf{x}_j \mid D)}{L_\theta(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)}\right). \tag{1.1}$$

Above, $L_\theta(\mathbf{X}_j = \mathbf{x} \mid D)$ is the likelihood of observing $\mathbf{X}_j = \mathbf{x} \in \mathbb{R}^n$ conditional on the masked data D, and θ are any unknown parameters which may affect the likelihood, such as coefficients in a generalized linear model (GLM) or parameters in a random forest or neural network. To aid intuition, observe that swapping $\mathbf{x}_j$ and $\widetilde{\mathbf{x}}_j$ flips the sign of $\mathrm{MLR}_j^{\mathrm{oracle}}$, and thus $\mathrm{MLR}_j^{\mathrm{oracle}}$ is indeed a valid knockoff statistic (as reviewed in Section 1.3). We also note that Katsevich and Ramdas (2020) previously argued the *vanilla* likelihood had certain (weaker) optimality properties: we reconcile these results in Sections 1.4 and 2.3.

2. Theoretical contribution: optimality of masked likelihood ratio (MLR) statistics. The exact optimal knockoff statistics (1.1) depend on the likelihood, which is unknown by assumption. To avoid this impossibility result, we settle for statistics which are average-case optimal over a user-specified prior π on the unknown parameters θ. Marginalizing over θ yields the *masked likelihood ratio (MLR) statistic*:

$$W_j^\star := \log\left(\frac{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j = \mathbf{x}_j \mid D)\right]}{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)\right]}\right). \tag{1.2}$$

Our main theoretical result (Theorem 2.1) shows that MLR statistics are asymptotically average-case optimal over π, i.e., they maximize the expected number of discoveries. The proof is involved,

[1]For brevity, in this section, we use the notation of model-X knockoffs to outline our results: analagous results for fixed-X knockoffs are available throughout the paper and appendix.

but our result allows for arbitrarily high-dimensional asymptotic regimes and allows the likelihood to take any form—in particular, we do not assume $\mathbf{y} \mid \mathbf{X}$ follows a linear model. Instead, the key assumption we make is that the signs of the MLR statistics satisfy a local dependency condition, similar in flavor to dependency conditions assumed on p-values in the multiple testing literature (Genovese and Wasserman, 2004; Storey et al., 2004; Ferreira and Zwinderman, 2006; Farcomeni, 2007). Crucially, our local dependency condition only depends on simple quantities which are (usually) easy to compute from the data, so it can be roughly diagnosed using the data at hand.

3. Methodological contribution: practical and powerful MLR statistics. Our third contribution is to demonstrate via simulations and three real data analyses that MLR statistics are powerful in practical settings. In particular, our theory only shows that MLR statistics are *average-case* optimal over a prior on the distribution of $y \mid X$ (although the analyst may choose any prior), so we aim to show that MLR statistics are powerful even when the prior is highly mispecified.[2]

- We develop concrete instantiations of MLR statistics based on (relatively) uninformative priors, including versions developed for use in Gaussian linear models, generalized additive models, and binary generalized linear models. We also develop efficient software in python to fit MLR feature statistics. Our implementation is extremely efficient, in many cases substantially faster than fitting a cross-validated lasso.

- In extensive empirical studies, we show that MLR statistics outperform other state-of-the-art feature statistics, often by wide margins, including settings where the prior is highly mispecified. Indeed, MLR statistics often nearly match the performance of the oracle procedure in Equation (1.1), showing that the mispecified prior provably has little effect. Furthermore, in settings where $\mathbf{y}$ has a highly nonlinear relationship with $\mathbf{X}$, MLR statistics also outperform feature statistics based on black-box machine learning algorithms.

- We replicate three knockoff-based analyses of drug resistance (Barber and Candès, 2015), financial factor selection (Challet et al., 2021), and graphical model discovery for gene expression data (Li and Maathuis, 2019), where we show that MLR statistics (using an uninformative prior) make up to an order of magnitude more discoveries than lasso-based competitors.

Overall, our results suggest that MLR statistics can substantially increase the power of knockoffs.

1.2 Notation

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$ denote the design matrix and response vector in a feature selection problem with n data points and p features. In settings where the design matrix is random, we let $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$ denote the observed values of the columns of the design matrix. For convenience, we let the non-bold versions $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ denote the features and response for an arbitrary single observation. For any integer k, define $[k] \coloneqq \{1, \ldots, k\}$. For any matrix $M \in \mathbb{R}^{m \times k}$ and any $J \subset [k]$, let M_J denote the columns of M corresponding to the indices in J. Similarly, M_{-J} denotes the columns of M which do not appear in J, and let M_{-j} denote all columns except column $j \in [k]$. For any matrices $M_1 \in \mathbb{R}^{n \times k_1}, M_2 \in \mathbb{R}^{n \times k_2}$, we define $[M_1, M_2] \in \mathbb{R}^{n \times (k_1 + k_2)}$ to be the column-wise concatenation of M_1, M_2. Let $[M_1, M_2]_{\text{swap}(j)}$ denote the matrix $[M_1, M_2]$ but with the jth column of M_1 and M_2 swapped: similarly, $[M_1, M_2]_{\text{swap}(J)}$ swaps all columns $j \in J$ of M_1 and M_2. Let I_n denote the $n \times n$ identity. Furthermore, for any vector $x \in \mathbb{R}^n$, we let

[2] Of course, knockoffs provably control the FDR even when the prior is mispecified.

$\bar{x}_k \coloneqq \frac{1}{k}\sum_{i=1}^{k} x_i$ be the sample mean of the first k elements of x. Furthermore, for $x \in \mathbb{R}^n$ and a permutation $\kappa : [n] \to [n]$, $\kappa(x)$ denotes the coordinates of x permuted according to κ, so that $\kappa(x)_i = x_{\kappa(i)}$.

1.3 Review of knockoffs

We start with a review of model-X (MX) knockoffs (Candès et al., 2018), which tests the hypotheses $H_j : X_j \perp\!\!\!\perp Y \mid X_{-j}$ assuming only that the data are i.i.d. and that the distribution of X is known. Applying MX knockoffs requires three steps.

Step 1: constructing knockoffs. Valid MX knockoffs $\widetilde{X}$ must satisfy two properties. First, X and $\widetilde{X}$ must be *pairwise exchangeable*, meaning $[X, \widetilde{X}]_{\text{swap}(j)} \overset{\text{d}}{=} [X, \widetilde{X}]$ for each $j \in [p]$. Second, we require that $\widetilde{X} \perp\!\!\!\perp Y \mid X$, which holds, (e.g.) if one constructs $\widetilde{X}$ without looking at Y. Informally, these constraints guarantee that $X_j, \widetilde{X}_j$ are "indistinguishable" under H_j. Sampling knockoffs can be challenging, but this problem is well studied (see, e.g., Bates et al. (2020)).

Step 2: fitting feature statistics. Next, one can use almost any machine learning algorithm to fit feature importances $Z = z([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}) \in \mathbb{R}^{2p}$, where Z_j and Z_{j+p} heuristically measure the "importance" of X_j and $\widetilde{X}_j$ in predicting Y. The only restriction on z is that swapping the positions of X_j and $\widetilde{X}_j$ must also swap the feature importances Z_j and Z_{j+p}, and swapping the positions of X_i and $\widetilde{X}_i$ does not change Z_j or Z_{j+p} for $i \neq j$. This restriction is satisfied by most machine learning algorithms, such as the lasso or various neural networks (Lu et al., 2018).

Given Z, we define the *feature statistics* $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}) \in \mathbb{R}^p$ via $W_j = f(Z_j, Z_{j+p})$, where f is any "antisymmetric" function, meaning $f(x, y) = -f(y, x)$. For example, the lasso coefficient difference (LCD) statistic sets $W_j = Z_j - Z_{j+p}$, where Z_j and Z_{j+p} are the absolute coefficients returned by a lasso model fit on $[\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}$. Intuitively, when W_j has a positive sign, this suggests that X_j is more important than $\widetilde{X}_j$ in predicting Y and thus is evidence against the null. Indeed, Step 1 and 2 guarantee that the signs of the null W_j are distributed as i.i.d. random coin flips.

Step 3: make rejections. Define the data-dependent threshold $T \coloneqq \inf\left\{t > 0 : \frac{\#\{j : W_j \leq -t\}+1}{\#\{W_j \geq t\} \leq q}\right\}$.[3] Then, reject $\hat{S} = \{j : W_j \geq T\}$, which guarantees finite-sample FDR control at level $q \in (0, 1)$. It is important to note that knockoffs only makes rejections when the feature statistics which have high absolute values are also consistently positive.

Lastly, we also briefly describe fixed-X knockoffs (Barber and Candès, 2015), which do not require knowledge of the distribution of X and control the FDR under the Gaussian linear model where $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$. In particular, FX knockoffs do not need to satisfy the constraints in Step 1: instead, $\widetilde{\mathbf{X}}$ must satisfy (i) $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X}$ and (ii) $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} - S$, for some diagonal matrix S satisfying $2\mathbf{X}^T\mathbf{X} \succ S$. The other difference between FX and MX knockoffs is that the feature importances Z can only depend on $\mathbf{y}$ only through $[\mathbf{X}, \widetilde{\mathbf{X}}]^T\mathbf{y}$, which permits the use of many machine learning models, but not all (for example, it prohibits the use of cross-validation). Other than these differences, fitting FX knockoffs uses the same three steps outlined above.

[3] By convention we define the infimum of the empty set to be ∞.

1.4 Related literature

The knockoffs literature contains a vast number of feature statistics, which can (roughly) be separated into three categories. First, some of the most common feature statistics are those based on penalized regression coefficients, notably the lasso signed maximum (LSM) and lasso coefficient difference (LCD) statistics of Barber and Candès (2015). Indeed, these lasso-based statistics are often used in applied work (Sesia et al., 2019) and have received a great deal of theoretical attention (Weinstein et al., 2017; Fan et al., 2020; Ke et al., 2020; Weinstein et al., 2020; Wang and Janson, 2021). Perhaps surprisingly, we argue in this paper that many of these statistics are estimating the wrong quantity, leading to substantially reduced power. Second, some previous works have introduced Bayesian knockoff statistics (see, e.g., Candès et al. (2018); Ren and Candès (2020)). MLR statistics have a Bayesian flavor, but they take a different form than previous statistics. Furthermore, they have a fundamentally different motivation than previous Bayesian statistics: the real innovation of MLR statistics is to attempt to estimate a (masked) likelihood ratio, and we only use a Bayesian framework to appropriately quantify our uncertainty about nuisance parameters. In contrast, previous works largely motivated Bayesian statistics as a way to incorporate prior information (Candès et al., 2018) or side information about which features are non-null (Ren and Candès, 2020). That said, it is worth noting that an important special case of MLR statistics is similar to the "BVS" statistics from Candès et al. (2018), as we discuss in Section 2.5. Third and finally, many feature statistics take advantage of "black-box" machine learning methods to assign variable importances (see, e.g., Lu et al. (2018); Gimenez et al. (2019)). Although MLR statistics are most easily computable when the analyst specifies a parametric model, our implementation of MLR statistics based on regression splines performs favorably compared to "black-box" feature statistics in Section 3.

Previous analyses of the power of knockoffs have largely focused on showing that coefficient-difference feature statistics are consistent under regularity conditions on $\mathbf{X}$ (Liu and Rigollet, 2019; Fan et al., 2020; Spector and Janson, 2022) or explicitly quantifying the power of coefficient-difference feature statistics assuming $\mathbf{X}$ has i.i.d. Gaussian entries (Weinstein et al., 2017, 2020; Wang and Janson, 2021). Ke et al. (2020) also derive a phase diagram for LCD statistics assuming $\mathbf{X}$ is blockwise orthogonal with a blocksize of 2. Our work has a different goal, which is to show that MLR statistics are asymptotically optimal, with particular focus on nontrivial settings where the asymptotic power of MLR statistics is strictly between 0 and 1. Furthermore, our analysis places no explicit restrictions on $y \mid X$, $\mathbf{X}$, or the dimensionality of the problem, in contrast to the aforementioned works, which exclusively focus on the case where $y \mid X$ follows a Gaussian linear model. Instead, the key assumption we make is that the signs of the MLR statistics satisfy a local dependency condition, similar to dependency conditions assumed on p-values in the multiple testing literature (Genovese and Wasserman, 2004; Storey et al., 2004; Ferreira and Zwinderman, 2006; Farcomeni, 2007). However, our proof technique is specific to knockoffs.

Our theory is close in spirit to that of Li and Fithian (2021), who developed knockoff*, an oracle statistic for FX knockoffs which provably maximizes power in finite samples. Indeed, the oracle masked likelihood ratio statistics are equivalent to knockoff* when $\mathbf{y} \mid \mathbf{X}$ follows a Gaussian linear model and $\widetilde{\mathbf{X}}$ are fixed-X knockoffs. To our knowledge, the only other work which attempts to prove optimality guarantees for knockoff statistics is Katsevich and Ramdas (2020), who showed that using the likelihood as the feature statistic maximizes $\mathbb{P}(W_j > 0)$ against a point alternative. We see our work as building on that of Katsevich and Ramdas (2020), as MLR statistics also have this property (averaging over the prior), although we go substantially farther and show that

MLR statistics maximize the number of discoveries of the overall knockoffs procedure. Another key difference is that vanilla likelihood statistics technically violate the pairwise exchangeability condition, whereas working with the masked likelihood guarantees FDR control. Lastly, we note that the oracle procedures derived in Li and Fithian (2021) and Katsevich and Ramdas (2020) cannot be used in practice since they depend on unknown parameters—to our knowledge, MLR statistics are the first computable knockoff statistics with explicit (if asymptotic) optimality guarantees on their power.

1.5 Outline

The rest of the paper proceeds as follows. In Section 2, we prove an equivalent formulation of the knockoffs methodology, introduce MLR statistics, and present the main theoretical results of the paper. We also give concrete suggestions on the choice of prior and discuss how to compute MLR statistics efficiently. In Section 3, we present simulations comparing the power of MLR statistics to common feature statistics from the literature. In Section 4, we apply MLR statistics to three real datasets previously analyzed using knockoffs. Finally, Section 5 concludes with a discussion of future directions.

2 Introducing masked likelihood ratio statistics

2.1 A motivating example

Before proving our core results, we begin by giving some intuition as to why standard feature statistics like the LCD and LSM statistics are sub-optimal. As notation, let $(\hat{\beta}^{(\lambda)}, \tilde{\beta}^{(\lambda)}) \in \mathbb{R}^{2p}$ denote the estimated lasso coefficients fit on $[\mathbf{X}, \widetilde{\mathbf{X}}]$ and $\mathbf{y}$ with regularization parameter λ. Furthermore, let $\hat{\lambda}_j$ (resp. $\tilde{\lambda}_j$) denote the smallest value of λ such that $\hat{\beta}_j^{(\lambda)} \neq 0$ (resp. $\tilde{\beta}_j^{(\lambda)} \neq 0$). Then the LCD and LSM statistics are defined as:

$$W_j^{\text{LCD}} = |\hat{\beta}_j^{(\lambda)}| - |\tilde{\beta}_j^{(\lambda)}|, \qquad W_j^{\text{LSM}} = \text{sign}(\hat{\lambda}_j - \tilde{\lambda}_j) \max(\hat{\lambda}_j, \tilde{\lambda}_j). \tag{2.1}$$

As a thought experiment, imagine that we observe a covariate $\mathbf{X}_j$ which appears to significantly influence $\mathbf{y}$: however, due to high correlations within $\mathbf{X}$, we must create a knockoff $\widetilde{\mathbf{X}}_j$ which is highly correlated with $\mathbf{X}_j$, such that the lasso is reasonably likely to select $\widetilde{\mathbf{X}}_j$ instead of $\mathbf{X}_j$. (Since the lasso induces sparsity, it is unlikely to select *both* $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$ if they are highly correlated.) In this case, W_j^{LCD} and W_j^{LSM} will both have large absolute values, since $\mathbf{X}_j$ appears significant, but W_j^{LCD} and W_j^{LSM} will have a reasonably high probability of being negative, since $\mathbf{X}_j \approx \widetilde{\mathbf{X}}_j$. This is a serious problem, since as shown in Figures 1 and 2, knockoffs can only make discoveries when the feature statistics W_j with the largest absolute values are consistently positive. Indeed, even when the LCD and LSM statistics only produce a few highly negative W-statistics, this can still prevent knockoffs from making *any* discoveries at all.

That said, it is important to note that this problem is largely avoidable. Indeed, if we observe that $\text{Corr}(X_j, \widetilde{X}_j)$ is very large and thus there is some risk that W_j is highly negative, we can simply "deprioritize" W_j by lowering its absolute value. This allows us to potentially discover W_j while reducing the risk that a large and negative W-statistic prevents the procedure from making many

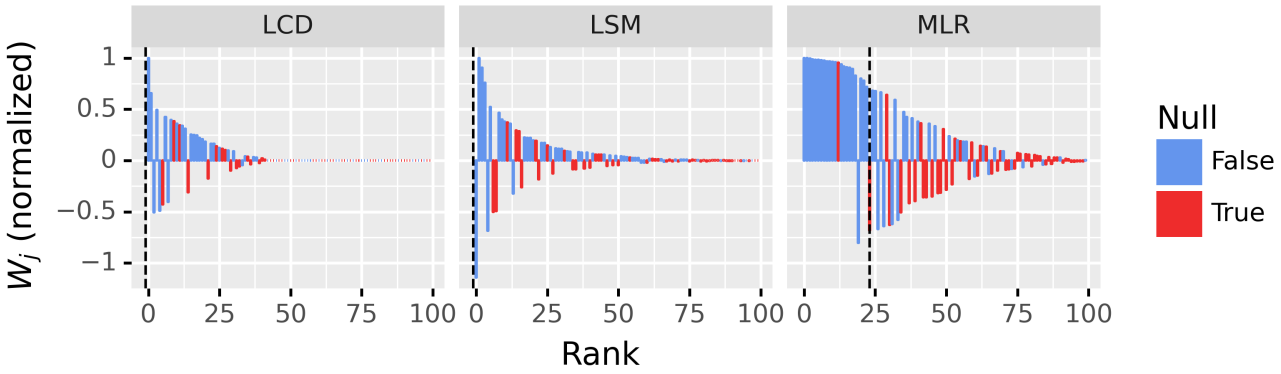

Figure 1: We plot the LCD, LSM, and MLR feature statistics sorted in descending order of absolute value. The data-dependent threshold is shown by the black line. This figure shows that when $\mathbf{X}$ is highly correlated, LCD and LSM statistics make few discoveries because they occasionally yield highly negative W-statistics for non-null variables which have low-quality knockoffs. In this setting, we fit fixed-X knockoffs and the rows of $\mathbf{X}$ follow a Gaussian AR1 process with correlations from Unif$([0.8, 0.99])$. For visualization, we normalize $\{W_j\}$ such that $|W_j| \leq 1$, which (provably) does not change the performance of knockoffs.

discoveries. The main point is as follows: while it is important to rank the W_j by some notion of the signal strength, the signal strength for knockoffs is not fully captured by the estimated coefficients and also depends on the (known) dependence structure among $[\mathbf{X}, \widetilde{\mathbf{X}}]$. In the next two sections, we argue that the masked likelihood ratio is the "correct" measure of signal strength.

2.2 Knockoffs as inference on masked data

In the previous section, we argued that to maximize power, W_j should have a large absolute value if and only if $\mathbb{P}(W_j > 0)$ is large. This raises the question: how we can know when $\mathbb{P}(W_j > 0)$ is large? Indeed, we will need to estimate $\mathbb{P}(W_j > 0)$ from the data, but we cannot use *all* the data for this purpose: for example, we cannot directly observe $\text{sign}(W)$ and use it to adjust $|W|$ without violating FDR control. To resolve this ambiguity, we reformulate knockoffs as inference on *masked data*, as defined below.

Definition 2.1. *Suppose we observe data* $\mathbf{X}, \mathbf{y}$*, knockoffs* $\widetilde{\mathbf{X}}$*, and independent random noise* U*. (*U *may be used to fit a randomized feature statistic.) The* masked data D *is defined as*

$$D = \begin{cases} (\mathbf{y}, \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p, U) & \textit{for model-X knockoffs} \\ (\mathbf{X}, \widetilde{\mathbf{X}}, \{\mathbf{X}_j^T\mathbf{y}, \widetilde{\mathbf{X}}_j^T\mathbf{y}\}_{j=1}^p, U) & \textit{for fixed-X knockoffs.} \end{cases} \tag{2.2}$$

As we will see in a moment, the masked data D is all of the data we are "allowed" to use when fitting a feature statistic W, and knockoffs will be powerful precisely when we can reliably recover the full data from D. For example, in the model-X case, D contains $\mathbf{y}$ and the unordered pairs of vectors $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$, and Proposition 2.1 tells us that ensuring W_j is positive is equivalent to recovering $\mathbf{X}_j$ from $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$.[4]

[4]Throughout the paper, we only consider W-statistics which are nonzero with probability one, because one can provably increase the power of knockoffs by ensuring that each coordinate of W is nonzero.

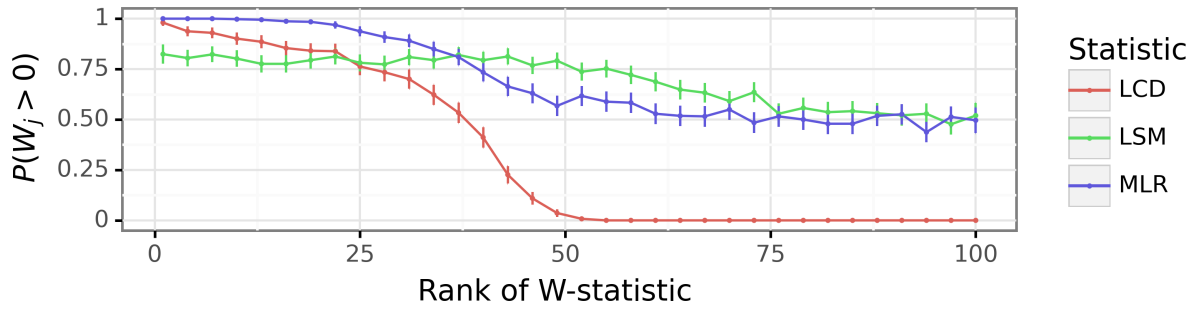

Figure 2: In the same setting as Figure 1, this plots the probability that the sorted W-statistics are positive. It shows that MLR statistics do a better job ensuring that the W-statistics with large absolute values are consistently positive.

Proposition 2.1. *Let $\widetilde{\mathbf{X}}$ be model-X knockoffs such that $\mathbf{X}_j \neq \widetilde{\mathbf{X}}_j$ for $j \in [p]$. Then $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ is a valid feature statistic if and only if (i) $|W|$ is a function of the masked data D and (ii) there exist* distinguishing *functions $\widehat{\mathbf{X}}_j = g_j(D)$ such that $W_j > 0$ if and only if $\widehat{\mathbf{X}}_j = \mathbf{X}_j$.*

In particular, Proposition 2.1 reformulates knockoffs as a guessing game, where ensuring $W_j > 0$ is equivalent to recovering $\mathbf{X}_j$ from $\mathbf{y}$ and $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p$. (Note that $\widehat{\mathbf{X}}_j$ is perhaps an unusual estimator, as it is a vector but can only take one of two values, namely $\widehat{\mathbf{X}}_j \in \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$.) If our "guess" is right, meaning $\widehat{\mathbf{X}}_j = \mathbf{X}_j$, then we are rewarded and $W_j > 0$: else $W_j < 0$. Furthermore, to avoid highly negative W-statistics, this suggests that we should only assign W_j a large absolute value when we are a confident that our "guess" $\widehat{\mathbf{X}}_j$ is correct. We discuss more implications of this result in the next section: for now, we obtain an analagous result for fixed-X knockoffs (similar to a result from Li and Fithian (2021)) by substituting $\{\mathbf{X}_j^T\mathbf{y}, \widetilde{\mathbf{X}}_j^T\mathbf{y}\}$ for $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$.

Proposition 2.2. *Let $\widetilde{\mathbf{X}}$ be fixed-X knockoffs satisfying $\mathbf{X}_j^T\mathbf{y} \neq \widetilde{\mathbf{X}}_j^T\mathbf{y}$ for $j \in [p]$. Then $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ is a valid feature statistic if and only if (i) $|W|$ is a function of D and (ii) there exist distinguishing functions $g_j(D)$ such that $W_j > 0$ if and only if $\mathbf{X}_j^T\mathbf{y} = g_j(D)$.*

Note that Propositions 2.1 and 2.2 hold for the original definition of knockoffs in Barber and Candès (2015); Candès et al. (2018): however, one could slightly extend the definition of knockoffs and augment D to contain any random variable which is independent of those specified in Equation (2.2). For example, in the fixed-X case, one can also include $\hat{\sigma}^2 = \|(I_n - H)\mathbf{y}\|_2^2$ where H is the projection matrix of $[\mathbf{X}, \widetilde{\mathbf{X}}]$ (Chen et al., 2019; Li and Fithian, 2021). Our proofs for the rest of the paper also apply to such extensions.

2.3 Introducing masked likelihood ratio (MLR) statistics

In this section, we introduce masked likelihood ratio (MLR) statistics. For brevity, this subsection focuses on the case of model-X knockoffs; the analagous results for fixed-X knockoffs merely replace $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$ with $\{\mathbf{X}_j^T\mathbf{y}, \widetilde{\mathbf{X}}_j^T\mathbf{y}\}$, as in Definition 2.2. Now, to build intuition, recall that Propositon 2.1 tells us that model-X knockoffs can be reformulated as the following guessing game. After observing the masked data $D = (\mathbf{y}, \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p)$, the analyst must do the following.

1. Step 1: using D, the analyst produces "guesses" $\widehat{\mathbf{X}}_j$ of the true value of $\mathbf{X}_j$ for each $j \in [p]$.

2. Step 2: the analyst may arbitrarily reorder the hypotheses and sequentially check which guesses $\widehat{\mathbf{X}}_j$ are correct (in the order they specified).
3. Step 3: let k be the maximum number such that the first $\lceil (k+1)/(1+q) \rceil$ guesses were correct. Then, the analyst can reject the null hypotheses corresponding to their first $\lceil (k+1)/(1+q) \rceil$ correct guesses while provably controlling the FDR at level q.

Note that in the traditional language of knockoffs, Step 1 determines $\text{sign}(W)$, Step 2 corresponds to choosing the absolute values $|W|$, and Step 3 is simply a description of the SeqStep procedure. However, this reformulation suggests a very intuitive strategy to maximize the number of discoveries:

1. In Step 1, the analyst should guess the value $\widehat{\mathbf{X}}_j \in \{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}$ which maximizes the likelihood of observing the masked data, denoted $L_\theta(D)$. Indeed, this is a standard binary decision problem, so this choice of $\widehat{\mathbf{X}}_j$ maximizes the chance that each guess is correct. Equivalently, this maximizes $\mathbb{P}(W_j > 0 \mid D)$ since $\mathbb{P}(W_j > 0 \mid D) = \mathbb{P}(\widehat{\mathbf{X}}_j = \mathbf{X}_j \mid D)$.
2. In Step 2, the analyst should rank the hypotheses by the likelihood that each guess $\widehat{\mathbf{X}}_j$ is correct. This ensures that for each k, the first k ranked hypotheses contain as many "good guesses" as possible, which thus maximizes the number of discoveries in Step 3. Equivalently, in the traditional notation of knockoffs, this means that $\{|W_j|\}_{j=1}^p$ has the same order as $\{\mathbb{P}(W_j > 0 \mid D)\}_{j=1}^p$

Both of these criteria are achieved by the Neyman-Pearson test statistic which tests $H_0 : \mathbf{X}_j = \widetilde{\mathbf{x}}_j$ against the alternative $H_a : \mathbf{X}_j = \mathbf{x}_j$ using D. We refer to this statistic as the (oracle) masked likelihood ratio statistic, defined below:

$$\text{MLR}_j^{\text{oracle}} = \log\left(\frac{L_\theta(\mathbf{X}_j = \mathbf{x}_j \mid D)}{L_\theta(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)}\right), \tag{1.1}$$

where $L_\theta(\mathbf{X}_j = \mathbf{x} \mid D)$ is the likelihood of observing $\mathbf{X}_j = \mathbf{x}$ conditional on D, and θ are any unknown parameters that affect the likelihood, such as the coefficients in a GLM or the parameters in a neural network. In a moment, we will verify that this achieves the criteria specified above, and in Section 2.4, we will also show that for a fixed distribution of $(\mathbf{X}, \mathbf{y})$, choosing $\text{MLR}_j^{\text{oracle}}$ as the test statistic asymptotically maximizes the expected number of discoveries under mild regularity conditions.

Unfortunately, $\text{MLR}_j^{\text{oracle}}$ is not useful, since it depends on θ, which is unknown. A heuristic solution would be to replace θ with an estimator $\hat{\theta}$ of θ, but we found that this "plug-in" approach performs poorly since the masked likelihood $L_\theta(D)$ is usually highly multimodal and non-convex (see Appendix D.1) for discussion). Instead, to appropriately account for uncertainty over θ, we recommend a Bayesian approach. Let π be any user-specified prior over the unknown parameters θ—in Section 3, we will give several choices of uninformative priors which perform well even when they are mispecified. We will settle for feature statistics which are *average-case* optimal over π, which is perhaps the best we could hope for, since the optimal statistics for fixed θ are effectively uncomputable. Note that while MLR statistics are only average-case optimal, we emphasize that this in no way compromises their (pointwise) ability to control the FDR, which is always guaranteed by the knockoff procedure. To quote Ren and Candès (2020), "wrong models [or priors] do not hurt FDR control!"

Given such a prior π, we define MLR statistics below.

Definition 2.2 (MLR statistics)**.** *Suppose we observe data* $\mathbf{y}, \mathbf{X}$ *with knockoffs* $\widetilde{\mathbf{X}}$*. Let* D *be the masked data as in Equation (2.2). For any prior* $\theta \sim \pi$ *on the masked likelihood* $L_\theta(D)$*, we define*

the model-X masked likelihood ratio (MLR) statistic by marginalizing over θ:

$$W_j^\star := \log\left(\frac{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j = \mathbf{x}_j \mid D)\right]}{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)\right]}\right) \text{ for model-X knockoffs.}^5 \tag{2.3}$$

The fixed-X MLR statistic is analagous but replaces $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$ with $\{\mathbf{X}_j^T\mathbf{y}, \widetilde{\mathbf{X}}_j^T\mathbf{y}\}$: in this case,

$$W_j^\star := \log\left(\frac{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j^T\mathbf{y} = \mathbf{x}_j^T\mathbf{y} \mid D)\right]}{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j^T\mathbf{y} = \widetilde{\mathbf{x}}_j^T\mathbf{y} \mid D)\right]}\right) \text{ for fixed-X knockoffs.} \tag{2.4}$$

Note that we can easily extend this definition to apply to *group* knockoffs, but for brevity, we defer this extension to Appendix E.

The MLR statistic is also closely related to the Neyman-Pearson test statistic which (in the model-X case) tests $H_0 : \mathbf{X}_j = \widetilde{\mathbf{x}}_j$ against the alternative $H_a : \mathbf{X}_j = \mathbf{x}_j$, except now, these null and alternative hypotheses implicitly marginalize over the unknown parameters θ according to π. As a result, we can easily verify that averaging over π, MLR statistics achieve the intuitive optimality criteria outlined at the beginning of this section. Note that the proposition below also applies to the oracle MLR statistics, since $\text{MLR}_j^{\text{oracle}} = W_j^\star$ in the special case where π is a point mass on some fixed θ.

Proposition 2.3. *Given data $\mathbf{y}, \mathbf{X}$ and knockoffs $\widetilde{\mathbf{X}}$, let $W^\star$ be the MLR statistics and let W be any other valid knockoff feature statistic. Then,*

$$\mathbb{P}(W_j^\star > 0 \mid D) \geq \mathbb{P}(W_j > 0 \mid D). \tag{2.5}$$

Furthermore, $\{|W_j^\star|\}_{j=1}^p$ has the same order as $\{\mathbb{P}(W_j^\star > 0 \mid D)\}_{j=1}^p$. More precisely,

$$\mathbb{P}(W_j^\star > 0 \mid D) = \frac{\exp(|W_j^\star|)}{1 + \exp(|W_j^\star|)}. \tag{2.6}$$

These results hold in the average case, i.e., they hold over the posterior distribution $(\mathbf{y}, \mathbf{X}, \theta) \mid D$.

Proof. First, we prove Equation (2.5). Recall by Proposition 2.1 that for any W_j, there exists a function $\widehat{\mathbf{X}}_j$ of D such that $W_j > 0$ if and only if $\widehat{\mathbf{X}}_j = \mathbf{X}_j$. Standard decision theory for binary decision problems tells us that $\mathbb{P}(W_j > 0)$ is maximized if $\widehat{\mathbf{X}}_j = \arg\max_{\mathbf{x}\in\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}} \mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid D)$, where the probability is taken over $(\mathbf{y}, \mathbf{X}, \theta) \mid D$. However, by the tower property, $\mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid D) = \mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_j = \mathbf{x} \mid D)\right]$, and thus $W_j^\star > 0$ if and only if $\mathbb{P}(\mathbf{X}_j = \mathbf{x}_j \mid D) > \mathbb{P}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)$. This proves that for $W_j^\star$, $\widehat{\mathbf{X}}_j = \arg\max_{\mathbf{x}\in\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}} \mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid D)$, which thus proves Equation (2.5).

To prove Equation (2.6), observe $\mathbb{E}_{\theta\sim\pi}[L(\mathbf{X}_j = \mathbf{x} \mid D)] = \mathbb{P}(\mathbf{X}_j = \mathbf{x}_j \mid D) = 1 - \mathbb{P}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)$, so

$$|W_j^\star| = \log\left(\frac{\max_{\mathbf{x}\in\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}} \mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid D)}{1 - \max_{\mathbf{x}\in\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}} \mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid D)}\right) = \log\left(\frac{\mathbb{P}(W_j^\star > 0)}{1 - \mathbb{P}(W_j^\star > 0)}\right),$$

where the second step uses the fact that $\mathbb{P}(W_j^\star > 0 \mid D) = \mathbb{P}\left(\mathbf{X}_j = \widehat{\mathbf{X}}_j \mid D\right)$ for $\widehat{\mathbf{X}}_j$ as defined above. This completes the proof for model-X knockoffs; the proof in the fixed-X case is analagous. □

[5] In the edge case where the masked likelihood ratio is exactly zero, we set $W_j^\star \overset{\text{ind}}{\sim} \text{Unif}(\{-\epsilon, +\epsilon\})$ where ϵ is chosen such that $\epsilon < |W_k^\star|$ for each k such that the MLR is nonzero.

Proposition 2.3 tells us that MLR statistics avoid the problem identified in Section 2.1. In particular, if $\mathbf{X}_j$ appears highly significant but $\widetilde{\mathbf{X}}_j$ is nearly indistinguishable from $\mathbf{X}_j$, the absolute value of $W_j^\star$ will still be quite small, since $\mathbf{x}_j \approx \widetilde{\mathbf{x}}_j$ intuitively implies $L_\theta(\mathbf{X}_j = \mathbf{x}_j \mid D) \approx L_\theta(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)$. As a result, MLR statistics are explicitly designed to avoid situations where $W_j^\star$ is highly negative, as confirmed by Figure 1. Indeed, Equation (2.6) tells us that the absolute values $|W_j^\star|$ have the same order as $\mathbb{P}(W_j^\star > 0 \mid D)$, so, MLR statistics will always prioritize the hypotheses "correctly."

Lastly, we pause to make two connections to the literature. First, Proposition 2.3 suggests a similarity between MLR statistics and a proposal in Ren and Candès (2020) to rank the hypotheses by $\mathbb{P}(W_j > 0 \mid |W_j|)$ (see their footnote 8). MLR statistics satisfy a similar property, except we condition on *all* of the masked data D, leading to provably higher power. Furthermore, this initial similarity is slightly misleading, since Ren and Candès (2020) had a very different setting and motivation, namely to incorporate side information for an "adaptive" extension of knockoffs. Indeed, Ren and Candès (2020) first compute LCD statistics and then introduce an auxiliary model to compute the probability a LCD statistic is positive given its magnitude, instead of computing the actual posterior of $W^\star \mid D$ as we do. Second, Proposition 2.3 is similar to Theorem 5 of Katsevich and Ramdas (2020), who show that the *unmasked* likelihood statistic maximizes $\mathbb{P}(W_j > 0)$; indeed, we see our work as building on theirs. However, there are two important differences. To start, the unmasked likelihood statistic does not satisfy pairwise exchangeability even though it is symmetric under the null (see Appendix C): in contrast, MLR statistics provably control the FDR. Second and more importantly, unlike in Katsevich and Ramdas (2020), MLR statistics are associated with guarantees on their *magnitudes*, allowing us to show much stronger theoretical results, as we do in the next section.

2.4 MLR statistics are asymptotically optimal

We now show that MLR statistics asymptotically maximize the expected number of discoveries. To start, it is worth noting that when the signs of $W^\star$ are independent conditional on D, $W^\star$ is *exactly* optimal in finite samples. This result is stated below—note that the following proposition is a generalization of Proposition 2 in Li and Fithian (2021), and the proofs of these propositions are effectively identical.

Proposition 2.4. *When $\{\mathbb{I}(W_j^\star > 0)\}_{j=1}^p$ are conditionally independent given D, then the MLR statistics $W^\star$ maximize the expected number of true discoveries in finite samples. Furthermore, if $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$ and π is a point mass on the true value of β, then $\{\mathbb{I}(W_j^\star > 0)\}_{j=1}^p$ are conditionally independent given D whenever $\mathbf{X}$ are fixed-X knockoffs or Gaussian conditional model-X knockoffs (Huang and Janson, 2020).*

Unfortunately, absent independence conditions, computing the exact Bayes-optimal statistic requires solving an intractable combinatorial optimization problem. However, we now show that under mild regularity conditions, accounting for such dependencies does not meaningfully increase power *asymptotically*, and thus MLR statistics are asymptotically optimal.

To this end, consider any asymptotic regime where we observe $\mathbf{X}^{(n)} \in \mathbb{R}^{n \times p_n}, \mathbf{y}^{(n)} \in \mathbb{R}^n$ and construct knockoffs $\widetilde{\mathbf{X}}^{(n)}$. For each n, let $L(\mathbf{y}^{(n)}; \mathbf{X}^{(n)}, \theta^{(n)})$ denote the likelihood of $\mathbf{y}^{(n)}$ given $\mathbf{X}^{(n)}$ and (unknown) parameters $\theta^{(n)}$ and let $\pi^{(n)}$ be a prior distribution on $\theta^{(n)}$. Let $D^{(n)}$ denote the masked data for knockoffs as defined in Section 2.2. We will analyze the limiting *empirical power* of a sequence of feature-statistics $W^{(n)} = w_n([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y}^{(n)})$, defined as the expected number of

discoveries normalized by the expected number of non-nulls $\kappa^{(n)}$. Formally, let $S^{(n)}(q)$ denote the rejection set of $W^{(n)}$ when controlling the FDR at level q. Then we define

$$\widetilde{\text{Power}}_q(w_n) = \frac{\mathbb{E}[|S^{(n)}(q)|]}{\kappa^{(n)}}, \tag{2.7}$$

where the expectation in the numerator is over $\mathbf{y}^{(n)}, \mathbf{X}^{(n)}, \theta^{(n)}$ given D. This notion of power is slightly unconventional, as it counts the number of discoveries instead of the number of *true* discoveries. Thus, we pause to make two remarks. First, in Appendix A.4, we introduce a modification of MLR statistics and give a proof sketch that this modification maximizes the expected number of *true* discoveries: however, computing this modification is prohibitively expensive. Since MLR statistics perform very well anyway, the difference between these procedures does not justify the cost in computation, and thus we choose to analyze $\widetilde{\text{Power}}$ as stated above. Second, we might hope that there is not too much difference between these metrics of power anyway. Indeed, let $T^{(n)}(q)$ and $V^{(n)}(q)$ count the number of true and false discoveries, and define the true positive rate $\text{TPR} = \frac{\mathbb{E}[T^{(n)}(q)]}{\kappa^{(n)}}$ and the marginal FDR $\text{mFDR} \coloneqq \frac{\mathbb{E}[V^{(n)}(q)]}{\mathbb{E}[|S^{(n)}(q)|]}$. Then we can write $\widetilde{\text{Power}} = \text{TPR}(1 - \text{mFDR})$: since knockoffs provably control the (non-marginal) FDR at level q, we might hope intuitively that $\widetilde{\text{Power}}$ is a good proxy for the TPR. That said, this is not a formal argument, so we use the tilde above $\widetilde{\text{Power}}$ to mark the difference between our measure of power and the conventional one.

As stated so far, this is a completely general asymptotic regime: we have made no assumptions whatsoever about the distribution of $\mathbf{y}^{(n)}, \mathbf{X}^{(n)}, \pi^{(n)}$, or the form of the likelihood. To show that MLR statistics maximize the expected number of discoveries, the main condition we need is that conditional on the masked data $D^{(n)}$, the signs of $W^\star$ are not too strongly dependent, and thus the successive averages of $\text{sign}(W^\star)$ converge to their conditional means. We note that the concept of a local dependence condition has appeared before in the multiple testing literature (Genovese and Wasserman, 2004; Storey et al., 2004; Ferreira and Zwinderman, 2006; Farcomeni, 2007), suggesting that this assumption is plausible. (That said, our proof technique is novel.) We will give further justification for this assumption in a moment—however, we first state Theorem 2.1, our main theoretical result.

Theorem 2.1. *Consider any arbitrarily high-dimensional asymptotic regime where we observe data* $\mathbf{X}^{(n)} \in \mathbb{R}^{n \times p_n}, \mathbf{y}^{(n)} \in \mathbb{R}^n$ *and construct knockoffs* $\widetilde{\mathbf{X}}^{(n)}$ *with* $D^{(n)}$ *denoting the masked data.Let* $W^\star = w_n^\star([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y}^{(n)})$ *denote the MLR statistics with respect to a prior* $\pi^{(n)}$ *on the parameters* $\theta^{(n)}$*. Let* $W = w_n([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y}^{(n)})$ *denote any other sequence of feature statistics.*

Assume only the following conditions. First, assume that $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star)$ *and* $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n)$ *exist for each* $q \in (0, 1)$*. Second, assume that* $\kappa^{(n)}$*, the expected number of non-nulls, grows faster than* $\log(p_n)^4$*, i.e.,* $\lim_{n\to\infty} \frac{\kappa^{(n)}}{\log(p_n)^4} = \infty$*. Finally, assume that conditional on* $D^{(n)}$*, the covariance between the signs of* $W^\star$ *decays exponentially. That is, there exist constants* $C \geq 0, \rho \in (0, 1)$ *such that*

$$|\operatorname{Cov}(\mathbb{I}(W_i^\star > 0), \mathbb{I}(W_j^\star > 0) \mid D^{(n)})| \leq C\rho^{|i-j|}. \tag{2.8}$$

Then for all but countably many $q \in (0, 1)$*,*

$$\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star) \geq \lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n). \tag{2.9}$$

Theorem 2.1 tells us that MLR statistics asymptotically maximize the number of expected discoveries in great generality: for example, we place no explicit restrictions on the relationship between $\mathbf{y}$

and $\mathbf{X}$ or the dimensionality. Indeed, the first two assumptions are quite weak: the first assumption merely guarantees that the limiting powers we aim to study actually exist, and the second is only a mild restriction on the sparsity regime. Indeed, our setting allows for many previously studied sparsity regimes, such as a polynomial sparsity model (Donoho and Jin, 2004; Ke et al., 2020) and the linear sparsity regime (Weinstein et al., 2017). Nonetheless, these assumptions can be substantially weakened at the cost of a more technical theorem statement, as discussed in Appendix A.3.

Equation (2.8), which asks that $\text{sign}(W^\star)$ is only locally dependent, is a stronger assumption. That said, there are a few reasons to think it is not too strong. First, this is (roughly speaking) a *checkable* condition. Indeed, computing MLR statistics usually requires sampling from the joint distribution of $\text{sign}(W^\star)$ conditional on D. This means that we can actually inspect $\text{Cov}(\text{sign}(W^\star) \mid D)$ in finite samples, allowing us to diagnose whether $\text{sign}(W^\star)$ are only locally dependent. Indeed, in Section 3, we will see that even when $\mathbf{X}$ is extremely highly correlated, the signs of $W^\star$ are nearly independent anyway. Furthermore, we expect the conclusion to be robust to mild violations of Equation (2.8), since from a theoretical perspective, the conclusion only requires that the successive averages of $\text{sign}(W^\star)$ obey a law of large numbers conditional on D (see Appendix A.3 for discussion). Thus, it should be possibly to heuristically assess whether the local dependence condition holds for any given problem.

Second, in some restricted settings, we can prove the local dependence condition. For example, Equation (2.8) holds in the special case where $\mathbf{y}, \mathbf{X}$ follow a Gaussian linear model and $\mathbf{X}^T\mathbf{X}$ is block-diagonal, similar to the setting of Ke et al. (2020), as stated below.

Proposition 2.5. *Suppose* $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$ *and* $\mathbf{X}^T\mathbf{X}$ *is a block-diagonal matrix with maximum block size* $M \in \mathbb{N}$. *Suppose* π *is any prior such that the coordinates of* β *are a priori independent and* σ^2 *is a known constant. Then if* $\widetilde{\mathbf{X}}$ *are either fixed-X knockoffs or conditional Gaussian model-X knockoffs (Huang and Janson, 2020), the coordinates of* $\text{sign}(W^\star)$ *are* M*-dependent conditional on* D, *implying that Equation (2.8) holds, e.g., with* $C = 2^M$ *and* $\rho = \frac{1}{2}$.

The block-diagonal assumption in Proposition 2.5 is quite restrictive, but intuitively, we expect that Equation (2.8) should hold whenever $\mathbf{X}$ exhibits only local dependencies. For example, if $\mathbf{X}_i, \mathbf{X}_j$ represent genetic variants on the genome which are very far apart, they are likely to be mostly uncorrelated with each other and thus the sign of $W_i^\star$ is probably not too dependent on the sign of $W_j^\star$. Indeed, we verify this intuition empirically in Section 3, where we show that the coordinates of $\mathbb{I}(\text{Cov}(W^\star) > 0)$ are almost completely conditionally uncorrelated even in settings where $\mathbf{X}$ is extremely highly correlated.

Perhaps the weakest aspect of Theorem 2.1 is its conclusion: MLR statistics are only average-case optimal with respect to a user-specified prior π on the parameters θ. If the true parameters θ are very different from the average case specified by the prior, MLR statistics may not perform well. With this motivation, in the next section, we suggest practical choices of π which performed well empirically, even when they were very poorly specified.

2.5 Computing MLR statistics

2.5.1 General strategy

In this section, we discuss how to compute $W_j^\star = \log\left(\frac{\mathbb{E}_{\theta\sim\pi}[L_\theta(\mathbf{X}_j=\mathbf{x}_j|D)]}{\mathbb{E}_{\theta\sim\pi}[L_\theta(\mathbf{X}_j=\widetilde{\mathbf{x}}_j|D)]}\right)$ under the assumption that for any fixed value of θ, we can compute the true likelihood $L_\theta(\mathbf{y} \mid \mathbf{X})$. The challenge of computing

$W_j^\star$ is that we must marginalize over both θ *and* the unknown values of $\mathbf{X}_{-j}$, as we only observe the unordered pairs $\{\mathbf{x}_{j'}, \widetilde{\mathbf{x}}_{j'}\}$ for $j' \neq j$.

This suggests a general approach based on Gibbs sampling, as described by Algorithm 1. Informally, let θ_j denote the coordinates of θ which determine whether or not $\mathbf{X}_j$ is non-null: for example, in a linear regression where $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$, θ_j corresponds to β_j. The algorithm is then as follows. First, after initializing values for θ and $\mathbf{X}$, we iteratively resample $\mathbf{X}_j \mid \mathbf{y}, \mathbf{X}_{-j}, \theta, D$ and $\theta_j \mid \mathbf{y}, \mathbf{X}, \theta_{-j}$. Resampling θ_j can be done using any off-the-shelf Bayesian Gibbs sampler, since this step is identical to a typical Bayesian regression problem. Resampling $\mathbf{X}_j \mid \mathbf{y}, \mathbf{X}_{-j}, \theta, D$ is also straightforward, since conditional on D, $\mathbf{X}_j$ must take one of the two values $\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}$. In particular,

$$\frac{\mathbb{P}(\mathbf{X}_j = \mathbf{x}_j \mid \mathbf{y}, \mathbf{X}_{-j}, \theta, D)}{\mathbb{P}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid \mathbf{y}, \mathbf{X}_{-j}, \theta, D)} = \frac{\pi(\theta)\mathbb{P}(\mathbf{X}_j = \mathbf{x}_j, \widetilde{\mathbf{X}}_j = \widetilde{\mathbf{x}}_j \mid \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}) L_\theta(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}_j, \mathbf{X}_{-j}, D)}{\pi(\theta)\mathbb{P}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j, \widetilde{\mathbf{X}}_j = \mathbf{x}_j \mid \mathbf{X}_{-j}, \widetilde{\mathbf{X}}_{-j}) L_\theta(\mathbf{y} \mid \mathbf{X}_j = \widetilde{\mathbf{x}}_j, \mathbf{X}_{-j}, D)} \tag{2.10}$$

$$= \frac{L_\theta(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}_j, \mathbf{X}_{-j}, D)}{L_\theta(\mathbf{y} \mid \mathbf{X}_j = \widetilde{\mathbf{x}}_j, \mathbf{X}_{-j}, D)}, \tag{2.11}$$

where the second step uses the pairwise exchangeability of $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$. The resulting "meta-algorithm" is summarized below in Algorithm 1. Of course, there are many natural variants of Algorithm 1: in particular, our implementations modify lines 4 and 5 of Algorithm 1 to marginalize over the possible values of θ_j when resampling from $X_j \mid X_{-j}, \mathbf{y}, \theta_{-j}, D$, which helps the Markov chain converge more quickly.

Algorithm 1 Gibbs sampling meta-algorithm to compute MLR statistics.

Input: Masked data $D = (\mathbf{y}, \{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}_{j=1}^p)$, a likelihood $L_\theta(\mathbf{y} \mid \mathbf{X})$ and a prior π on θ.

1: Initialize $\mathbf{X}_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\})$ and $\theta \sim \pi$.
2: **for** $i = 1, 2, \ldots, n_{\text{sample}}$ **do**
3: **for** $j = 1, \ldots, p$ **do**:
4: Set $\eta_j^{(i)} = \log\left(L_\theta(\mathbf{y} \mid \mathbf{X}_{-j}, \mathbf{X}_j = \mathbf{x}_j)\right) - \log\left(L_\theta(\mathbf{y} \mid \mathbf{X}_{-j}, \mathbf{X}_j = \widetilde{\mathbf{x}}_j)\right)$.
5: Sample $\mathbf{X}_j \mid \mathbf{X}_{-j}, \mathbf{y}, \theta, D$, using Equation (2.11) and $\eta_j^{(i)}$.
6: Sample $\theta_j \mid \mathbf{y}, \mathbf{X}, \theta_{-j}$.
7: Resample any hyperparameters (e.g., σ^2 in linear regression).
8: For each j, return $W_j^\star = \log\left(\sum_{i=1}^{n_{\text{sample}}} \frac{\exp(\eta_j^{(i)})}{1+\exp(\eta_j^{(i)})}\right) - \log\left(\sum_{i=1}^{n_{\text{sample}}} \frac{1}{1+\exp(\eta_j^{(i)})}\right)$.

In a moment, we will briefly describe a few particular modeling choices which performed well in a variety of settings. Before doing this, it is important to note that in Gaussian linear models where there is a sparse "spike-and-slab" prior on the coefficients β, Algorithm 1 is similar in flavor to the "Bayesian Variable Selection" (BVS) feature statistic from Candès et al. (2018), although there are substantial differences in the Gibbs sampler and some differences in the final estimand. Broadly, we see our work as complimentary to theirs; however, aside from technical details, a main difference is that Candès et al. (2018) seemed to argue that the main advantage of BVS was to incorporate accurate prior information. In contrast, we argue that a main advantage of MLR statistics is that they are estimating the right *estimand*, and thus MLR statistics can be very powerful even when using mispecified priors (see Section 3). Of course, BVS statistics are only similar to a single instantiation of MLR statistics, although admittedly a very important one!

2.5.2 Sparse priors for generalized additive models and binary GLMs

We start by considering a generalized additive model of the form

$$Y \mid X \sim \mathcal{N}\left(\sum_{j=1}^{p} \phi_j(X_j)^T \beta^{(j)}, \sigma^2\right) \tag{2.12}$$

where $\phi_j : \mathbb{R} \to \mathbb{R}^{d_j}$ is a prespecified set of basis functions and $\beta^{(j)} \in \mathbb{R}^{p_j}$ are linear coefficients. For example, in a Gaussian linear model, ϕ_j is the identity function so $\mathbb{E}[Y \mid X] = X^T\beta$. More generally, in Section 3, we will take $\phi_j(\cdot)$ to be the basis representation of d-degree regression splines with K knots (see Hastie et al. (2001) for a review). We picked this model because it can flexibly model nonlinear relationships between $\mathbf{y}$ and $\mathbf{X}$ while also allowing efficient computation of MLR statistics.

For the prior, we assume $\beta^{(j)} = 0 \in \mathbb{R}^{p_j}$ a priori with probability p_0, and otherwise $\beta^{(j)} \sim \mathcal{N}(0, \tau^2 I_{p_j})$. This group-sparse prior is effectively a "two-groups" model, as $\mathbf{X}_j$ is null if and only if $\beta^{(j)} = 0$. Since the sparsity p_0 and signal size τ^2 are typically not known a priori, we use the conjugate hyperpriors $\tau^2 \sim \text{invGamma}(a_\tau, b_\tau), \sigma^2 \sim \text{invGamma}(a_\sigma, b_\sigma)$ and $p_0 \sim \text{Beta}(a_0, b_0)$. As we will see in Section 3, using these hyperpriors will allow us to adaptively estimate the sparsity level. Using standard techniques for sampling from "spike-and-slab" models (George and McCulloch, 1997), we can compute MLR statistics in at most $O(n_{\text{iter}}np)$ operations (assuming that the dimensionality of ϕ_j is fixed). This complexity is comparable to the cost of fitting the LASSO, which is roughly $O(n_{\text{iter}}np)$ using coordinate descent or $O(np^2)$ using the LARS algorithm (Efron et al., 2004), and it is usually faster than the cost of computing Gaussian model-X or fixed-X knockoffs (which is usually $O(np^2 + p^3)$). Please see Appendix D for a detailed derivation of the Gibbs updates for this model.

Lastly, we can easily extend this algorithm to binary responses. In particular, using standard data augmentation strategies developed in Albert and Chib (1993); Polson et al. (2013), we can compute Gibbs updates in the same computational complexity when $\mathbb{P}(Y = 1 \mid X) = s\left(\sum_{j=1}^{p} \phi_j(X_j)^T \beta^{(j)}\right)$, where s is either the probit or logistic link function. See Appendix D for further details.

3 Simulations

In this section, we show via simulations that MLR statistics are powerful in a variety of settings. At the outset, we emphasize that in every simulation in this section, MLR statistics do *not* have accurate prior information: indeed, we use exactly the same hyperpriors to compute MLR statistics in every single plot. Furthermore, we consider settings where $\mathbf{X}$ is highly correlated in order to test whether MLR statistics perform well even when the "local dependence" assumption from Theorem 2.1 may not hold. Nonetheless, the MLR statistics perform very well, suggesting that they are robust to misspecification of the prior and strong dependencies in $\mathbf{X}$.

Throughout this section, we control the FDR at level $q = 0.05$ unless otherwise specified. All plots have two standard-deviation error bars, although sometimes the bars are so small they are not visible. In each plot in this section, knockoffs provably control the FDR, so we only plot power. All simulation code is available at `https://github.com/amspector100/mlr_knockoff_paper`.

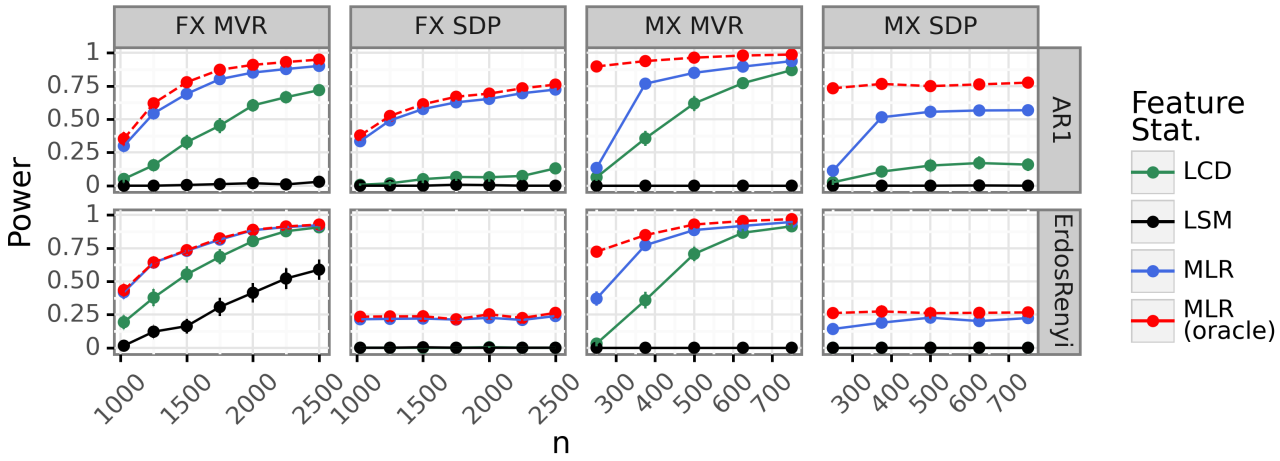

Figure 3: Power of MLR, LCD, and LSM statistics in a sparse Gaussian linear model with $p = 500$ and 50 non-nulls. Note that when computing fixed-X knockoffs, MLR statistics almost exactly match the power of the oracle procedure which provably upper bounds the power of any knockoff feature statistic. For MX knockoffs, MLR statistics are slightly less powerful than the oracle, although they are still very powerful compared to the lasso-based statistics. Note that perhaps surprisingly, the power of knockoffs can be roughly constant in n in the "SDP" setting: this is because SDP knockoffs sometimes have identifiability issues (Spector and Janson, 2022). See Appendix F for precise simulation details.

3.1 Gaussian linear models

In this section, we assume $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, I_n)$ for sparse β. We draw $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ for two choices of Σ: by default, we take Σ to correspond to a highly correlated nonstationary AR(1) process, inspired by real design matrices in genetic studies. However, we also analyze a setting where Σ is 80% sparse with 20% of its entries drawn to be nonzero uniformly at random, in accordance with an ErdosRenyi (ER) procedure, so in this setting $\mathbf{X}$ does not exhibit local dependencies. We also compute both "SDP" and "MVR" knockoffs (Spector and Janson, 2022; Candès et al., 2018) to show that MLR statistics perform well in both cases. Please see Appendix F for specific simulation details.

We compare four types of feature statistics. First, we compute MLR statistics using the identity basis functions as specified in Section 2.5—as this is our default choice, and in plots, "MLR" refers to this version of MLR statistics. Second, we compute the "gold-standard" LCD and LSM statistics as described in Section 2.1. Lastly, we compute the oracle MLR statistics—note that for fixed-X knockoffs, the oracle MLR statistics are equivalent to "knockoff$\star$" procedure from Li and Fithian (2021), which is a provable finite-sample upper-bound on the power of *any* knockoff feature statistic. Figure 3 shows the results while varying n in both a low-dimensional setting (using fixed-X knockoffs) and high-dimensional setting (using MX knockoffs). It shows that MLR statistics are substantially more powerful than the lasso-based statistics and, in the fixed-X case, MLR statistics almost perfectly match the power of the oracle procedure. Indeed, this result holds even for the "ErdosRenyi" covariance matrix, where $\mathbf{X}$ exhibits strong non-local dependencies (in contrast to the theoretical assumptions in Theorem 2.1). Furthermore, 4 shows that MLR statistics are quite efficient, often faster than computing a cross-validated lasso in the model-X case, and comparable to the cost of computing knockoffs in the fixed-X case.

Next, we analyze the performance of MLR statistics in a setting where the prior is very mispecified.

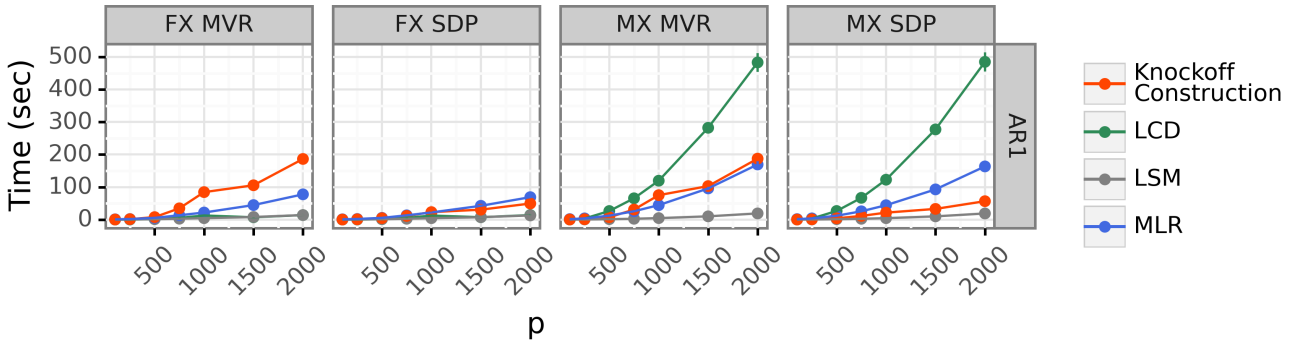

Figure 4: This figure shows the computation time for various feature statistics in the same setting as Figure 3, as well as the cost of computing knockoffs. It shows that MLR statistics are competitive with state-of-the-art feature statistics (in the model-X case) or comparable to the cost of computing knockoffs (in the fixed-X case).

In Figure 5, we vary the sparsity (proportion of non-nulls) between 5% and 40%, and we consider settings where (i) the non-null coefficients are heavy-tailed and drawn as i.i.d. Laplace random variables and (ii) the non-nulls are "light-tailed" and drawn as i.i.d. $\text{Unif}([-1/2, -1/4] \cup [1/4, 1/2])$ random variables. In both settings, the MLR prior assumes the non-null coefficients are i.i.d. $\mathcal{N}(0, \tau^2)$. Nonetheless, as shown by Figure 5, MLR statistics still consistently outperform the lasso-based statistics and nearly match the performance of the oracle. This result is of particular interest in the fixed-X setting, where one cannot use cross validation to adaptively tune hyperparameters to the sparsity level—in contrast, MLR statistics match the power of the oracle for all sparsity levels.

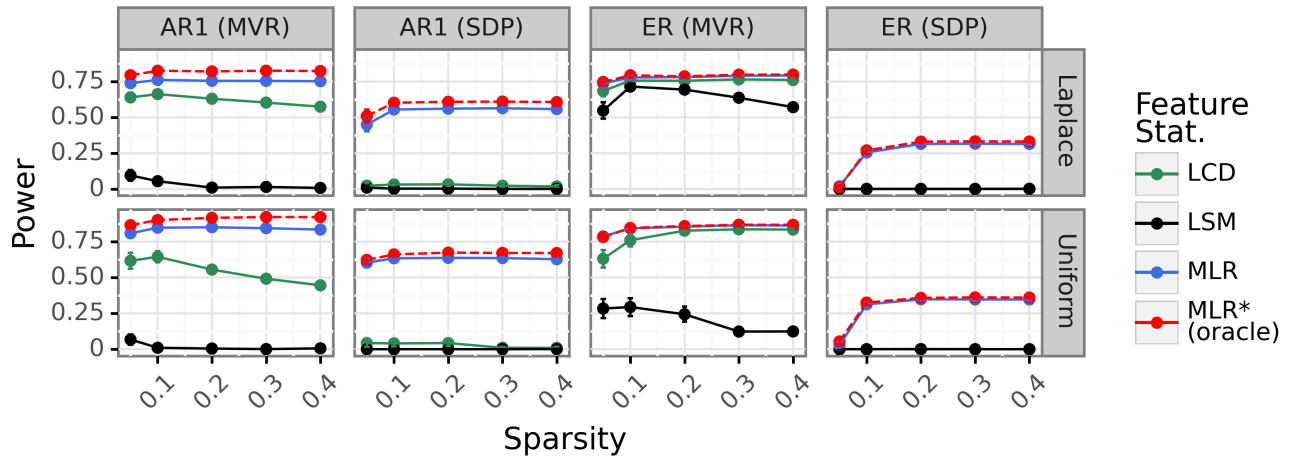

Figure 5: This figure shows the power of MLR, LCD, and LSM statistics when varying the sparsity level and drawing the non-null coefficients from a heavy-tailed (Laplace) and light-tailed (Uniform) distribution, with $p = 500$ and $n = 1250$. The setting is otherwise identical to the AR1 setting from Figure 3. It shows that the MLR statistics perform well despite using the same (misspecified) prior in every setting.

Lastly, we verify that the local dependence condition we assumed in Theorem 2.1 holds empirically. We consider the AR(1) setting but modify the parameters so that $\mathbf{X}$ is extremely highly correlated, with adjacent correlations drawn as i.i.d. $\text{Beta}(50, 1)$ variables. We also consider a setting where $\mathbf{X}$ is equicorrelated with correlation 95%. In both settings, Figure 6 shows that the signs of $W^\star$ are essentially completely uncorrelated conditional on D despite the extreme correlation in $\mathbf{X}$, supporting the local dependence assumption from Section 2.

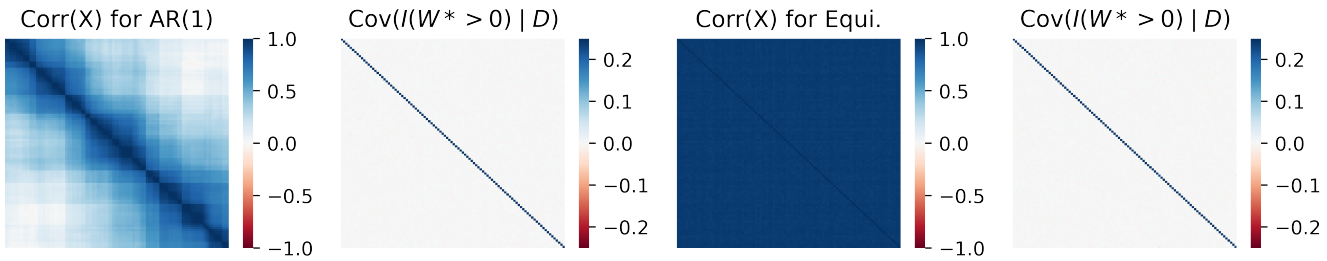

Figure 6: In the AR(1) and equicorrelated settings, this plot shows both the correlation matrix of $\mathbf{X}$ as well as the conditional covariance of signs of the MLR statistic $W^\star$, computed using the Gibbs sampler for the MLR statistic using model-X knockoffs. It shows that even when $\mathbf{X}$ is very highly correlated, the signs of $W^\star$ are essentially uncorrelated.

3.2 Generalized additive models

We now consider generalized additive models, where $Y \mid X \sim \mathcal{N}(h(X)^T\beta, \epsilon)$ for some non-linear function h applied element-wise to X. We run simulations in the AR(1) setting from Section 3.1 with four choices of h: $h(x) = \sin(x), h(x) = \cos(x), h(x) = x^2, h(x) = x^3$. We compare six feature statistics: the linear MLR statistics, MLR based on cubic regression splines with one knot, the LCD, a random forest with swap importances as in Gimenez et al. (2019), and DeepPINK (Lu et al., 2018), a feature statistic based on a feedforward neural network. We note that this setting is much more challenging than the linear regression setting, since these feature statistics must learn (or approximate) the function h in addition to estimating the coefficients β. For this reason, our simulations in this section are low-dimensional with $n > p$, and we should not expect any feature statistic to match the performance of the oracle MLR statistics.

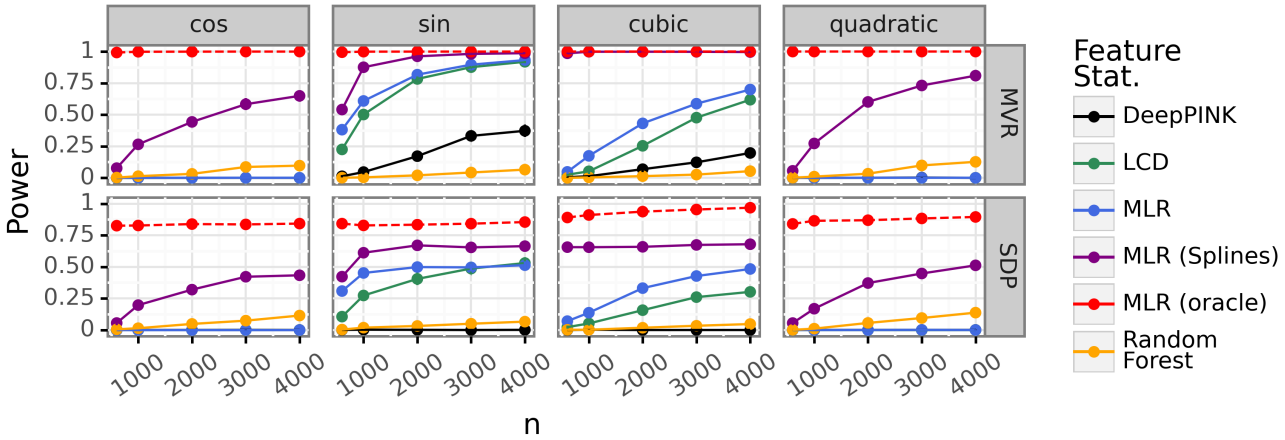

Figure 7: This plot shows the power of MLR statistics and competitors in generalized additive models where $Y \mid X \sim \mathcal{N}(h(X)^T\beta, \epsilon)$, for h applied elementwise to X. The x-facets show the choice of h, and the y-facets show results for both MVR and SDP knockoffs; note $p = 200$ and there are 60 non-nulls. For this plot only, we choose $q = 0.1$ because several of the competitor statistics made almost no discoveries at $q = 0.05$. See Appendix F for the corresponding plot with $q = 0.05$.

Figure 7 shows that the "MLR (splines)" feature statistic uniformly outperforms every other feature statistic, often by wide margins. Note that the linear MLR and LCD statistics are powerless in the cos and quadratic settings, since in this case h is an even function and thus the non-null features

have no linear relationship with the response. However, when in the sin and cubic settings, the linear MLR statistics outperform the LCD statistics, suggesting that even when the linear model is misspecified, linear MLR statistics can be powerful as long as there is some linear effect. That said, if one believes $\mathbf{y}$ may have a highly nonlinear relationship with $\mathbf{X}$, the we recommend using the MLR statistic based on splines. As expected, no statistic matches the power of the oracle MLR statistic.

3.3 Logistic regression

Lastly, we now consider the setting of logistic regression, so $Y \mid X \sim \text{Bern}(s(X^T\beta))$ where s is the sigmoid function. We run the same simulation setting as Figure 3, except that now Y is binary and we consider low-dimensional settings, since inference in logistic regression is generally more challenging than in linear regression. The results are shown by Figure 8, which shows that MLR statistics generally perform better than the LCD, although there is a substantial gap between the performances of MLR statistics and the oracle MLR statistics. That said, it is worth noting that the MLR statistics take roughly three times longer to compute than the LCD in this setting due to the data-augmentation step in the Gibbs sampler for binary regression (see Appendix D). It may be possible to speed up the Gibbs sampler, but we leave this possibility to future work.

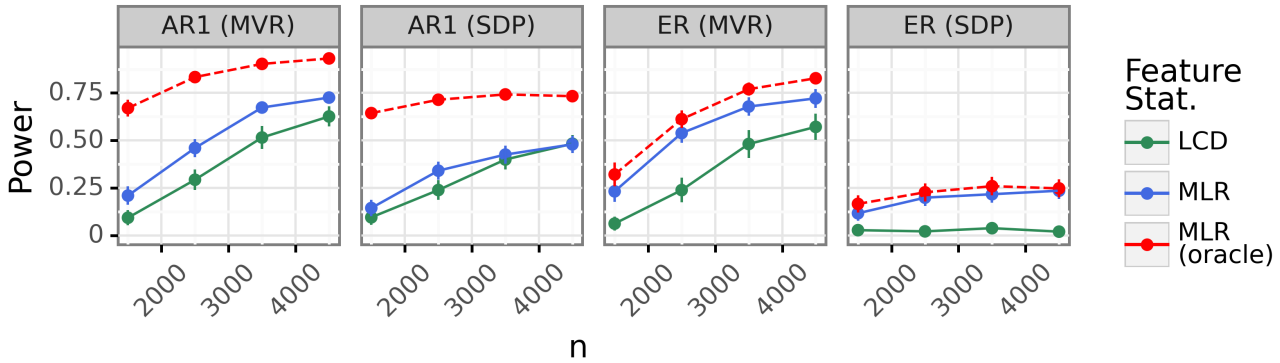

Figure 8: This plot shows the power of MLR statistics compared to the cross-validated LCD in logistic regression, with $p = 500$, 50 non-nulls, and n varied between 1500 and 4500. The setting is otherwise identical to that of Figure 3.

4 Real applications

In this section, we apply MLR statistics to three real datasets which have been previously analyzed using knockoffs. In each case, MLR statistics have comparable or higher power than competitor statistics. All code and data are available at `https://github.com/amspector100/mlr_knockoff_paper`.

4.1 HIV drug resistance

We begin with the HIV drug-resistance dataset from Rhee et al. (2006), which was previously analyzed by (e.g.) Barber and Candès (2015) using knockoffs. The dataset consists of genotype

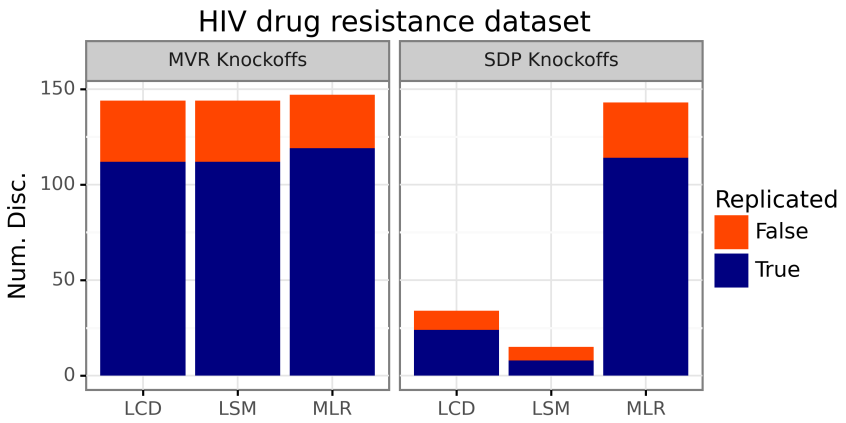

Figure 9: This figure shows the total number of discoveries made by the LCD, LSM, and MLR feature statistics in the HIV drug-resistance dataset from Rhee et al. (2006), summed across all 21 drugs.

data from roughly 750 HIV samples as well as drug resistance measurements for 21 different drugs, and the goal of our analysis is to discover genetic variants which affect drug resistance for each of the drugs. An advantage of this dataset is that Rhee et al. (2005) published treatment-selected mutation panels for exactly this setting, so we can check whether any discoveries made by knockoffs are corroborated by this separate analysis.

We preprocess and model the data in exactly the same way as Barber and Candès (2015), and following Barber and Candès (2015), we apply fixed-X knockoffs with LCD, LSM, and MLR statistics and FDR level $q = 0.05$. For both MVR and SDP knockoffs, Figure 9 shows the total number of discoveries made by each statistic, stratified by whether each discovery is corroborated by Rhee et al. (2005). For SDP knockoffs, the MLR statistics make nearly an order of magnitude more discoveries than the competitor methods with a comparable corroboration rate. For MVR knockoffs, all three statistics perform roughly equally well, although MLR statistics make slightly more discoveries with a slightly higher corroboration rate. Overall, in this setting, MLR statistics are competitive with and sometimes substantially outperform the lasso-based statistics. See Appendix G for specific results for each drug.

4.2 Financial factor selection

Next, we consider a "fund replication" dataset inspired by Challet et al. (2021). In finance, analysts often aim to select a few key factors which drive the performance of an asset such as an index fund or hedge fund. Challet et al. (2021) applied knockoffs to factor selection, and as a benchmark, they applied fixed-X knockoffs with the LCD to test which US equities explained the performance of an index fund for the energy sector (ticker name XLE). Since the XLE index fund is essentially a weighted combination of a known list of US equities, Challet et al. (2021) were able to tell whether each discovery made by knockoffs was a true or false positive.

We perform the same analysis for ten index funds describing key sectors of the US economy, including index funds for energy (XLE), technology (XLK), real estate (XLRE), and (see Appendix G for the full list). In our analysis, $\mathbf{y}$ is the daily log return of the index fund and $\mathbf{X}$ consists of the daily log returns of each stock in the S&P 500 since 2013, so $p \approx 500$ and $n \approx 2300$, although for several of the index funds, fewer than ten years of data is available. We compute fixed-X knockoffs using both MVR and SDP knockoffs and apply LCD, LSM, and MLR statistics. Figure 10 shows the total

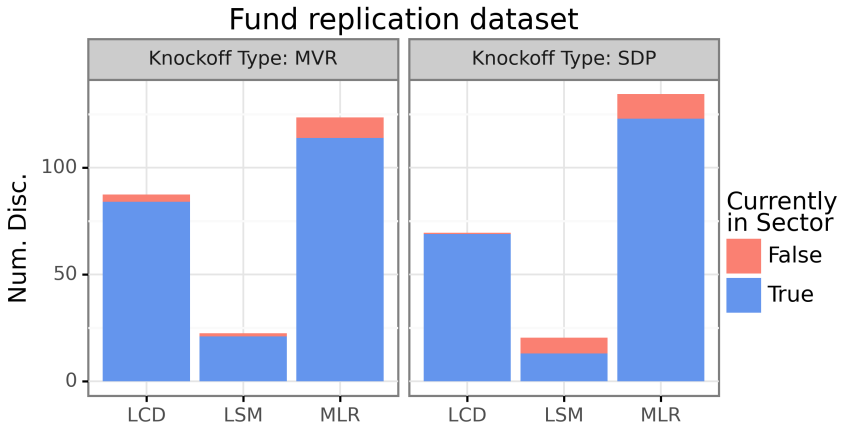

Figure 10: This figure shows the total number of discoveries made by each method in the fund-replication dataset inspired by Challet et al. (2021), summed across all ten index funds. See Appendix G for a table showing that the average FDP for each method is below the nominal level of $q = 0.05$.

number of true and false discoveries summed across all ten index funds with FDR level $q = 0.05$. In particular, Figure 10 shows that MLR statistics make substantially more discoveries than either the LCD or LSM statistics: indeed, MLR statistics make 35% and 78% more discoveries than the LCD statistic for MVR and SDP knockoffs (respectively), and the LSM statistic makes more than 5 times fewer discoveries than the MLR statistics. We also note that the FDP (averaged across all ten index funds) is well below 5% for each method analyzed in Figure 10, as shown in Appendix G. Thus, MLR statistics substantially outperform the lasso-based statistics in this analysis.

4.3 Graphical model discovery for gene networks

Lastly, we consider the problem of recovering a gene network from single-cell RNAseq data. Our analysis is inspired by that of Li and Maathuis (2019), who analyze a single-cell RNAseq dataset from Zheng et al. (2017) by modeling the gene expression log-counts as a Gaussian graphical model (see Li and Maathuis (2019) for justification of the Gaussian assumption and the precise data processing steps). In particular, Li and Maathuis (2019) developed an extension of fixed-X knockoffs to detect edges in Gaussian graphical models while controlling the false discovery rate across discovered edges, and they applied this to the RNAseq data for the 50 genes with the highest variance Zheng et al. (2017).

We analyze the same dataset using the same model, but we instead compare the performance of LCD, LSM, and MLR statistics for both MVR and SDP fixed-X knockoffs. Since no methods made any discoveries at level $q = 0.05$, we plot the number of discoveries as a function of q, which we vary between 0 and 0.5. Figure 11 shows the results, namely that the MLR statistics make the most discoveries for nearly every value of q, although often by a relatively small margin. That said, for small values of q, the LSM statistic performs poorly, and for large values of q, the LCD statistic performs poorly, whereas the MLR statistic is the most consistently powerful.

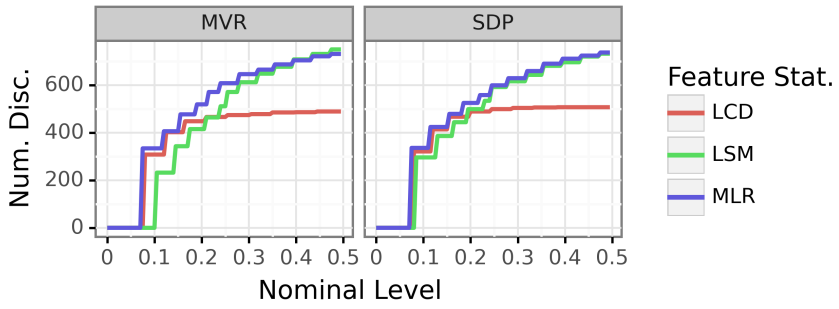

Figure 11: This figure shows the number of discoveries made by LCD, LSM, and MLR statistics when used to detect edges in a Gaussian graphical model for gene expression data, as in Li and Maathuis (2019).

5 Discussion

This paper introduced masked likelihood ratio statistics, a class of knockoff statistics which are provably asymptotically optimal under mild regularity conditions. We show in simulations and three data applications that in addition to having appealing theoretical properties, MLR statistics are efficient and powerful in practical applications. However, our work leaves open several possible directions for future research.

- Our theory shows that MLR statistics are asymptotically *average-case* optimal over a user-specified prior on the unknown parameters. However, it might be worthwhile to develop *minimax-optimal* knockoff-statistics, e.g., by computing a "least-favorable" prior.

- A limitation of our theory is that it requires a "local dependency" condition which is challenging to verify analytically, although our local dependency condition can be diagnosed using the data at hand. It might be interesting to investigate (i) precisely when this local dependency condition holds and (ii) whether MLR statistics are still in any way optimal when it fails to hold.

- Our paper only considers a few instantiations of MLR statistics designed for linear models, generalized additive models, and binary generalized linear models. However, other classes of MLR statistics could be more powerful or more computationally efficient. For example, it is not clear how to extend the MLR statistics in Section 2.5 to allow them to detect interactions between features without substantially increasing the computational burden. Similarly, it might be worthwhile to improve the efficiency of MLR statistics in the case where $\mathbf{y}$ is binary. We leave such questions to future work.

6 Acknowledgements

The authors would like to thank John Cherian, Kevin Guo, Lucas Janson, and Lihua Lei for valuable comments. A.S. is partially supported by the Two Sigma Graduate Fellowship Fund and a Graduate Research Fellowship from the National Science Foundation. W.F. is partially supported by the NSF DMS-1916220 and a Hellman Fellowship from Berkeley.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

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

Bates, S., Candès, E., Janson, L., and Wang, W. (2020). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 0(0):1–15.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577.

Challet, D., Bongiorno, C., and Pelletier, G. (2021). Financial factors selection with knockoffs: Fund replication, explanatory and prediction networks. *Physica A: Statistical Mechanics and its Applications*, 580:126105.

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, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In Balcan, M. F. and Weinberger, K. Q., editors, *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1851–1859, New York, New York, USA. PMLR.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962 – 994.

Doukhan, P. and Neumann, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407 – 499.

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.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics*, 34(2):275–297.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini–Hochberg method. *The Annals of Statistics*, 34(4):1827 – 1849.

Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035 – 1061.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Huang, D. and Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *Ann. Statist.*, 48(5):3021–3042.

Katsevich, E. and Ramdas, A. (2020). On the power of conditional independence testing under model-x.

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*.

Li, J. and Maathuis, M. H. (2019). Ggm knockoff filter: False discovery rate control for gaussian graphical models.

Li, X. and Fithian, W. (2021). Whiteout: when do fixed-x knockoffs fail?

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*, volume 32. 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.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ren, Z. and Candès, E. J. (2020). Knockoffs with side information. *Annals of Applied Statistics*.

Rhee, S.-Y., Fessel, W. J., Zolopa, A. R., Hurley, L., Liu, T., Taylor, J., Nguyen, D. P., Slome, S., Klein, D., Horberg, M., Flamm, J., Follansbee, S., Schapiro, J. M., and Shafer, R. W. (2005). Hiv-1 protease and reverse-transcriptase mutations: Correlations with antiretroviral therapy in subtype b isolates and implications for drug-resistance surveillance. *The Journal of Infectious Diseases*, 192(3):456–465.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360.

Sechidis, K., Kormaksson, M., and Ohlssen, D. (2021). Using knockoffs for controlled predictive biomarker identification. *Statistics in Medicine*, 40(25):5453–5473.

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.

Spector, A. and Janson, L. (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252 – 276.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Wang, W. and Janson, L. (2021). A high-dimensional power analysis of the conditional randomization test and knockoffs. *Biometrika*.

Weinstein, A., Barber, R. F., and Candès, E. J. (2017). A power analysis for knockoffs under Gaussian designs. *IEEE Transactions on Information Theory*.

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv*.

Zheng, G., Terry, J., Belgrader, P., Ryvkin, P., Bent, Z., Wilson, R., Ziraldo, S., Wheeler, T., McDermott, G., Zhu, J., Gregory, M., Shuga, J., Montesclaros, L., Underwood, J., Masquelier, D., Nishimura, S., Schnall-Levin, M., Wyatt, P., Hindson, C., Bharadwaj, R., Wong, A., Ness, K., Beppu, L., Deeg, H., McFarland, C., Loeb, K., Valente, W., Ericson, N., Stevens, E., Radich, J., Mikkelsen, T., Hindson, B., and Bielas, J. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*.

A Main proofs

In this section, we prove the main results of the paper.

A.1 Knockoffs as inference on masked data

We start by proving Propositions 2.1, 2.2, and a related corollary which will be useful when proving Theorem 2.1. It may be helpful to recall that throughout, we only consider feature statistics W which are nonzero with probability one, since such feature statistics are provably inadmissible (one can always make more discoveries by adding a tiny amount of noise to the entries of W which are zeros).

Proposition 2.1. *Let $\widetilde{\mathbf{X}}$ be model-X knockoffs such that $\mathbf{X}_j \neq \widetilde{\mathbf{X}}_j$ for $j \in [p]$. Then $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ is a valid feature statistic if and only if (i) $|W|$ is a function of the masked data D and (ii) there exist* distinguishing *functions $\widehat{\mathbf{X}}_j = g_j(D)$ such that $W_j > 0$ if and only if $\widehat{\mathbf{X}}_j = \mathbf{X}_j$.*

Proof. Forward direction: Suppose W is a valid feature statistic; we will now show conditions (i) and (ii). To show (i), note that observing $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p$ is equivalent to observing $[\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(J)}$ for some unobserved $J \subset [p]$ chosen uniformly at random, since we can always randomly assign $\mathbf{X}_j, \widetilde{\mathbf{X}}_j$ to the left and right slots, respectively. Define $[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}] \coloneqq [\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(J)}$ and let $W' = w([\mathbf{X}^{(1)}, \mathbf{X}^{(2)}], \mathbf{y})$. Then by the swap invariance property of knockoffs, we have that $|W| = |W'|$, and W' is a function of D: thus, $|W|$ is a function of D as desired.

To show (ii), we construct g_j as follows:

$$\widehat{\mathbf{X}}_j = g_j(\mathbf{y}, \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p) = g_j(\mathbf{y}, [\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(J)}) = \begin{cases} \mathbf{X}_j^{(1)} & W_j' > 0 \\ \mathbf{X}_j^{(2)} & W_j' < 0. \end{cases}$$

Note that g_j is well-defined (does not depend on J) because w is by definition antisymmetric. Indeed, there are two cases. In the first case, if $j \notin J$, we note (i) $\mathbf{X}_j^{(1)} = \mathbf{X}_j$ by definition of $\mathbf{X}^{(1)}$ and (ii) we have that $W_j' = W_j$ by the antisymmetry property of w. Thus $\widehat{\mathbf{X}}_j = \mathbf{X}_j^{(1)} = \mathbf{X}_j$ if and only if $W_j > 0$. The second case is analagous: if $j \in J$, then $W_j' = -W_j$, so $\widehat{\mathbf{X}}_j = \mathbf{X}_j^{(2)} = \mathbf{X}_j$ if and only if $W_j > 0$. In both cases, $W_j > 0$ if and only if $\widehat{\mathbf{X}}_j \neq \mathbf{X}_j$, proving (ii).

Backwards direction: To show $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ is a valid feature statistic, it suffices to show the flip-sign property, namely that $W' \coloneqq w([\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(J)}, \mathbf{y}) = -1_J \odot W$, where $\odot$ denotes elementwise multiplication and -1_J is the vector of all ones but with negative ones at the indices in J. To do this, note that D is invariant to swaps of $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, so $|W| = |W'|$ because by assumption $|W|, |W'|$ are a function of D. Furthermore, for any $j \in [p]$, we have that $W_j > 0$ if and only if $\widehat{\mathbf{X}}_j = \mathbf{X}_j$; however, since $\widehat{\mathbf{X}}_j$ is also a function of D, we have that $\text{sign}(W_j) = \text{sign}(W_j')$ if and only if $j \notin J$. This completes the proof. □

The proof of Proposition 2.2 is identical to the proof of Proposition 2.1, so we omit it for brevity. However, the following corollary of Propositions 2.1 and 2.2 will be important when proving Theorem 2.1.

Corollary A.1. *Let W, W' be two knockoff feature-statistics. Then in the same setting as Propositions 2.1 and 2.2, the event* $\text{sign}(W_j) = \text{sign}(W'_j)$ *is a deterministic function of the masked data D.*

Proof. As usual, we give the proof for the model-X case, and the fixed-X case is analagous. By Proposition (2.1), there exist deterministic functions $\widehat{\mathbf{X}}_j = g_j(D)$, $\widehat{\mathbf{X}}'_j = g'_j(D)$ such that $W_j > 0 \Leftrightarrow \widehat{\mathbf{X}}_j = \mathbf{X}_j$ and $W'_j > 0 \Leftrightarrow \widehat{\mathbf{X}}'_j = \mathbf{X}_j$. (Note this does not exclude the possibility of randomized feature statistics since D includes auxiliary random noise.) Since $\widehat{\mathbf{X}}_j, \widehat{\mathbf{X}}'_j$ must take one of exactly two distinct values, this implies that

$$\text{sign}(W_j) = \text{sign}(W'_j) \Leftrightarrow \widehat{\mathbf{X}}_j = \widehat{\mathbf{X}}'_j \Leftrightarrow g_j(D) = g'_j(D),$$

where we note that the right-most expression is clearly a deterministic function of D. This completes the proof. □

A.2 Proof of Theorem 2.1

The main idea behind Theorem 2.1 is that we will compare the power of a knockoff feature statistic W to the power of a "soft" version of the SeqStep procedure, which depends only on the conditional expectation of $\text{sign}(W)$ instead of the realized values of $\text{sign}(W)$. Roughly speaking, if $\text{sign}(W)$ obey a strong law of large numbers, these two procedures will have the same power asymptotically.

To make this precise, for a knockoff feature-statistic W, let $\text{sorted}(W)$ denote W sorted in decreasing order of its absolute values, and let $R = \mathbb{I}(\text{sorted}(W) > 0) \in \{0,1\}^p$ be the vector indicating where $\text{sorted}(W)$ has positive entries. The number of discoveries made by knockoffs only depends on R. Indeed, for any vector $\eta \in [0,1]^p$ and any desired FDR level $q \in (0,1)$, define

$$\psi_q(\eta) \coloneqq \max_{k \in [p]} \left\{ k : \frac{k - k\bar{\eta}_k + 1}{k\bar{\eta}_k} \leq q \right\} \text{ and } \tau_q(\eta) = \left\lceil \frac{\psi_q(\eta) + 1}{1+q} \right\rceil, \tag{A.1}$$

where by convention we set $\frac{x}{0} = \infty$ for any $x \in \mathbb{R}_{>0}$. It turns out that knockoffs makes exactly $\tau_q(R)$ discoveries. For brevity, we refer the reader to Lemma B.3 of Spector and Janson (2022) for a formal proof of this: however, to see this intuitively, note that $k - k\bar{R}_k + 1$ (resp. $k\bar{R}_k$) counts the number of negative (resp. positive) entries in the first k coordinates of $\text{sorted}(W)$, so this definition lines up with the definition of the data-dependent threshold in Section 1.3.

Now, let $\delta \coloneqq \mathbb{E}[R \mid D] \in [0,1]^p$ be the conditional expectation of R given the data we are "allowed" to observe when computing W, where $D = \{\mathbf{y}, \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p\}$ for model-X knockoffs and $D = \{\mathbf{X}, \widetilde{\mathbf{X}}, \{\mathbf{X}_j^T\mathbf{y}, \widetilde{\mathbf{X}}_j^T\mathbf{y}\}_{j=1}^p\}$ for fixed-X knockoffs. The "soft" version of SeqStep simply applies the functions ψ_q and τ_q to δ instead of R. Intuitively speaking, our goal will be to apply a law of large numbers to show the following asymptotic result:

$$|\tau_q(\delta) - \tau_q(R)| = o_p\left(\#\text{ of non-nulls}\right).$$

Once we have shown this, it will be straightforward to show that MLR statistics are asymptotically optimal, since MLR statistics maximize $\tau_q(\delta)$ in finite samples.

We now begin to prove Theorem 2.1 in earnest. In particular, the following pair of lemmas tells us that if $\bar{R}_k$ converges uniformly to $\bar{\delta}_k$, then $\tau_q(\delta) \approx \tau_q(R)$.

Lemma A.1. *Let $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ be any feature statistic with $R, \delta, \psi_q, \tau_q$ as defined earlier. Fix any $k_0 \in [p]$ and sufficiently small $\epsilon > 0$ such that $\eta = 3(1+q)\epsilon < 1$. Define the event*

$$A_{k_0,\epsilon} = \left\{ \max_{k_0 \le k \le p} |\bar{R}_k - \bar{\delta}_k| \le \epsilon \right\}.$$

Then on the event $A_{k_0,\epsilon}$, we have that $\tau_{q-\eta}(R) - k_0 \le \tau_q(\delta) \le \tau_{q+\eta}(R) + k_0$. In particular, since $\tau_q(\delta), \tau_q(R) \le p$, this implies that

$$|\tau_q(R) - \tau_q(\delta)| \le p\mathbb{I}(A^c_{k_0,\epsilon}) + \tau_{q+\eta}(R) - \tau_{q-\eta}(R) + k_0. \tag{A.2}$$

Proof. The proof is entirely algebraic (there is no probablistic content). To start, define the sets

$$\mathcal{R} = \left\{ k \in [p] : \frac{k - k\bar{R}_k + 1}{k\bar{R}_k} \le q + \eta \right\} \text{ and } \mathcal{D} = \left\{ k \in [p] : \frac{k - k\bar{\delta}_k + 1}{k\bar{\delta}_k} \le q \right\}$$

and recall that $\psi_q(R) = \max(\mathcal{R})$, $\psi_{q+\eta}(\delta) = \max(\mathcal{D})$. To analyze the difference between these quantities, fix any $k \in \mathcal{D} \setminus \mathcal{R}$. Then by definition of $\mathcal{D}$ and $\mathcal{R}$, we know

$$\frac{k - k\bar{\delta}_k + 1}{k\bar{\delta}_k} \le q < q + \eta < \frac{k - k\bar{R}_k + 1}{k\bar{R}_k}.$$

However, Lemma A.2 (proved in a moment) tells us that this implies the following:

$$\bar{\delta}_k - \bar{R}_k \ge \frac{\eta}{3(1+q)} = \frac{3(1+q)\epsilon}{3(1+q)} = \epsilon.$$

However, on the event $A_{k_0,\epsilon}$ this cannot occur for any $k \ge k_0$. Therefore, on the event $A_{k_0,\epsilon}$, $\mathcal{R} \setminus \mathcal{D} \subset \{1, \dots, k_0 - 1\}$. This implies that

$$\psi_q(\delta) - \psi_{q+\eta}(R) = \max(\mathcal{D}) - \max(\mathcal{R}) \le \begin{cases} 0 & \max(\mathcal{D}) \ge k_0 \\ k_0 - 1 & \max(\mathcal{D}) < k_0. \end{cases} \tag{A.3}$$

We can combine these conditions by writing that $\psi_q(\delta) - \psi_{q+\eta}(R) \le k_0 - 1$. Using the definition of $\tau_q(\cdot)$, we conclude

$$\begin{aligned} \tau_q(\delta) - \tau_{q+\eta}(\delta) &= \left\lceil \frac{\psi_q(R)}{1+q} \right\rceil - \left\lceil \frac{\psi_{q+\eta}(\delta)}{1+q+\eta} \right\rceil \\ &\le 1 + \frac{\psi_q(R)}{1+q} - \frac{\psi_{q+\eta}(\delta)}{1+q+\eta} \\ &= 1 + \psi_q(R) - \psi_{q+\eta}(\delta) \\ &\le 1 + k_0 - 1 \\ &= k_0. \end{aligned}$$

This proves the upper bound, namely that $\tau_q(\delta) \le \tau_{q+\eta}(R) + k_0$. To prove the lower bound, note that we can swap the role of R and δ and apply the upper bound to $q' = q - \eta$. Then if we take $\eta' = 3(1+q')\epsilon < \eta < 1$, applying the upper bound yields

$$\tau_{q'}(R) \le \tau_{q'+\eta'}(\delta) + k_0.$$

Observe that $\tau_q(\cdot)$ is monotone in q, and since $\eta' < \eta$, we have that (1) $q - \eta \le q' = q - \eta'$ and (2) $q' + \eta' = q - \eta + \eta' < q$. Therefore, by monotonicity, we conclude

$$\tau_{q-\eta}(R) \le \tau_{q'}(R) \le \tau_{q'+\eta'}(\delta) + k_0 \le \tau_q(\delta) + k_0.$$

Subtracting k_0 from both sides, this implies the lower bound $\tau_{q-\eta}(R) - k_0 \le \tau_q(\delta)$. □

Lemma A.2. *For any $x, y \in [0,1]$, $k \in \mathbb{N}$, and any $\gamma \in (0,1)$, suppose that $\frac{1+k-kx}{kx} \leq q < q+\gamma \leq \frac{1+k-ky}{ky}$. Then*

$$x - y \geq \frac{\gamma}{(1+q)(1+q+\gamma)} \geq \frac{\gamma}{3(1+q)}.$$

Proof. Note that we cannot have $x = 0$, since it would violate the strict inequality. For $x > 0$, we have that

$$\frac{1+k-kx}{kx} \leq q \implies 1+k-kx \leq kqx \implies x \geq \frac{k+1}{k(1+q)}. \tag{A.4}$$

Now, there are two cases. If $y = 0$, the inequality holds because we have that

$$x - y = x = \frac{k+1}{k} \cdot \frac{1}{1+q} \geq \frac{\gamma}{3(1+q)}.$$

Alternatively, if $y > 0$, we observe similarly to before that

$$\frac{1+k-ky}{ky} \geq q+\gamma \implies y \leq \frac{k+1}{k(1+q+\gamma)}. \tag{A.5}$$

Combining Equations (A.4)–(A.5) yields the result:

$$x - y \geq \frac{k+1}{k}\left(\frac{1}{1+q} - \frac{1}{1+q+\gamma}\right) = \frac{k+1}{k}\frac{1}{(1+q+\gamma)(1+q)} \geq \frac{1}{3(1+q)}.$$

□

We are now ready to prove Theorem 2.1. As a reminder, we consider an asymptotic regime with data $\mathbf{X}^{(n)} \in \mathbb{R}^{n \times p_n}, \mathbf{y}^{(n)} \in \mathbb{R}^n$ and knockoffs $\widetilde{\mathbf{X}}^{(n)}$, where the likelihood is denoted by $L(\mathbf{y}^{(n)}; \mathbf{X}^{(n)}, \theta^{(n)})$ and $\pi^{(n)}$ is a prior on the unknown parameters $\theta^{(n)}$. We let $D^{(n)}$ denote the masked data for knockoffs as defined in Section 2.2 and let $\kappa^{(n)}$ denote the expected number of non-nulls under $\pi^{(n)}$. We will analyze the limiting *empirical power* of feature-statistics $W^{(n)} = w_n([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y})$ with rejection set $S^{(n)}(q)$, defined as the expected number of discoveries divided by the expected number of non-nulls:

$$\widetilde{\text{Power}}_q(w_n) = \frac{\mathbb{E}[|S^{(n)}(q)|]}{\kappa^{(n)}}. \tag{2.7}$$

For convenience, we restate Theorem 2.1 and then prove it.

Theorem 2.1. *For each n, let $W^\star = w_n^\star([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y}^{(n)})$ denote the MLR statistics with respect to $\pi^{(n)}$ and let $W = w_n([\mathbf{X}^{(n)}, \widetilde{\mathbf{X}}^{(n)}], \mathbf{y}^{(n)})$ denote any other sequence of feature-statistics. Let $(\mathbf{y}^{(n)}, \mathbf{X}^{(n)}, \theta^{(n)})$ be any sequence such that the following three conditions hold:*

- *Assume $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star)$ and $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n)$ exist for each $q \in (0,1)$.*
- *The expected number of non-nulls grows faster than $\log(p_n)^4$. Formally, assume that for some $\gamma > 0$, $\lim_{n\to\infty} \frac{\kappa^{(n)}}{\log(p_n)^{4+\gamma}} = \infty$.*
- *Assume that conditional on $D^{(n)}$, the covariance between the signs of $W^\star$ decays exponentially. That is, there exist constants $C \geq 0, \rho \in (0,1)$ such that*

$$|\operatorname{Cov}(\mathbb{I}(W_i^\star > 0), \mathbb{I}(W_j^\star > 0) \mid D^{(n)})| \leq C\rho^{|i-j|}. \tag{2.8}$$

Then for all but countably many $q \in (0,1)$,

$$\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star) \geq \lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n). \tag{A.6}$$

Proof. The proof proceeds in three main steps: before beginning these steps, however, we first we outline the overall strategy of the proof and introduce some initial notation. In particular, following Lemma A.1, let $R = \mathbb{I}(\text{sorted}(W) > 0)$ and $R^\star = \mathbb{I}(\text{sorted}(W^\star) > 0)$, and let $\delta = \mathbb{E}[W \mid D^{(n)}]$ and $\delta^\star = \mathbb{E}[W^\star \mid D^{(n)}]$ be their conditional expectations. (Note that $W, R, \delta, W^\star, R^\star$ and $\delta^\star$ all depend on n—however, we omit this dependency to lighten the notation.) As in Equation (A.1), we can write the number of discoveries made by W and $W^\star$ as a function of $R^\star$ and R:

$$\widetilde{\text{Power}}_q(w_n^\star) - \widetilde{\text{Power}}_q(w_n) = \frac{\mathbb{E}\left[\tau_q(R^\star)\right]}{\kappa^{(n)}} - \frac{\mathbb{E}\left[\tau_q(R)\right]}{\kappa^{(n)}}.$$

We will show that the limit of this quantity is nonnegative, and the main idea is to make the approximations $\tau_q(R^\star) \approx \tau_q(\delta^\star)$ and $\tau_q(R) \approx \tau_q(\delta)$. In particular, we can decompose

$$\begin{aligned}
\widetilde{\text{Power}}_q(w_n^\star) - \widetilde{\text{Power}}_q(w_n) &= \frac{\mathbb{E}\left[\tau_q(R^\star) - \tau_q(R)\right]}{\kappa^{(n)}} \\
&= \frac{\mathbb{E}[\tau_q(R^\star) - \tau_q(\delta^\star)]}{\kappa^{(n)}} + \frac{\mathbb{E}[\tau_q(\delta^\star) - \tau_q(\delta)]}{\kappa^{(n)}} + \frac{\mathbb{E}\left[\tau_q(\delta) - \tau_q(R)\right]}{\kappa^{(n)}} \\
&\geq \frac{\mathbb{E}[\tau_q(\delta^\star) - \tau_q(\delta)]}{\kappa^{(n)}} - \frac{\mathbb{E}|\tau_q(R^\star) - \tau_q(\delta^\star)|}{\kappa^{(n)}} - \frac{\mathbb{E}|\tau_q(\delta) - \tau_q(R)|}{\kappa^{(n)}}.
\end{aligned} \tag{A.7}$$

In particular, Step 1 of the proof is to show that $\tau_q(\delta^\star) \geq \tau_q(\delta)$ holds deterministically, for fixed n. This implies that the first term of Equation (A.7) is nonnegative for fixed n. In Step 2, we show that as $n \to \infty$, the second and third term Equation (A.7) vanish. In Step 3, we combine these results and take limits to yield the final result.

Step 1: In this step, we show that $\tau_q(\delta^\star) \geq \tau_q(\delta)$ holds deterministically for fixed n. To do this, it suffices to show that $\bar{\delta}_k^\star \geq \bar{\delta}_k$ for each $k \in [p_n]$. To see this, recall that

$$\tau_q(\delta^\star) = \max_{k\in[p_n]} \left\{ \frac{k - k\bar{\delta}_k^\star + 1}{k\bar{\delta}_k^\star} \leq q \right\} \text{ and } \tau_q(\delta) = \max_{k\in[p_n]} \left\{ \frac{k - k\bar{\delta}_k + 1}{k\bar{\delta}_k} \leq q \right\}. \tag{A.8}$$

Since the function $\gamma = \frac{k-k\gamma+1}{k\gamma}$ is decreasing in γ, $\bar{\delta}_k^\star \geq \bar{\delta}_k$ implies that $\frac{k-k\bar{\delta}_k^\star+1}{k\bar{\delta}_k^\star} \leq \frac{k-k\bar{\delta}_k+1}{k\bar{\delta}_k}$ for each k, and therefore $\tau_q(\delta^\star) \geq \tau_q(\delta)$. Thus, it suffices to show that $\bar{\delta}_k^\star \geq \bar{\delta}_k$ holds for each k.

To do this, we first argue that conditional on $D^{(n)}$, $R^\star$ is a deterministic function of R. Recall that according to Corollary A.1, the event $\text{sign}(W_j) \neq \text{sign}(W_j^\star)$ is completely determined by the masked data $D^{(n)}$. Furthermore, since $R^\star$ and R are random permutations of the vectors $\mathbb{I}(W > 0)$ and $\mathbb{I}(W^\star > 0)$ where the random permutations only depend on $|W|$ and $|W^\star|$, this implies there exists a random vector $\xi \in \{0,1\}^{p_n}$ and a random permutation $\sigma \in S_{p_n}$ such that $R^\star = \xi \oplus \sigma(R)$ and ξ, σ are deterministic conditional on $D^{(n)}$. The intuition here is that following Proposition 2.1, fitting a feature statistic W is equivalent to observing $D^{(n)}$, assigning an ordering to the features, and then guessing which one of $\{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}$ is the true feature and which is a knockoff, where $W_j > 0$ if and only if this "guess" is correct. Since these decisions are made as deterministic functions of $D^{(n)}$, $W^\star$ can only be different than W in that it sometimes makes different guesses, flipping the sign of W (as represented by ξ), and is in a different order (as represented by σ).

Now, since ξ and σ are deterministic functions of $D^{(n)}$, this implies that

$$\delta_i^\star = \mathbb{E}[R_i^\star \mid D^{(n)}] = \mathbb{E}[\xi_i \oplus R_{\sigma(i)} \mid D^{(n)}] = \begin{cases} 1 - \delta_{\sigma(i)} & \xi_i = 1 \\ \delta_{\sigma(i)} & \xi_i = 0. \end{cases}$$

However, by construction, $\mathbb{E}[R_i^\star \mid D^{(n)}] = \mathbb{P}(\text{sorted}(W^\star)_i > 0 \mid D^{(n)})$, and Proposition 2.3 tells us that $\mathbb{P}(W_i^\star > 0 \mid D^{(n)}) \geq 0.5$ for all $i \in [p_n]$: since the ordering of $W^\star$ is deterministic conditional on $D^{(n)}$, this also implies $\mathbb{P}(\text{sorted}(W^\star)_i > 0 \mid D^{(n)})$. Therefore, $\delta_i^\star \geq 0.5$ and thus $\delta_i^\star \geq \delta_{\sigma(i)}$ for each $i \in [p_n]$. Additionally, by construction $W^\star$ ensures that $\delta_1^\star \geq \delta_2^\star \geq \cdots \geq \delta_{p_n}^\star$. If $\delta_{(1)}, \ldots, \delta_{(p_n)}$ are the order statistics of δ in decreasing order, this implies that $\delta_i^\star \geq \delta_{(i)}$ for all i. Therefore,

$$\bar{\delta}_k^\star = \frac{1}{k}\sum_{i=1}^k \delta_i^\star \geq \frac{1}{k}\sum_{i=1}^k \delta_{(i)} \geq \frac{1}{k}\sum_{i=1}^n \delta_i.$$

By the previous analysis, this proves that $\tau_q(\delta^\star) \geq \tau_q(\delta)$.

Step 2: In this step, we show that $\frac{\mathbb{E}|\tau_q(\delta^\star) - \tau_q(R^\star)|}{\kappa^{(n)}} \to 0$ for all but countably many $q \in (0,1)$, as well as the analagous result for R and δ. We start by proving the result for $R^\star$ and $\delta^\star$, and we begin by applying Lemma A.1. Indeed, Lemma A.1 tells us that if we fix any $k_n \in [p_n]$ and any $\epsilon_n > 0$,

$$A_n = \left\{ \max_{k_n \leq k \leq p_n} |\bar{R}_k^\star - \bar{\delta}_k^\star| \leq \epsilon_n \right\},$$

and

$$|\tau_q(R^\star) - \tau_q(\delta^\star)| \leq p_n \mathbb{I}(A_n^c) + \tau_{q+\eta_n}(R^\star) - \tau_{q-\eta_n}(R^\star) + k_n.$$

where $\eta_n = 3(1+q)\epsilon_n$. Therefore,

$$\frac{\mathbb{E}|\tau_q(R^\star) - \tau_q(\delta^\star)|}{\kappa^{(n)}} \leq p_n \mathbb{P}(A_n^c) + \frac{k_n}{\kappa^{(n)}} + \frac{\mathbb{E}[\tau_{q+\eta_n}(R^\star)] - \mathbb{E}[\tau_{q-\eta_n}(R^\star)]}{\kappa^{(n)}}. \tag{A.9}$$

We now analyze these terms in order: while doing so, we will choose sequences $\{k_n\}, \{\epsilon_n\}$ which guarantee the desired result. Note that eventually, our choice of $\{\epsilon_n\}$ will depend on q, so the convergence is not necessarily uniform, but that will not be a problem for our proof.

First term: To start, we will first apply a finite-sample concentration result to bound $\mathbb{P}(A_n^c)$. In particular, we show in Corollary B.1 that if $X_1, \ldots, X_n$ are mean-zero, $[-1,1]$-valued random variables satisfying the exponential decay condition from Equation (2.8), then there exists a universal constant $C' > 0$ depending only on C and ρ such that

$$\mathbb{P}\left(\max_{n_0 \leq i \leq n} |\bar{X}_k| \geq t \right) \leq n \exp(-C' t^{1/4} n_0^{1/4}). \tag{A.10}$$

Furthermore, Corollary B.1 shows that this result holds even if we permute $X_1, \ldots, X_n$ according to some arbitrary *fixed* permutation σ. Now, observe that conditional on $D^{(n)}$, $R_j^\star - \delta_j^\star$ is a zero-mean, $[-1,1]$-valued random variable which is a fixed permutation of $\mathbb{I}(W^\star > 0)$ minus its (conditional) expectation. Since $\mathbb{I}(W^\star > 0)$ obeys the exponential decay condition in Equation (2.8), $R_j^\star - \delta_j^\star$ must as well: therefore, we conclude

$$\mathbb{P}(A_n^c \mid D^{(n)}) \leq p_n \exp(-C' \epsilon_n^{1/4} k_n^{1/4}) \tag{A.11}$$

which implies by the tower property that $p_n \mathbb{P}(A_n^c) \leq p_n^2 \exp(-C' \epsilon_n^{1/4} k_n^{1/4})$. Now, suppose we take

$$k_n = \left\lceil \log(p_n)^{4+\gamma} \right\rceil \text{ and } \epsilon_n = \Omega\left(\frac{1}{\log(p_n)^{\gamma/2}}\right).$$

Then observe that as as $n \to \infty$, $k_n^{1/4} \epsilon_n^{1/4} = \Omega(\log(p_n)^{1+\gamma/2})$, and thus

$$\log(p_n \mathbb{P}(A_n^c)) \leq 2\log(p_n) - \Omega\left(\log(p_n)^{1+\gamma/2}\right) \to -\infty.$$

Therefore, for any sequence satisfying this condition, we have that $p_n \mathbb{P}(A_n^c) \to 0$.

Second term: This term is easy, as we assume in the statement that $\frac{k_n}{\kappa^{(n)}} \sim \frac{\log(p_n)^{4+\gamma}}{\kappa^{(n)}} \to 0$.

Third term: We will now show that for all but countably many $q \in (0,1)$, there exists a sequence $\{\epsilon_n\}$ satisfying (i) $\epsilon_n = \Omega\left(\frac{1}{\log(p_n)^{\gamma/2}}\right)$ and (ii) $\limsup_{n\to\infty} \frac{\mathbb{E}[\tau_{q+\eta_n}(R^\star)] - \mathbb{E}[\tau_{q-\eta_n}(R^\star)]}{\kappa^{(n)}} \leq v$ for *any* $v > 0$.

To do this, recall by assumption that for all $q \in (0,1)$, we have that $\lim_{n\to\infty} \frac{\mathbb{E}[\tau_q(R^\star)]}{\kappa^{(n)}}$ exists and converges to some (extended) real number $L(q)$. Furthermore, we show in Lemma B.2 that $L(q)$ is always finite—this is intuitively a consequence of the fact that knockoffs controls the false discovery rate, and thus the number of discoveries cannot exceed the number of non-nulls by more than a constant factor. Crucially, since $\tau_q(R^\star)$ is increasing in q, the function $L(q)$ is increasing in q for all $q \in (0,1)$: therefore, it is continuous on $(0,1)$ except on a countable set.

Supposing that q is a continuity point of $L(q)$, there exists some $\beta > 0$ such that $|q - q'| \leq \beta \implies |L(q) - L(q')| \leq v/2$. Take ϵ_n to be the constant sequence $\epsilon_n = \frac{\beta}{3(1+q)}$ such that $\eta_n = \beta$ for all n. Then since η_n is constant, we have that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{\mathbb{E}[\tau_{q+\eta_n}(R^\star)] - \mathbb{E}[\tau_{q-\eta_n}(R^\star)]}{\kappa^{(n)}} &= L(q+\beta) - L(q-\beta) \quad \text{because } \frac{\mathbb{E}[\tau_q(R^\star)]}{\kappa^{(n)}} \to L(q) \text{ pointwise} \\
&\leq v. \qquad \text{by continuity}
\end{aligned}$$

Combining the results for all three terms, we see the following: for each $v > 0$, there exist sequences $\{k_n\}, \{\epsilon_n\}$ guaranteeing that

$$\lim_{n\to\infty} \frac{\mathbb{E}|\tau_q(R^\star) - \tau_q(\delta^\star)|}{\kappa^{(n)}} \leq v.$$

Since this holds for all $v > 0$, we conclude $\lim_{n\to\infty} \frac{\mathbb{E}|\tau_q(R^\star) - \tau_q(\delta^\star)|}{\kappa^{(n)}} = 0$ as desired.

Lastly in the step, we need to show the same result for R and δ in place of $R^\star$ and $\delta^\star$. However, the proof for for R and δ is identical to the proof for $R^\star$ and $\delta^\star$. The one subtlety worth mentioning is that we do not directly assume the exponential decay condition in Equation (2.8) for W. However, as we argued in Step 1, we can write $\mathbb{I}(W > 0) = \xi \oplus \mathbb{I}(W^\star > 0)$ for some random vector $\xi \in \{0,1\}^{p_n}$ which is a deterministic function of $D^{(n)}$. As a result, we have that

$$|\operatorname{Cov}(\mathbb{I}(W_i > 0), \mathbb{I}(W_j > 0) \mid D^{(n)})| = |\operatorname{Cov}(\mathbb{I}(W_i^\star > 0), \mathbb{I}(W_j^\star > 0) \mid D^{(n)})| \leq C\rho^{|i-j|}.$$

Thus, we also conclude that $\lim_{n\to\infty} \frac{\mathbb{E}|\tau_q(R) - \tau_q(\delta)|}{\kappa^{(n)}} = 0$.

Step 3: Finishing the proof. Recall Equation (A.7), which states that

$$\widetilde{\text{Power}}_q(w_n^\star) - \widetilde{\text{Power}}_q(w_n) \geq \frac{\mathbb{E}[\tau_q(\delta^\star) - \tau_q(\delta)]}{\kappa^{(n)}} - \frac{\mathbb{E}|\tau_q(R^\star) - \tau_q(\delta^\star)|}{\kappa^{(n)}} - \frac{\mathbb{E}|\tau_q(\delta) - \tau_q(R)|}{\kappa^{(n)}}. \quad \text{(A.7)}$$

In Step 1, we showed that $\tau_q(\delta^\star) \geq \tau_q(\delta)$ for fixed n. Furthermore, in Step 2, we showed that the second two terms vanish asymptotically. As a result, we take limits and conclude

$$\liminf_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star) - \widetilde{\text{Power}}_q(w_n) \geq 0.$$

Furthermore, since we assume that the limits $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star), \lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n)$ exist, this implies that

$$\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n^\star) - \lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n) \geq 0.$$

This concludes the proof.

□

A.3 Further discussion of assumptions in Theorem 2.1

In this section, we discuss a few ways to relax the assumptions in Theorem 2.1.

First, we can easily relax the assumption that the limits $L(q) \coloneqq \lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n)$ and $L^\star(q) \coloneqq \lim_{n\to\infty} \widetilde{\text{Power}}_n(w_n^\star)$ exist for each $q \in (0,1)$. Indeed, the proof of Theorem 2.1 only uses this assumption to argue that there exists a sequence $\eta_n \to 0$ such that $L(q+\eta_n) \to L(q), L(q-\eta_n) \to L(q)$ (and similarly for $L^\star(q)$). Thus, we do not need the limits $L(q)$ to exist for every $q \in (0,1)$: in contrast, the result of Theorem 2.1 will hold (e.g.) for any q such that $L(\cdot), L^\star(\cdot)$ are continuous at q. Intuitively, this means that the result in Theorem 2.1 holds except at points q which delineate a "phase transition," where the power of knockoffs jumps in a discontinuous fashion as q increases.

Second, it is important to note that the precise form of the local dependency condition (2.8) is not crucial. Indeed, the proof of Theorem 2.1 only uses this condition to show that the partial sums of $\mathbb{I}(W^\star > 0)$ converge to their conditional mean given D. To be precise, fix any permutation $\kappa : [p] \to [p]$ and let $R = \mathbb{I}(\kappa(W^\star) > 0)$ where $\kappa(W^\star)$ permutes $W^\star$ according to κ. Let $\delta = \mathbb{E}[R \mid D]$. Then the proof of Theorem 2.1 will go through exactly as written if we replace Equation (2.8) with the following condition:

$$\mathbb{P}\left(\max_{k_n \leq k \leq p_n} |\bar{R}_k - \bar{\delta}_k| \mid D\right) = o(p_n^{-1}) \tag{A.12}$$

where k_n is some sequence satisfying $k_n \to \infty$ and $\frac{k_n}{\kappa^{(n)}} \to 0$.

The upshot is this: under any condition where each permutation of $\mathbb{I}(W^\star > 0)$ obeys a certain strong law of large numbers, we should expect Theorem 2.1 to hold. Although it is perhaps slightly unusual to require that a strong law holds for any fixed permutation of a vector, usually there is a "worst-case" permutation where if Equation (A.12) holds for some choice of κ, then it holds for every choice of κ. For example, in Corollary B.1, we show that if Equation (B.1) holds, then it suffices to show Equation (A.12) in the case where κ is the identity permutation, since the identity permutation places the most correlated coordinates of $W^\star$ next to each other.

A.4 Maximizing the expected number of true discoveries

One weakness of Theorem 2.1 is that it shows that MLR statistics maximize the (normalized) expected number of discoveries, which is not exactly the same as maximizing the expected number

of *true* discoveries. In this section, we introduce a modification of MLR statistics and give a proof sketch that they maximize the expected number of true discoveries. However, computing these modified MLR statistics is extremely computationally expensive, so we prefer to use the MLR statistics defined in the paper, as they perform quite well anyway.

Throughout this section, we use the notation introduced in Section A.2. As a reminder, for any feature statistic W, let $R = \mathbb{I}(\text{sorted}(W) > 0)$, let $\delta = \mathbb{E}[R \mid D]$, and let $\psi_q(\cdot)$ be as defined in Equation (A.1) so that MLR statistics make $\tau_q(R) = \left\lceil \frac{\psi_q(R)+1}{1+q} \right\rceil$ discoveries. The key idea behind the proof of Theorem 2.1 is to observe that:

1. $\tau_q(R)$ only depends on the successive partial averages of R, denoted $\bar{R}_k = \frac{1}{k}\sum_{i=1}^{k} R_i$.
2. As $p \to \infty$, $\bar{R}_k \overset{\text{a.s.}}{\to} \bar{\delta}_k$ under suitable assumptions. Thus, $\tau_q(R) \approx \tau_q(\delta)$.
3. If $R^\star = \mathbb{I}(\text{sorted}(W^\star) > 0)$ are MLR statistics with $\delta^\star = \mathbb{E}[R^\star \mid D]$, then $R^\star$ is asymptotically optimal because $\tau_q(\delta^\star) \geq \tau_q(\delta)$ holds in finite samples for any choice of δ. In particular, this holds because MLR statistics ensure $\delta^\star$ is in descending order.

To show a similar result for the number of *true* discoveries, we can now effectively repeat the three steps used in the proof of Theorem 2.1. To do this, let I_j be the indicator that the feature corresponding the the jth coordinate of R_j is non-null, and let $B_j = \mathbb{I}(I_j = 1, R_j = 1)$ be the indicator that $\text{sorted}(W)_j > 0$ *and* that the corresponding feature is non-null. Let $b = \mathbb{E}[B \mid D]$. Then:

1. Let $T_q(R, B)$ denote the number of *true* discoveries. We claim that $T_q(R, B)$ is a function of the successive partial means of R and B. To see this, recall that the knockoffs procedure applied to W will make $\tau_q(R)$ discoveries, and in particular it will make discoveries corresponding to any of the first $\psi_q(R)$ coordinates of R which are positive. Therefore,

$$T_q(R, B) = \sum_{j=1}^{\psi_q(R)} B_j = \psi_q(R) \cdot \frac{1}{\psi_q(R)} \sum_{j=1}^{\psi_q(R)} B_j.$$

   Since $\psi_q(R)$ only depends on the successive averages of R and the second term is itself a successive partial average of $\{B_j\}$, this finishes the first step.

2. The second step in the "proof sketch" is to show that as $p \to \infty$, $\bar{B}_k \overset{\text{a.s.}}{\to} \bar{b}_k, \bar{R}_k \overset{\text{a.s.}}{\to} \bar{\delta}_k$ and therefore $T_q(R, B) \approx T_q(\delta, b)$. This can be done using the same techniques as Theorem 2.1, although it requires an extra assumption that B also obeys the local dependency condition (2.8). However, just like the original local dependency condition, this condition also only depends on the posterior of B, so it can be diagnosed using the data as hand.

3. To complete the proof, we need to define a modified MLR statistic W' with corresponding R', δ', b' such that $T_q(\delta', b') \geq T_q(\delta, b)$ holds in finite samples for any other feature statistic W. It is easy to see that W' must have the same *signs* as the original MLR statistics $W^\star$, since the signs of $W^\star$ maximize $\delta^\star$ and $b^\star$ coordinatewise. However, the *absolute values* of W' may differ from those of $W^\star$, since it is not always true that sorting δ in decreasing order maximizes $T_q(\delta, b)$. Since changing the absolute values of $W^\star$ merely permutes $b^\star$ and $\delta^\star$, the modified MLR statistic must solve the following combinatorial optimization problem:

$$\kappa^\star = \underset{\kappa:[p]\to[p]}{\arg\max}\, T_q(\kappa(\delta^\star), \kappa(b^\star)) = \underset{\kappa:[p]\to[p]}{\arg\max} \sum_{j=1}^{\psi_q(\kappa(\delta^\star))} b^\star_{\kappa(j)}. \tag{A.13}$$

Having found this optimal choice of $\kappa^\star$, we can construct the modified MLR statistics setting $W' \approx W^\star$, except we change the magnitudes of W' so that the ordering implied by its absolute values agrees with the one implied by $\kappa^\star$, i.e., $\mathbb{E}[\text{sorted}(W') > 0 \mid D] = \kappa^\star(\delta^\star)$.

Intuitively, we do not expect the modified MLR statistics W' to look very different from W. To see this, note that in the special case where $\delta^\star$ has the same order as $b^\star$, MLR statistics exactly maximize $T_q(\delta^\star, b^\star)$. Usually, $\delta^\star$ will have a very similar order to $b^\star$, because

$$\delta_j^\star = \mathbb{P}(\text{sorted}(W^\star)_j > 0 \mid D) \text{ and } b^\star = \mathbb{P}(\text{sorted}(W^\star)_j > 0, I_j > 0 \mid D).$$

Indeed, the pairwise exchangeability of knockoffs guarantees that $\delta_j^\star$ can only be large whenever $I_j = 1$ with high probability, so $\delta_j^\star$ and $b_j^\star$ are extremely closely related. Intuitively, this makes sense: a knockoff W-statistic will only be highly likely to be positive if feature j has a large posterior probability of being non-null. This suggests that MLR statistics will approximately maximize the expected number of true discoveries, even though we did not prove this rigorously.

A.5 Verifying the local dependence assumption in a simple setting

We now prove Proposition 2.5, which verifies the local dependency condition (2.8) in the setting where $\mathbf{X}^T\mathbf{X}$ is block-diagonal and σ^2 is known.

Proposition 2.5. *Suppose $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$ and $\mathbf{X}^T\mathbf{X}$ is a block-diagonal matrix with maximum block size $M \in \mathbb{N}$. Suppose π is any prior such that the coordinates of β are a priori independent and σ^2 is a known constant. Then if $\widetilde{\mathbf{X}}$ are either fixed-X knockoffs or conditional Gaussian model-X knockoffs (Huang and Janson, 2020), the coordinates of $\text{sign}(W^\star)$ are M-dependent conditional on D, implying that Equation (2.8) holds, e.g., with $C = 2^M$ and $\rho = \frac{1}{2}$.*

Proof. Define $R \coloneqq \mathbb{I}(W > 0)$. We will prove the stronger result that if $J_1, \dots, J_m \subset [p]$ are a partition of $[p]$ corresponding to the blocks of $\mathbf{X}^T\mathbf{X}$, then $R_{J_1}, \dots, R_{J_m}$ are jointly independent conditional on D. As notation, suppose without loss of generality that $J_1, \dots, J_m$ are contiguous subsets and $\mathbf{X}^T\mathbf{X} = \text{diag}\{\Sigma_1, \dots, \Sigma_m\}$ for $\Sigma_i \in \mathbb{R}^{|J_i| \times |J_i|}$.

We give the proof for model-X knockoffs; as always, the proof for fixed-X knockoffs is quite similar. Recall by Proposition 2.1 that we can write $R_j = \mathbb{I}(W_j > 0) = \mathbb{I}(\mathbf{X}_j = \widehat{\mathbf{X}}_j)$ where $\widehat{\mathbf{X}}_j$ is a function of the masked data D. Therefore, to show $R_{J_1}, \dots, R_{J_m}$ are independent conditional on D, it suffices to show $X_{J_1}, \dots, X_{J_m}$ are conditionally independent given D. To do this, it will first be useful to note that for any value of $\mathbf{X}$ which is consistent with D,

$$\begin{aligned} L_{\beta,\sigma}(\mathbf{y} \mid \mathbf{X}) &\propto \exp\left(-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||_2^2\right) \\ &\propto \exp\left(\frac{2\beta^T\mathbf{X}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{X}\beta}{2\sigma^2}\right) \\ &\propto \prod_{i=1}^m \exp\left(\frac{2\beta_{J_i}^T\mathbf{X}_{J_i}^T\mathbf{y} - \beta_{J_i}^T\Sigma_i\beta_{J_i}}{2\sigma^2}\right). \end{aligned}$$

A subtle but very important observation in the calculation above is that we can verify that $\mathbf{X}^T\mathbf{X} = \text{diag}\{\Sigma_1, \dots, \Sigma_m\}$ having only observed D without observing $\mathbf{X}$. Indeed, this follows because for conditional Gaussian MX knockoffs, $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X}$ and $\widetilde{\mathbf{X}}^T\mathbf{X}$ only differs from $\mathbf{X}^T\mathbf{X}$ on the main

diagonal (just like in the fixed-X case). With this observation in mind, let $p(\cdot \mid \cdot)$ denote an arbitrary conditional density, and observe

$$\begin{aligned}
p(\mathbf{X} \mid D) &\propto p(\mathbf{X}, \mathbf{y} \mid \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p) \\
&= p(\mathbf{X} \mid \{\mathbf{X}_j, \widetilde{\mathbf{X}}_j\}_{j=1}^p) p(\mathbf{y} \mid \mathbf{X}) \qquad \text{since } \mathbf{y} \perp\!\!\!\perp \widetilde{\mathbf{X}} \mid \mathbf{X} \\
&= \frac{1}{2^p} p(\mathbf{y} \mid \mathbf{X}) \qquad \text{by pairwise exchangeability} \\
&\propto \int_\beta p(\beta) p(\mathbf{y} \mid \mathbf{X}, \beta) d\beta \\
&\propto \int_\beta \prod_{i=1}^m p(\beta_{J_i}) \exp\left(\frac{-\beta_{J_i}^T \Sigma_i \beta_{J_i}}{2\sigma^2}\right) \exp\left(\frac{\beta_{J_i}^T \mathbf{X}_{J_i}^T \mathbf{y}}{\sigma^2}\right) d\beta \\
&\propto \int_{\beta_{J_1}} \dots, \int_{\beta_{J_m}} \prod_{i=1}^m p(\beta_{J_i}) \exp\left(\frac{-\beta_{J_i}^T \Sigma_i \beta_{J_i}}{2\sigma^2}\right) \exp\left(\frac{\beta_{J_i}^T \mathbf{X}_{J_i}^T \mathbf{y}}{\sigma^2}\right) d\beta_{J_1}\, \beta_{J_2} \dots\, d\beta_{J_m}.
\end{aligned}$$

At this point, we can iteratively pull out parts of the product. In particular, define the following function:

$$q_i(\mathbf{X}_{j_i}) \coloneqq \int_{\beta_{J_i}} p(\beta_{J_i}) \exp\left(\frac{-\beta_{J_i}^T \Sigma_i \beta_{J_i}}{2\sigma^2}\right) \exp\left(\frac{\beta_{J_i}^T \mathbf{X}_{J_i}^T \mathbf{y}}{\sigma^2}\right) d\beta_{J_i}.$$

Since $\mathbf{y}, \sigma^2$ and Σ_i are fixed, $q_i(\mathbf{X}_{j_i})$ is a deterministic function of $\mathbf{X}_{j_i}$ that does not depend on β_{-J_i}. Therefore, we can iteratively integrate as below:

$$\begin{aligned}
p(\mathbf{X} \mid D) &\propto \int_{\beta_{J_1}} \dots, \int_{\beta_{J_m}} \prod_{i=1}^m p(\beta_{J_i}) \exp\left(\frac{-\beta_{J_i}^T \Sigma_i \beta_{J_i}}{2\sigma^2}\right) \exp\left(\frac{\beta_{J_i}^T \mathbf{X}_{J_i}^T \mathbf{y}}{\sigma^2}\right) d\beta_{J_1}\, d\beta_{J_2} \dots\, d\beta_{J_m} \\
&= \int_{\beta_{J_1}} \cdots \int_{\beta_{J_{m-1}}} \prod_{i=1}^{m-1} p(\beta_{J_i}) \exp\left(\frac{-\beta_{J_i}^T \Sigma_i \beta_{J_i}}{2\sigma^2}\right) \exp\left(\frac{\beta_{J_i}^T \mathbf{X}_{J_i}^T \mathbf{y}}{\sigma^2}\right) q_m(\mathbf{X}_{j_m}) d\beta_{J_1}\, d\beta_{J_2} \dots\, d\beta_{J_{m-1}} \\
&= \prod_{i=1}^m q_i(\mathbf{X}_{j_i}).
\end{aligned}$$

This shows that $\mathbf{X}_{J_1}, \dots, \mathbf{X}_{J_m} \mid D$ are jointly (conditionally) independent since their density factors, thus completing the proof. For fixed-X knockoffs, the proof is very similar as one can show that the density of $p(\mathbf{X}^T \mathbf{y} \mid D)$ factors into blocks. □

B Technical proofs

B.1 Key concentration results

The proof of Theorem 2.1 relies on the fact that the successive averages of the vector $\mathbb{I}(\text{sorted}(W) > 0) \in \mathbb{R}^p$ converge uniformly to their conditional expectation given the masked data $D^{(n)}$. In this section, we give a brief proof of this result, which is essentially an application of Theorem 1 from Doukhan and Neumann (2007). For convenience, we first restate a special case of this theorem (namely, the case where the random variables in question are bounded and we have bounds on pairwise correlations) before proving the corollary we use in Theorem 2.1.

Theorem B.1 (Doukhan and Neumann (2007)). *Suppose that $X_1, \dots, X_n$ are mean-zero random variables taking values in $[-1, 1]$ such that $\mathrm{Var}(\bar{X}_n) \le C_0 n$ for a constant $C_0 > 0$. Let $L_1, L_2 < \infty$ be constants such that for any $i \le j$,*

$$|\operatorname{Cov}(X_i, X_j)| \le 4\varphi(j - i)$$

where $\{\varphi(k)\}_{k\in\mathbb{N}}$ is a nonincreasing sequence satisfyng

$$\sum_{s=0}^{\infty} (s+1)^k \varphi(s) \le L_1 L_2^k k! \text{ for all } k \ge 0.$$

Then for all $t \ge 0$, there exists a universal constant $C_1 > 0$ only depending on C_0, L_1 and L_2

$$\mathbb{P}\left(\bar{X}_n \ge t\right) \le \exp\left(-\frac{t^2}{C_0 n + C_1 t^{7/4} n^{7/4}}\right) \le \exp\left(-C' t^{1/4} n^{1/4}\right),$$

where C' is a universal constant only depending on C_0, L_1, L_2.

If we take $\varphi(s) = c\rho^s$, this yields the following corollary.

Corollary B.1. *Suppose that $X_1, \dots, X_n$ are mean-zero random variables taking values in $[-1, 1]$. Suppose that for some $C \ge 0, \rho \in (0, 1)$, the sequence satisfies*

$$|\operatorname{Cov}(X_i, X_j)| \le C\rho^{|i-j|}. \tag{B.1}$$

Then there exists a universal constant C' depending only on C and ρ such that

$$\mathbb{P}(\bar{X}_n \ge t) \le \exp\left(-C' t^{1/4} n^{1/4}\right). \tag{B.2}$$

Furthermore, let $\pi : [n] \to [n]$ be any permutation. For $k \le n$, define $\bar{X}_k^{(\pi)} := \frac{1}{k}\sum_{i=1}^{k} X_{\pi(i)}$ to be the sample mean of the first k random variables after permuting $(X_1, \dots, X_n)$ according to π. Then for any $n_0 \in \mathbb{N}, t \ge 0$,

$$\sup_{\pi\in S_n} \mathbb{P}\left(\max_{n_0 \le i \le n} |\bar{X}_k^{(\pi)}| \ge t\right) \le n \exp(-C' t^{1/4} n_0^{1/4}). \tag{B.3}$$

where S_n is the symmetric group.

Proof. The proof of Equation (B.2) follows an observation of Doukhan and Neumann (2007), where we note $\varphi(s) = C\exp(-as)$ for $a = -\log(\rho)$. Then

$$\sum_{s=0}^{\infty} (s+1)^k \exp(-as) \le \sum_{s=0}^{\infty} \prod_{i=1}^{k} (s+i)\exp(-as) = \frac{d^k}{dp^k}\left(\frac{1}{1-p}\right)\Big|_{p=\exp(-a)} = k! \frac{1}{(1-\exp(-a))^k}.$$

As a result, $\sum_{s=0}^{\infty}(s+1)^k\varphi(s) \le C\left(\frac{1}{(1-\exp(-a))}\right)^k k!$, so we take $L_1 = \frac{1}{(1-\exp(-a))}$ and $L_2 = C$. Lastly, we observe that employing another geometric series argument,

$$\mathrm{Var}(\bar{X}_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) \le \sum_{i=1}^{n} C \sum_{j=1}^{n} \rho^{|i-j|} \le nC\frac{2}{1-\rho}.$$

Thus, we take $C_0 = \frac{2C}{1-\rho}$ and apply Theorem B.1, which yields the first result. To prove Equation (B.3), the main idea is that we can apply Equation (B.2) to each sample mean $|\bar{X}_k^{(\pi)}|$, at which point the Equation (B.3) follows from a union bound.

To prove this, note that if we rearrange $(X_{\pi(1)}, \dots, X_{\pi(k)})$ into their "original order," then these variables satisfy the condition in Equation (B.1). Formally, let $A = \{\pi(1), \dots, \pi(k)\}$ and let $\nu : A \to A$ be the permutation such that $\nu(\pi(i)) > \nu(\pi(j))$ if and only if $i > j$, for $i, j \in [k]$. Then define $Y_i = X_{\nu(\pi(i))}$ for $i \in [k]$, and note that

$$|\operatorname{Cov}(Y_i, Y_j)| = |\operatorname{Cov}(X_{\nu(\pi(i))}, X_{\nu(\pi(j))})| \le C\rho^{|\nu(\pi(i))-\nu(\pi(j))|} \le C\rho^{|i-j|}.$$

Here, we observe that Y_i and Y_j are only $|\nu(\pi(i)) - \nu(\pi(j))|$ apart in the original sequence of $X_1, \dots, X_n$. By construction, applying $\nu(\pi(\cdot))$ to the whole sequence can only delete $n-k$ elements of the sequence without rearranging any of the remaining elements, so therefore Y_i and Y_j can be at most $|\nu(\pi(i)) - \nu(\pi(j))|$ apart in the subsequence $\{Y_1, \dots, Y_{n_k}\}$. Thus, $|i-j| \le |\nu(\pi(i)) - \nu(\pi(j))|$.

This analysis implies by Equation (B.2) that for any $\pi \in S_n$,

$$\mathbb{P}\left(\max_{n_0 \le k \le n} |\bar{X}_k^{(\pi)}| \ge t\right) \le \sum_{k=n_0}^{n} \mathbb{P}(|\bar{X}_k^{(\pi)}| \ge t) \le \sum_{k=n_0}^{n} \exp(-C' t^{1/4} k^{1/4}) \le n \exp(-C' t^{1/4} n_0^{1/4}).$$

This completes the proof. □

B.2 Bounds on the expected number of false discoveries

The proof of Theorem 2.1 relied on the fact that $\lim_{n\to\infty} \widetilde{\text{Power}}_q(w_n)$ is finite whenever it exists. This is a simple consequence of the following Lemma, proved below.

Lemma B.2. *For any data-generating process $(\mathbf{X}, \mathbf{y})$, any valid knockoffs $\widetilde{\mathbf{X}}$, any prior π, and any valid knockoff statistic $W = w([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$, $\widetilde{\text{Power}}_q(w) \le 1 + C(q)$ where $C(q)$ is a finite constant depending only on q.*

Proof. Fix any $q > 0$. If $L_\theta(\mathbf{y} \mid \mathbf{X})$ is the likelihood, we will prove this conditionally on θ. In particular, let $\mathcal{H}_0(\theta)$ denote the set of null features, let S be the rejection set and let $M(\theta) = p - |\mathcal{H}_0(\theta)|$ denote the number of non-nulls. If $V = |S \cap \mathcal{H}_0(\theta)|$ is the number of false discoveries, it suffices to show

$$\mathbb{E}\left[\frac{V}{M(\theta)} \mid \theta\right] \le C(q). \tag{B.4}$$

Proving B.4 proves the claim because it implies by the tower property that $\mathbb{E}[V] \le C(q)\mathbb{E}[M(\theta)]$. Therefore, since $|S| \le V + M(\theta)$, (B.4) implies

$$\widetilde{\text{Power}}_q(w) = \frac{\mathbb{E}[|S|]}{\mathbb{E}[M(\theta)]} \le 1 + \frac{\mathbb{E}[V]}{\mathbb{E}[M(\theta)]} \le 1 + C(q).$$

Thus, it suffices to prove (B.4), and the rest of the proof will hold conditional on θ, so we are essentially in the fully frequentist setting. Thus, for the rest of the proof, we will abbreviate $M(\theta)$ as M. We will also assume the "worst-case" values for the non-null coordinates of W_j: in particular, let W' denote W but with all of the non-null coordinates replaced with the value ∞, and let V'

be the number of false discoveries made when applying SeqStep to W'. These are the "worst-case" values in the sense that $V' \geq V$ deterministically (see Spector and Janson (2022), Lemma B.4), so it suffices to show that $\mathbb{E}[V'] \leq C(q)M$.

As notation, let $U = \text{sign}(\text{sorted}(W'))$ denote the signs of W' when sorted in descending order of absolute value. Following the notation in Equation (A.1), let $\psi(U) = \max\left\{k : \frac{k - k\bar{U}_k + 1}{k\bar{U}_k} \leq q\right\}$, where $\bar{U}_k = \frac{1}{k} = \sum_{i=1}^{\min(k,p)} U_i$. This ensures that $\tau := \left\lceil \frac{\psi(U)+1}{1+q} \right\rceil \leq \psi(U)$ is the number of discoveries made by knockoffs (Spector and Janson (2022), Lemma B.3). To prove the Lemma, we observe that it suffices to show $\mathbb{E}[\psi(U)] \leq C(q)$. To do this, let $K = \left\lceil \frac{M+1}{1+q} \right\rceil$ and fix any integer $c > 0$. Observe that

$$\begin{aligned}\mathbb{E}[\psi(U)] &\leq cK\mathbb{P}(\psi(U) \leq cK) + \sum_{k=cK}^{\infty} k\mathbb{P}(\psi(U) = k) \\ &\leq cK + \sum_{k=cK}^{\infty} k\mathbb{P}\left(\text{Bin}(k - M, 1/2) \geq \left\lceil \frac{k+1}{1+q} \right\rceil - M\right).\end{aligned}$$

where the second line follows because the event $\psi(U) = k$ implies that at least $\left\lceil \frac{k+1}{1+q} \right\rceil$ of the first k coordinates of U are positive. Crucially, the knockoff flip-sign property guarantees that conditional on θ, the null coordinates of U'_j are i.i.d. random signs conditional on the values of the non-null coordinates of U'. Thus, doing simple arithmetic, in the first k coordinates of U, there are $k - M$ null i.i.d. signs, of which at least $\left\lceil \frac{k+1}{1+q} \right\rceil - M$ must be positive, yielding the expression above. At this point, we will apply Hoeffding's inequality. In particular, choose any $c > \frac{1}{2}\left(\frac{1}{1+q} - \frac{1}{2}\right)^{-1}$, which ensures that for $k \geq cK$, $\frac{k-M}{2} \leq \left\lceil \frac{k+1}{1+q} \right\rceil - M$. Indeed, we can verify

$$\frac{k+1}{1+q} - M - \frac{k-M}{2} \geq k\left(\frac{1}{1+q} - \frac{1}{2}\right) - \frac{M}{2} \geq cM\left(\frac{1}{1+q} - \frac{1}{2}\right) - \frac{M}{2} \geq 0$$

where the last inequality follows by the choice of c. Having chosen this c, we may apply Hoeffding's inequality:

$$\mathbb{E}[\psi(U)] \leq cK + \sum_{k=cK}^{\infty} k\exp\left(-2\left(\left\lceil \frac{k+1}{1+q} \right\rceil - \frac{M}{2}\right)^2\right) \leq cK + \sum_{\ell=0}^{\infty}(\ell + cK)\exp\left(-\frac{2\ell^2}{(1+q)^2}\right).$$

Note that the sums $\sum_{\ell=0}^{\infty} \ell\exp(-2\ell^2/(1+q)^2)$ and $\sum_{\ell=0}^{\infty}\exp(-2\ell^2/(1+q)^2)$ are both convergent. As a result, $\mathbb{E}[\psi(U)]$ is bounded by a constant multiple of $cK \sim \frac{c}{1+q}M$, where the constant depends on q but nothing else. Since $\psi(U) \geq \tau \geq V' \geq V$ as previously argued, this completes the proof. □

C Comparison to the unmasked likelihood ratio

In this section, we compare MLR statistics to the earlier *unmasked* likelihood statistic introduced by Katsevich and Ramdas (2020), which this work builds upon. The upshot is that unmasked likelihood statistics give the most powerful "binary p-values," as shown by Katsevich and Ramdas

(2020), but do not yield jointly valid knockoff feature statistics in the sense required for the FDR control proof in Candès et al. (2018).

In particular, we call a statistic $T_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ a *marginally symmetric knockoff statistic* if T_j satisfies $T_j([\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(j)}, \mathbf{y}) = -T_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$. Under the null, T_j is marginally symmetric, so the quantity $p_j = \frac{1}{2} + \frac{1}{2}\mathbb{I}(T_j \leq 0)$ is a valid *"binary p-value"* which only takes values in $\{1/2, 1\}$. Theorem 5 of Katsevich and Ramdas (2020) shows that for any marginally symmetric knockoff statistic, $\mathbb{P}(p_j = 1/2) = \mathbb{P}(T_j > 0)$ is maximized if $T_j > 0 \Leftrightarrow L(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}_j, \mathbf{X}_{-j}) > L(\mathbf{y} \mid \mathbf{X}_j = \widetilde{\mathbf{x}}_j, \mathbf{X}_{-j})$. As such, one might initially hope to use the unmasked likelihood ratio as a knockoff statistic:

$$W_j^{\text{unmasked}} = \log\left(\frac{L_\theta(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}_j, \mathbf{X}_{-j})}{L_\theta(\mathbf{y} \mid \mathbf{X}_j = \widetilde{\mathbf{x}}_j, \mathbf{X}_{-j})}\right).$$

However, a marginally symmetric knockoff statistic is not necessarily a valid knockoff feature statistic, which must satisfy the following stronger property (Barber and Candès, 2015; Candès et al., 2018):

$$W_j([\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(J)}, \mathbf{y}) = \begin{cases} W_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}) & j \notin J \\ -W_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}) & j \in J, \end{cases}$$

for any $J \subset [p]$. This flip-sign property guarantees that the signs of the null coordinates of W are *jointly* i.i.d. and symmetric. However, the unmasked likelihood statistic does not satisfy this property, as changing the value of $\mathbf{X}_i$ for $i \neq j$ will often change the value of the likelihood $L_\theta(\mathbf{y} \mid \mathbf{X} = \mathbf{x}_j, \mathbf{X}_{-j})$.

D Methodological and computational details for MLR statistics

D.1 Why not use a plug-in estimator?

Since the oracle MLR statistics depend on any unknown parameters θ which affect the likelihood, one natural choice of feature-statistic would be to "plug in" an estimator $\hat{\theta}$ in place of θ, where $\hat{\theta}$ (e.g.) maximizes the regularized masked likelihood $L_\theta(D)$. In particular, we define the plug-in MLR statistic as

$$W_j \coloneqq \log\left(\frac{L_{\hat{\theta}}(\mathbf{X}_j = \mathbf{x}_j \mid D)}{L_{\hat{\theta}}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)}\right) \text{ for } \hat{\theta} = \arg\max_\theta L_\theta(D) + \mathcal{P}(\theta)$$

where $\mathcal{P}$ is some penalty on θ. In our explorations, we found that the plug-in statistic performed reasonably well, but not nearly as well as the MLR statistics defined in the paper. Our understanding of this phenomenon is that $L_\theta(D)$ is typically highly multimodal and non-convex. Intuitively, this is because the estimated value of θ depends very much on the value of $\mathbf{X}_j$, which is not known when we only observe D. For example, if $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, I_n)$, we should expect the estimate of β_j to be different in the two settings where (i) $\mathbf{X}_j = \mathbf{x}_j$ and (ii) $\mathbf{X}_j = \widetilde{\mathbf{x}}_j$. This has two consequences:

1. Computing $\hat{\theta} = \arg\max_\theta L_\theta(D)$ is likely to be expensive. That said, this challenge is not insurmountable, since we could use the EM algorithm to account for the fact that we do not know the value of $\mathbf{X}$ (in the model-X case) or $\mathbf{X}^T\mathbf{y}$ (in the fixed-X case).

2. More importantly, using the plug-in estimator $\hat{\theta}$ leads to lower power because it does not appropriately account for our uncertainty about $\hat{\theta}$. As an example, there are known settings in Gaussian linear models where the best estimate of β_j is positive when $\mathbf{X}_j = \mathbf{x}_j$ but the best estimate is negative when $\mathbf{X}_j = \widetilde{\mathbf{x}}_j$ (see Chen et al. (2019); Spector and Janson (2022)). It might be the case that the maximum likelihood of $\mathbf{y}$ is very similar whether or not we choose $\mathbf{X}_j = \mathbf{x}_j$ or $\mathbf{X}_j = \widetilde{\mathbf{x}}_j$, which would indicate that we cannot with confidence distinguish between the feature and the knockoff. However, perhaps the likelihood is *slightly* higher when we choose $\mathbf{X}_j = \widetilde{\mathbf{x}}_j$, so that the estimated coefficient $\hat{\beta}_j$ is highly negative. This will cause $L_{\hat{\beta}}(\mathbf{X}_j = \widetilde{\mathbf{x}}_j \mid D)$ to be much larger than $L_{\hat{\beta}}(\mathbf{X}_j = \mathbf{x}_j \mid D)$, since $\mathbf{x}_j$ has a positive linear relationship with $\mathbf{y}$, whereas $\hat{\beta}_j < 0$, leading to a large and negative value of the plug-in statistic W_j, which substantially reduces the power of knockoffs (see Figure 1). However, an MLR statistic which marginalizes over β would avoid this problem by also accounting for the case where $\beta_j > 0$.

To see formally that $L_\theta(D)$ is non-convex, let us consider a simple example where $\widetilde{\mathbf{X}}$ are fixed-X knockoffs and $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, I_n)$. In this setting, the masked likelihood has a particularly simple form, since usually, exactly computing the masked likelihood often requires summing over the 2^p possible values of $\mathbf{X}^T\mathbf{y}$. However, in the fixed-X case, it can be shown (Li and Fithian, 2021) that observing the masked data D is equivalent to observing $[\mathbf{X}, \widetilde{\mathbf{X}}]$ and the random variables $(|\tilde{\beta}|, \xi)$, where

$$\tilde{\beta} = S^{-1}(\mathbf{X} - \widetilde{\mathbf{X}})^T\mathbf{y} \text{ and } \xi = \frac{1}{2}(\mathbf{X} + \widetilde{\mathbf{X}})^T\mathbf{y}, \tag{D.1}$$

where $S \succ 0$ is the diagonal matrix satisfying $S = \mathbf{X}^T\mathbf{X} - \widetilde{\mathbf{X}}^T\mathbf{X}$. Both $\tilde{\beta}$ and ξ are Gaussian, so we can easily derive the masked log-likelihood in this setting. Indeed, let $A = \mathbf{X}^T\mathbf{X} - S/2$. Then ignoring additive constants,

$$\log L_\beta(D) = \beta^T\xi - \frac{1}{2}\beta^T A\beta - \sum_{j=1}^{p}\left[\frac{S_{j,j}\beta_j^2}{4} + \log\cosh\left(-\frac{1}{2}\beta_j|\tilde{\beta}_j|S_{j,j}\right)\right] \tag{D.2}$$

The function $x \mapsto ax^2 - \log\cosh(x)$ is convex for $a \geq \frac{1}{2}$, so the sum in (D.2) is concave if $\tilde{\beta}_j^2 \leq \frac{1}{2S_{j,j}}$ for all j. This is extremely unlikely to be the case, since $\tilde{\beta}_j \overset{\text{ind}}{\sim} \mathcal{N}\left(\beta_j, 2S_{j,j}^{-1}\right)$. Thus, even in the simple setting of fixed-X knockoffs, we should expect $\log L_\beta(D)$ to be highly multimodal.

D.2 Gibbs sampling updates for MLR statistics

In this section, we derive the Gibbs sampling updates for class of MLR statistics defined in Section 2.5.2. First, for convenience, we restate the model and choice of π.

D.2.1 Model and prior

First, we consider the model-X case. For each $j \in [p]$, let $\phi_j(\mathbf{X}_j) \in \mathbb{R}^{n \times p_j}$ denote any vector of prespecified basis functions applied to $\mathbf{X}_j$. We assume the following additive model:

$$\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2 \sim \mathcal{N}\left(\sum_{j=1}^{p}\phi_j(\mathbf{X}_j)\beta^{(j)}, \sigma^2 I_n\right)$$

with the following prior on $\beta^{(j)} \in \mathbb{R}^{p_j}$:

$$\beta^{(j)} \overset{\text{ind}}{\sim} \begin{cases} 0 \in \mathbb{R}^{p_j} & \text{w.p. } p_0 \\ \mathcal{N}(0, \tau^2 I_{p_j}) & \text{w.p. } 1 - p_0. \end{cases}$$

with the usual hyperpriors

$$\tau^2 \sim \text{invGamma}(a_\tau, b_\tau), \sigma^2 \sim \text{invGamma}(a_\sigma, b_\sigma) \text{ and } p_0 \sim \text{Beta}(a_0, b_0).$$

This is effectively a *group* spike-and-slab prior on $\beta^{(j)}$ which ensures group sparsity of $\beta^{(j)}$, meaning that either the whole vector equals zero or the whole vector is nonzero. We use this group spike-and-slab prior for two reasons. First, it reflects the intuition that ϕ_j is meant to represent only a single feature and thus $\beta^{(j)}$ will likely be entirely sparse (if $\mathbf{X}_j$ is truly null) or entirely non-sparse. Second, and more importantly, the group sparsity will substantially improve computational efficiency in the Gibbs sampler.

Lastly, for the fixed-X case, we assume exactly the same model but with the basis functions $\phi_j(\cdot)$ chosen to be the identity. Thus, this model is a typical spike-and-slab Gaussian linear model in the fixed-X case (George and McCulloch, 1997). It is worth noting that our implementation for the fixed-X case actually uses a slightly more general Gaussian mixture model as the prior on β_j, where the density $p(\beta_j) = \sum_{k=1}^m p_k \mathcal{N}(\beta_j; 0, \tau_k^2)$ for hyperpriors $\tau_0 = 0, \tau_k \overset{\text{ind}}{\sim} \text{invGamma}(a_k, b_k)$, and $(p_0, \ldots, p_m) \sim \text{Dir}(\alpha)$. However, for brevity, we only derive the Gibbs updates for the case of two mixture components.

D.2.2 Gibbs sampling updates

Following Section 2.5, we now review the details of the MLR Gibbs sampler which samples from the posterior of $(\mathbf{X}, \beta)$ given the masked data $D = \{\mathbf{y}, \{\mathbf{x}_j, \widetilde{\mathbf{x}}_j\}_{j=1}^p\}$.[6] As notation, let β denote the concatenation of $\{\beta^{(j)}\}_{j=1}^p$, let $\beta^{(-j)}$ denote all of the coordinates of β except those of $\beta^{(j)}$, let γ_j denote the indicator that $\beta^{(j)} \neq 0$, and let $\phi(\mathbf{X}) \in \mathbb{R}^{n \times \sum_j p_j}$ denote all of the basis functions concatenated together. Also note that although this section mostly uses the language of model-X knockoffs, when the basis functions $\phi_j(\cdot)$ are the identity, the Gibbs updates we are about to describe satisfy the sufficiency property required for fixed-X statistics, and indeed the resulting Gibbs sampler is actually a valid implementation of the fixed-X MLR statistic.

To improve the convergence of the Gibbs sampler, we slightly modify the meta-algorithm in Algorithm 1 to marginalize over the value of $\beta^{(j)}$ when resampling $\mathbf{X}_j$. To be precise, this means that instead of sampling $\mathbf{X}_j \mid \mathbf{X}_{-j}, \beta, \sigma^2$, we sample $\mathbf{X}_j \mid \mathbf{X}_{-j}, \beta^{(-j)}$. We derive this update in three steps, and along the way we derive the update for $\beta^{(j)} \mid \mathbf{X}, \beta^{(-j)}, D$.

Step 1: First, we derive the update for $\gamma_j \mid \mathbf{X}, \beta^{(-j)}, D$. Observe

$$\frac{\mathbb{P}(\gamma_j = 0 \mid \mathbf{X}, \beta^{(-j)}, D)}{\mathbb{P}(\gamma_j = 1 \mid \mathbf{X}, \beta^{(-j)}, D)} = \frac{p_0 p(\mathbf{y} \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} = 0)}{(1 - p_0) p(\mathbf{y} \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} \neq 0)}.$$

Analyzing the numerator is easy, as the model specifies that if we let $\mathbf{r} = \mathbf{y} - \phi(\mathbf{X}_{-j})\beta^{(-j)}$, then

$$p(\mathbf{y} \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} = 0) \propto \det(\sigma^2 I_n)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{r}\|_2^2\right).$$

[6] This is a standard derivation, but we review it here for the reader's convenience.

For the denominator, observe that $\mathbf{r}, \beta^{(j)} \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} \neq 0$ is jointly Gaussian: in particular,

$$(\beta^{(j)}, \mathbf{r}) \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} \neq 0 \sim \mathcal{N}\left(0, \begin{bmatrix} \tau^2 I_{p_j} & \tau\phi_j(\mathbf{X}_j)^T \\ \tau\phi_j(\mathbf{X}_j) & \tau^2\phi_j(\mathbf{X}_j)\phi_j(\mathbf{X}_j)^T + \sigma^2 I_n \end{bmatrix}\right). \tag{D.3}$$

To lighten notation, let $Q_j \coloneqq I_{p_j} + \frac{\tau^2}{\sigma^2}\phi(\mathbf{X}_j)^T\phi(\mathbf{X}_j)$. Using the above expression plus the Woodbury identity applied to the density of $\mathbf{y} \mid \mathbf{X}, \beta^{(-j)}, \beta^{(j)} \neq 0$, we conclude

$$\frac{\mathbb{P}(\gamma_j = 0 \mid \mathbf{X}, \beta^{(-j)}, D)}{\mathbb{P}(\gamma_j = 1 \mid \mathbf{X}, \beta^{(-j)}, D)} = \frac{p_0}{1-p_0}\det(Q_j)^{1/2}\exp\left(-\frac{\tau^2}{2\sigma^4}\mathbf{r}^T\phi_j(\mathbf{X}_j)Q_j^{-1}\phi_j(\mathbf{X}_j)^T\mathbf{r}\right).$$

Since Q_j is a $p_j \times p_j$ matrix, this quantity can be computed relatively efficiently.

Step 2: Next, we derive the distribution of $\beta^{(j)} \mid \mathbf{y}, \mathbf{X}, \beta^{(-j)}, \gamma_j$. Of course, the case where $\gamma_j = 0$ is trivial since then $\beta^{(j)} = 0$ by definition: in the alternative case, note from Equation (D.3) that we have

$$\beta^{(j)} \mid \mathbf{y}, \mathbf{X}, \beta^{(-j)}, \gamma_j = 1 \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2}\phi_j^T\mathbf{r} - \frac{\tau^4}{\sigma^4}\phi_j^T\phi_j Q_j^{-1}\phi_j^T\mathbf{r}, \tau^2 I_{p_j} - \frac{\tau^4}{\sigma^2}\phi_j^T\phi_j + \frac{\tau^6}{\sigma^4}\phi_j^T\phi_j Q_j^{-1}\phi_j^T\phi_j\right),$$

where above, we use ϕ_j as shorthand for $\phi_j(\mathbf{X}_j)$ to lighten notation.

Step 3: Lastly, we derive the update for $\mathbf{X}_j$ given $\mathbf{X}_{-j}, \beta^{(-j)}, D$. In particular, for any vector $\mathbf{x}$, let $\kappa(\mathbf{x}) \coloneqq \mathbb{P}(\gamma = 0 \mid \mathbf{X}_j = \mathbf{x}, \mathbf{X}_{-j}, \beta^{(-j)})$. Then by the law of total probability and the same Woodbury calculations as before,

$$\begin{aligned}
\mathbb{P}(\mathbf{X}_j = \mathbf{x} \mid \mathbf{X}_{-j}, \beta^{(-j)}, D) \propto & p(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}, \mathbf{X}_{-j}, \beta^{(-j)}) \\
&= \kappa(\mathbf{x})p(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}, \mathbf{X}_{-j}, \beta^{(-j)}, \beta^{(j)} = 0) \\
&+ (1-\kappa(\mathbf{x}))p(\mathbf{y} \mid \mathbf{X}_j = \mathbf{x}, \mathbf{X}_{-j}, \beta^{(-j)}, \beta^{(j)} \neq 0) \\
&\propto \kappa(\mathbf{x})\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{r}\|_2^2\right) \\
&+ (1-\kappa(\mathbf{x}))\det(Q_j(\mathbf{x}))^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{r}\|_2^2 + \frac{\tau^2}{2\sigma^4}\mathbf{r}^T\phi_j(\mathbf{x})^T Q_j(\mathbf{x})^{-1}\phi_j(\mathbf{x})^T\mathbf{r}\right) \\
&\propto \kappa(\mathbf{x}) + (1-\kappa(\mathbf{x}))\det(Q_j(\mathbf{x}))^{-1/2}\exp\left(\frac{\tau^2}{2\sigma^4}\mathbf{r}^T\phi_j(\mathbf{x})^T Q_j(\mathbf{x})^{-1}\phi_j(\mathbf{x})^T\mathbf{r}\right)
\end{aligned}$$

where above $Q_j(\mathbf{x}) = I_{p_j} + \frac{\tau^2}{\sigma^2}\phi_j(\mathbf{x})^T\phi_j(\mathbf{x})$ as before.

The only other sampling steps required in the Gibbs sampler are to sample from the conditional distributions of σ^2, τ^2 and p_0; however, this is straightforward since we use conjugate hyperpriors for each of these parameters.

D.2.3 Extension to binary regression

We can easily extend the Gibbs sampler in the preceding section to handle the case the response is binary via data-augmentation. Indeed, let us start by considering the case of Probit regression, which means we observe $\mathbf{z} = \mathbb{I}(\mathbf{y} \geq 0) \in \{0,1\}^n$ instead of the continuous outcome $\mathbf{y}$. Following Albert and Chib (1993), we note that distribution of $\mathbf{y} \mid \mathbf{z}, \mathbf{X}, \beta$ is truncated normal, namely

$$\mathbf{y}_i \mid \mathbf{z}, \mathbf{X}, \beta \overset{\text{ind}}{\sim} \begin{cases} \text{TruncNorm}(\mu_i, \sigma^2; (0, \infty)) & \mathbf{z}_i = 1 \\ \text{TruncNorm}(\mu_i, \sigma^2; (-\infty, 0) & \mathbf{z}_i = 0, \end{cases} \tag{D.4}$$

where $\mu = \phi(\mathbf{X})\beta = \mathbb{E}[\mathbf{y} \mid \mathbf{X}, \beta]$. Thus, when we observe a binary response $\mathbf{z}$ instead of the continuous response $\mathbf{y}$, we can employ the same Gibbs sampler as in Section D.2.2 except that after updating $\beta^{(j)} \mid \mathbf{X}, \beta^{(-j)}, \mathbf{y}$, we resample the latent variables $\mathbf{y}$ according to Equation (D.4), which takes $O(n)$ computation per iteration (since we can continuously update the value of μ whenever we update $\mathbf{X}$ or β in $O(n)$ operations as well). As a result, the computational complexity of this algorithm is the same as that of the algorithm in Section D.2.2. A similar formulation based on PolyGamma random variables is available for the case of logistic regression (see Polson et al. (2013)).

E MLR statistics for group knockoffs

In this section, we describe how MLR statistics extend to the setting of group knockoffs (Dai and Barber, 2016). In particular, for a partition $G_1, \dots, G_m \subset [p]$, group knockoffs allow analysts to test the *group* null hypotheses $H_{G_j} : X_{G_j} \perp\!\!\!\perp Y \mid X_{-G_j}$, which can be useful in settings where $\mathbf{X}$ is highly correlated and there is not enough data to discover individual null variables. In particular, knockoffs $\widetilde{\mathbf{X}}$ are model-X *group* knockoffs if they satisfy the *group* pairwise-exchangeability condition $[\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(G_j)} \overset{\text{d}}{=} [\mathbf{X}, \widetilde{\mathbf{X}}]$ for each $j \in [m]$. Similarly, $\widetilde{\mathbf{X}}$ are fixed-X group knockoffs if (i) $\mathbf{X}^T\mathbf{X} = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ and (ii) $S = \mathbf{X}^T\mathbf{X} - \widetilde{\mathbf{X}}^T\mathbf{X}$ is block-diagonal, where the blocks correspond to groups $G_1, \dots, G_m$. Given group knockoffs, one computes a single knockoff feature-statistic for each group.

MLR statistics extend naturally to the group knockoff setting because we can treat each group of features X_{G_j} as a single compound feature. In particular, the masked data for group knockoffs is

$$D = \begin{cases} (\mathbf{y}, \{\mathbf{X}_{G_j}, \widetilde{\mathbf{X}}_{G_j}\}_{j=1}^m) & \text{for model-X knockoffs} \\ (\mathbf{X}, \widetilde{\mathbf{X}}, \{\mathbf{X}_{G_j}^T\mathbf{y}, \widetilde{\mathbf{X}}_{G_j}^T\mathbf{y}\}_{j=1}^m) & \text{for fixed-X knockoffs,} \end{cases} \tag{E.1}$$

and the corresponding MLR statistics are

$$W_j^\star = \log\left(\frac{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_{G_j} = \mathbf{x}_{G_j} \mid D)\right]}{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_{G_j} = \widetilde{\mathbf{x}}_{G_j} \mid D)\right]}\right) \text{ for model-X knockoffs}$$

$$W_j^\star = \log\left(\frac{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_{G_j}^T\mathbf{y} = \mathbf{x}_{G_j}^T\mathbf{y} \mid D)\right]}{\mathbb{E}_{\theta\sim\pi}\left[L_\theta(\mathbf{X}_{G_j}^T\mathbf{y} = \widetilde{\mathbf{x}}_{G_j}^T\mathbf{y} \mid D)\right]}\right) \text{ for fixed-X knockoffs.}$$

Throughout the paper, we have proved several optimality properties of MLR statistics, and if we treat X_{G_j} as a single compound feature, all of these theoretical results (namely Proposition 2.3 and Theorem 2.1) immediately apply to group MLR statistics as well.

To compute group MLR statistics, we can use exactly the same Gibbs sampling strategy as in Section D.2—indeed, one can just treat X_{G_j} as a basis representation of a single compound feature and use exactly the same equations as derived previously.

F Additional details for the simulations

In this section, we describe the simulation settings in Section 3, and we also give the corresponding plot to Figure 7 which shows the results when $q = 0.05$. To start, we describe the simulation settings for each plot.

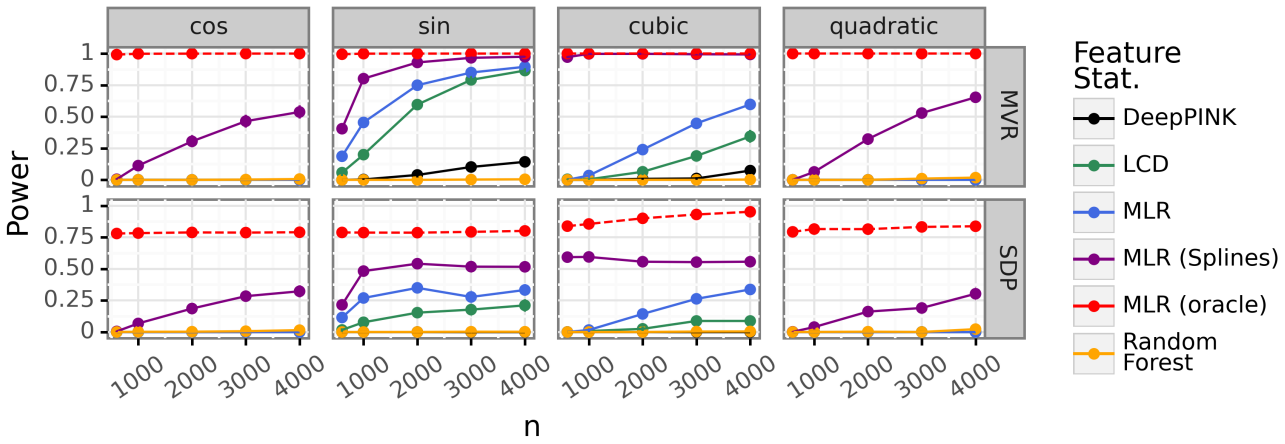

Figure 12: This plot is identical to Figure 7 except it shows the results for $q = 0.05$.

1. Sampling $\mathbf{X}$: We sample $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ in all simulations, with two choices of Σ. First, in the "AR(1)" setting, we take Σ to correspond to a nonstationary AR(1) Gaussian Markov chain, so $\mathbf{X}$ has i.i.d. rows satisfying $X_j \mid X_1, \ldots, X_{j-1} \sim \mathcal{N}(\rho_j X_{j-1}, 1)$ with $\rho_j \overset{\text{i.i.d.}}{\sim} \min(0.99, \text{Beta}(5, 1))$. Note that the AR(1) setting is the default used in any plot where the covariance matrix is not specified. Second, in the "ErdosRenyi" (ER) setting, we sample the random matrix V with entries $V_{ij} \overset{\text{i.i.d.}}{\sim} \text{Unif}((-1, -0.1) \cup (0.1, 1))$ and then set $\Sigma = (V + V^T) + (0.1 + \lambda_{\min}(V^T + V))I_p$ and then rescale Σ to be a correlation matrix.

2. Sampling β: Unless otherwise specified in the plot, we randomly choose $s = 10\%$ of the entries of β to be nonzero and sample the nonzero entries as i.i.d. $\text{Unif}([-\tau, -\tau/2] \cup [\tau/2, \tau])$ random variables with $\tau = 0.5$ by default. The exceptions are: (1) in Figure 5, we set $\tau = 0.3$, vary s between 0.05 and 0.4 as shown in the plot, and in some panels sample the non-null coefficients as Laplace(τ) random variables, (2) in Figure 7 we take $\tau = 2$ and $s = 0.3$, and (3) in Figure 8 we take $\tau = 1$.

3. Sampling $\mathbf{y}$: Throughout we sample $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, I_n)$, with only two exceptions. First, in Figure 7, we sample $\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(h(\mathbf{X})\beta, I_n)$ where h is a nonlinear function applied elementwise to $\mathbf{X}$, for $h(x) = \sin(x), h(x) = \cos(x), h(x) = x^2$ and $h(x) = x^3$. Second, in Figure 8, $\mathbf{y}$ is binary and $\mathbb{P}(Y = 1 \mid X) = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)}$.

4. Sampling knockoffs: We sample MVR and SDP Gaussian knockoffs using the default parameters from `knockpy` version 1.3, both in the fixed-X and model-X case. Note that in the model-X case, we use the true covariance matrix Σ to sample knockoffs, thus guaranteeing finite-sample FDR control.

5. Fitting feature statistics: We fit the following types of feature statistics throughout the simulations: LCD statistics, LSM statistics, a random forest with swap importances (Gimenez et al., 2019), DeepPINK (Lu et al., 2018), MLR statistics (linear variant), MLR statistics with splines, and the MLR oracle. In all cases we use the default hyperparameters from knockpy version 1.3, and we do not adjust the hyperparameters, so that the MLR statistics do not have well-specified priors. The exception is that the MLR oracle has access to the underlying data-generating process and the true coefficients β, which is why it is an "oracle."

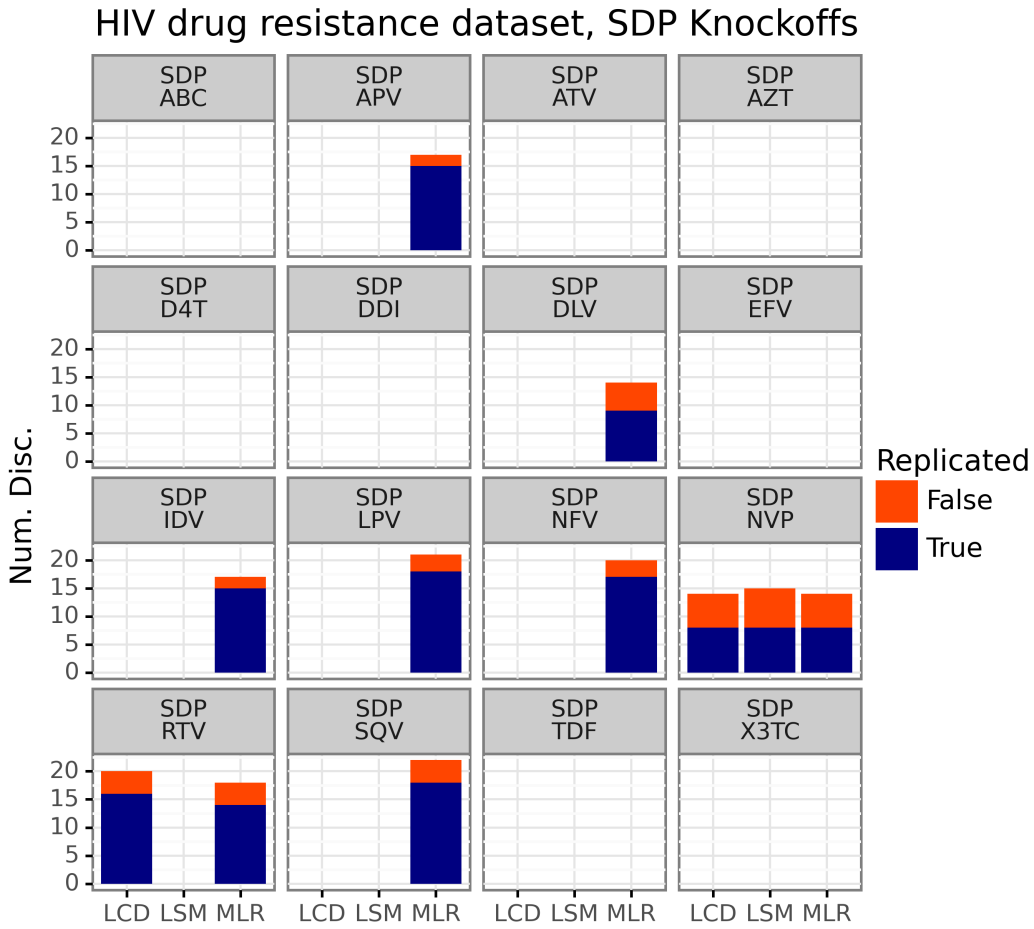

Figure 13: This figure shows the number of discoveries made by each feature statistic for each drug in the HIV drug resistance dataset.

Now, recall that in Figure 7, we showed the results for $q = 0.1$ because several competitor feature statistics made no discoveries at $q = 0.05$. Figure 12 is corresponding plot for $q = 0.05$.

G Additional results for the real data applications

G.1 HIV drug resistance

For the HIV drug resistance application, Figures 13 and 14 show the number of discoveries made by each feature statistic for SDP and MVR knockoffs, respectively, stratified by the drug in question. Note that the specific data analysis is identical to that of Barber and Candès (2015) and Fithian and Lei (2020) other than the choice of feature statistic—see either of those papers or `https://github.com/amspector100/mlr_knockoff_paper` for more details.

G.2 Financial factor selection

We now present a few additional details for the financial factor selection analysis from Section 4.2. First, we list the ten index funds we analyze, which are: XLB (materials), XLC (communication

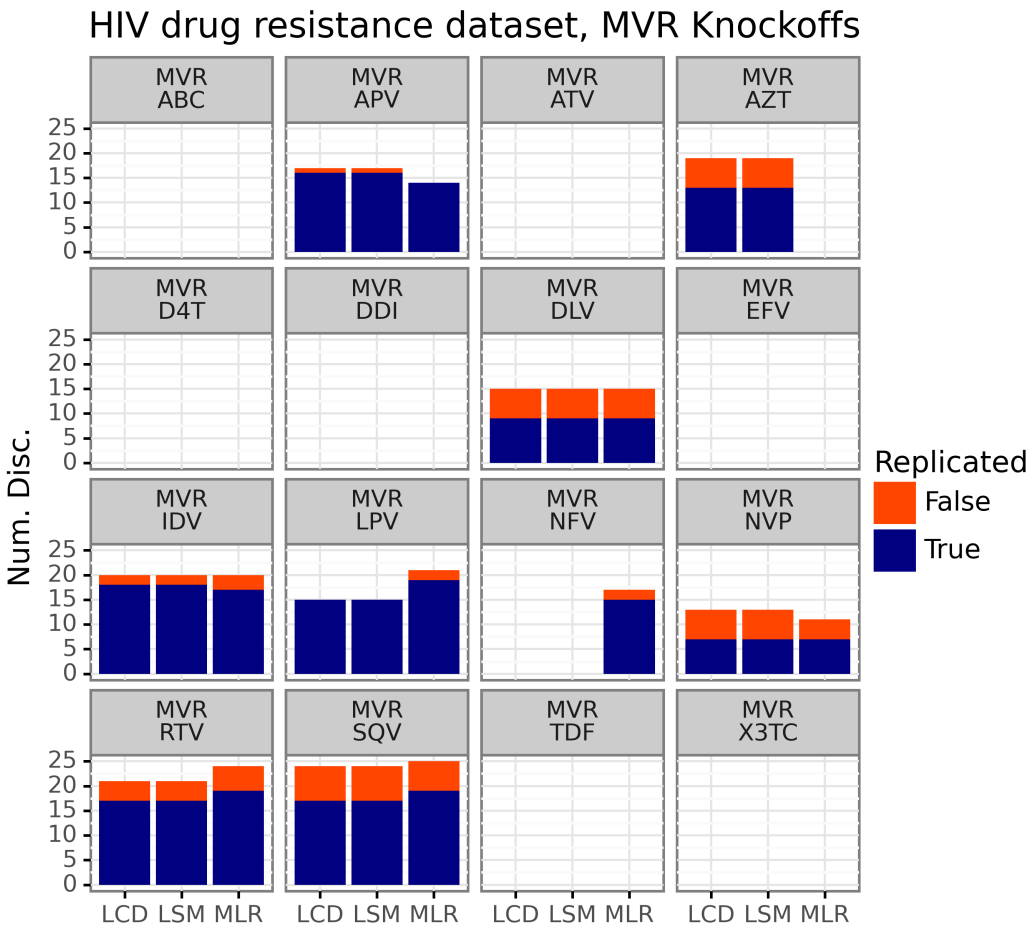

Figure 14: This figure shows the number of discoveries made by each feature statistic for each drug in the HIV drug resistance dataset.

Knockoff Type	Feature Stat.	Average FDP
MVR	LCD	0.013636
	LSM	0.004545
	MLR	0.038571
SDP	LCD	0.000000
	LSM	0.035000
	MLR	0.039002

Table 1: This table shows the average FDP, defined above, for each method in the financial factor selection analysis from Section 4.2.

services), XLE (energy), XLF (financials), XLK (information technology), XLP (consumer staples), XLRE (real estate), XLU (utilities), XLV (health care), and XLY (consumer discretionary). Second, for each feature statistic, Table 1 shows the average realized FDP across all ten analyses—as desired, the average FDP for each method is lower than the nominal level of $q = 0.05\%$.