JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–18
https://doi.org/10.1080/01621459.2022.2060113

False Discovery Rate Control via Data Splitting

Chenguang Dai*[a], Buyu Lin*[a], Xin Xing[b], and Jun S. Liu[a]

[a]Department of Statistics, Harvard University, Cambridge, MA; [b]Department of Statistics, Virginia Tech, Blacksburg, VA

ABSTRACT

Selecting relevant features associated with a given response variable is an important problem in many scientific fields. Quantifying quality and uncertainty of a selection result via false discovery rate (FDR) control has been of recent interest. This article introduces a data-splitting method (referred to as "DS") to asymptotically control the FDR while maintaining a high power. For each feature, DS constructs a test statistic by estimating two independent regression coefficients via data splitting. FDR control is achieved by taking advantage of the statistic's property that, for any null feature, its sampling distribution is symmetric about zero; whereas for a relevant feature, its sampling distribution has a positive mean. Furthermore, a Multiple Data Splitting (MDS) method is proposed to stabilize the selection result and boost the power. Surprisingly, with the FDR under control, MDS not only helps overcome the power loss caused by data splitting, but also results in a lower variance of the false discovery proportion (FDP) compared with all other methods in consideration. Extensive simulation studies and a real-data application show that the proposed methods are robust to the unknown distribution of features, easy to implement and computationally efficient, and are often the most powerful ones among competitors especially when the signals are weak and correlations or partial correlations among features are high. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Accepted March 2022

KEYWORDS
Data perturbation; FDR control; Feature selection; Graphical model; Linear model

1. Introduction

1.1. Background: FDR Control in Regression Models

Scientific researchers nowadays often have the privilege of collecting a large number of explanatory features targeting a specific response variable. For instance, population geneticists often profile thousands of single nucleotide polymorphisms (SNPs) in genome-wide association studies. A ubiquitous belief, however, is that the response variable depends only on a small fraction of the collected features. It is thus important to identify these relevant features so that the computability of the downstream analysis, the reproducibility of the reported results, and the interpretability of the scientific findings can be greatly enhanced. Throughout, we denote the explanatory features as $\{X_1, \ldots, X_p\}$, with p being potentially large, and denote the response variable as y.

Many advances in feature selection methods for regression analyses have been made in the past few decades, such as stepwise regressions (Efroymson 1960), Lasso regression (Tibshirani 1996), and Bayesian variable selection methods (O'Hara and Sillanpää 2009). A desirable property of a feature selection procedure is its capability of controlling the number of false positives, which can be mathematically formulated as controlling the false discovery rate (FDR) (Benjamini and Hochberg 1995), that is, keeping the FDR below a targeted level. FDR is defined as

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}], \quad \mathrm{FDP} = \frac{\#\{j : j \in S_0,\, j \in \widehat{S}\}}{\#\{j \in \widehat{S}\} \vee 1},$$

where S_0 denotes the index set of the null features (irrelevant features, see the formal definition in Section 2). $\widehat{S}$ denotes the set of the selected features, and FDP stands for "false discovery proportion." The expectation is taken with respect to the randomness in both the data and the selection procedure (if it is stochastic).

One popular class of FDR control methods is based on the Benjamin-Hochberg (BHq) procedure (Benjamini and Hochberg 1995). BHq requires p-values and guarantees exact FDR control when all the p-values are independent. Benjamini and Yekutieli (2001) generalized BHq to handle dependent p-values. They proved that BHq achieves FDR control under positive dependence, and also works under any arbitrary dependence structure if a shrinkage of the control level by $\sum_{j=1}^{p} 1/j$ is applied. Further discussions on generalizing BHq can be found in Sarkar (2002) for stepwise multiple testing procedures with positive dependence, Storey, Taylor, and Siegmund (2004) for weak dependence, Wu (2008) and Clarke and Hall (2009) for Markov models and linear processes.

Another class of powerful methods is based on the "knockoff filter" idea, which achieves FDR control by creating "knockoff" features in a similar spirit as adding spike-in controls in

CONTACT Jun S. Liu jliu@stat.harvard.edu Department of Statistics, Harvard University, Cambridge, MA.
*These authors contribute equally to this work.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

biological experiments instead of resorting to p-values. Barber and Candès (2015) first proposed the fixed-design knockoff filter, which achieves exact FDR control for low-dimensional Gaussian linear models regardless of the dependency structure among features. The model-X knockoff filter (Candès et al. 2018) further extends the applicability of knockoff filtering to high-dimensional problems and can be applied without having to know the underlying true relationship between the response and features. However, it requires the exact knowledge of the joint distribution of features. If this distribution is unknown, Barber, Candès, and Samworth (2020) showed that the inflation of the FDR is proportional to the estimation error in the conditional distribution of X_j given $\boldsymbol{X}_{-j} \stackrel{\text{def}}{=} \{X_1, \ldots, X_p\} \backslash \{X_j\}$. For details on how to generate good knockoff features, see Romano, Sesia, and Candès (2019), Jordon, Yoon, and Schaar (2019) (using deep generative models) and Bates et al. (2020) (using sequential MCMC algorithms). Huang and Janson (2020) generalized the model-X knockoff filter using conditioning to allow features following an exponential family distribution with unknown parameters. Further developments include the multilayer knockoff filter (Katsevich and Sabatti 2019) for FDR control at both group and individual levels, and DeepPINK (Lu et al. 2018) which models the relationship between the response and features by a neural network. Successful applications of the knockoff filter in genetics have been recently reported (Sesia, Sabatti, and Candès 2018; Sesia et al. 2020).

Data splitting has long been used for evaluating statistical predictions (e.g., cross-validation) (Stone 1974) and selecting efficient test statistics (Moran 1973; Cox 1975). Lately, data splitting has been employed to overcome difficulties in high-dimensional statistical inference. For example, Wasserman and Roeder (2009) proposed to split the data into three parts to implement a three-stage method: fits a suite of candidate models to the first part of the data; uses the second part of the data to select one of the models; and eliminates the null features based on hypothesis testing using the third part of the data. Other practices of data splitting in feature selection/multiple hypotheses testing can be found in Rubin, Dudoit, and der Laan (2006) (estimating the optimal cutoff for test statistics), Meinshausen, Meier, and Bühlmann (2009) (aggregating p-values obtained via repeated sample splitting), and Ignatiadis et al. (2016) (determining proper weights for individual hypotheses). More recently, Barber and Candès (2019) extended the applicability of the fixed-design knockoff filter to high-dimensional linear models via data splitting, in which the first part of the data is used to screen out enough null features so that the knockoff filter can be applied to the selected features using the second part of the data.

FDR control introduced by Benjamini and Hochberg (1995) is formulated as a sampling property of the procedure. All aforementioned methods including our proposed ones take this frequentist point of view. Empirical Bayes views of FDR control have also been studied in the literature, such as the local FDR control method of Efron (2005), which has been successfully applied to analyze microarray data (Efron et al. 2001). The local FDR control framework is more delicate in the sense that it attaches each hypothesis/feature a probabilistic quantification of being null based on efficient density estimation of the test statistics. It is worth noting that there is also a Bayesian interpretation of the positive FDR (i.e., $\mathbb{E}[\text{FDP}||\widehat{S}| > 0]$), as pointed out by Storey (2003).

1.2. Main Ideas

In high-dimensional regression, it can be challenging to either construct valid p-values (even asymptotically) or estimate accurately the joint distribution of features, thus, limiting the applicability of both BHq and the model-X knockoff filter. The data-splitting framework proposed here appears to fill in this gap. We focus here on a single data-splitting procedure (DS, henceforth) and its refinement the multiple data-splitting procedure (MDS, henceforth).

Although the motivation of most existing data-splitting methods is to handle the high dimensionality (e.g., to obtain valid p-values or apply the fixed-design knockoff filter), DS aims at obtaining two independent regression coefficients for each feature via two potentially different statistical fitting procedures applied to each part of the data. Using the two estimates, DS then constructs a test statistic M_j for each feature X_j, referred to as the "mirror statistic" or "symmetric statistic" in the literature, which should possess the following two key properties as illustrated in Figure 1:

(P1) A feature with a larger positive mirror statistic is more likely to be a relevant feature.
(P2) The sampling distribution of the mirror statistic of any null feature is symmetric about zero.

Property (P1) suggests that we can rank the importance of each feature by its mirror statistic, and select those features with mirror statistics larger than a cutoff (τ in Figure 1). Property (P2) implies that we can estimate (conservatively) the number of false positives, that is, $\#\{j : j \in S_0,\ M_j > \tau\}$, using the left tail of the distribution, $\#\{j : M_j < -\tau\}$.

When $p < n/2$, coefficient estimates for both parts/splits of the data can be obtained by the ordinary least squares (OLS). For high-dimensional cases, we can implement a Lasso+OLS procedure, which screens out some null features by Lasso using one part of the data and obtains the OLS estimates for the selected features using the other part of the data. The mirror statistics are then constructed using the Lasso and OLS estimates.

Although sharing similar properties (see Section 3.2 in Candès et al. 2018), the mirror statistic and the knockoff statistic are motivated by different philosophies. Specifically, the mirror statistic is free from constructing the matching "fake," which can be challenging and computationally expensive in high-dimensional settings. This also leads to different use of conditions and theoretical implications. Our framework relies on the model assumption of $y|\boldsymbol{X}$ for asymptotic FDP and FDR control, whereas the knockoff filter requires (almost) the exact knowledge of the joint distribution of features to achieve finite sample FDR control. DS also differs from the work on inference after selection (Berk et al. 2013; Lockhart et al. 2014; Lee et al. 2016), which attempts to construct p-values and confidence intervals for the selected features conditioning on the selected model. The objects of the inference thus depend on the initial selection step. In contrast, DS (including its refinement MDS) aims at selecting the relevant features with a reasonably low

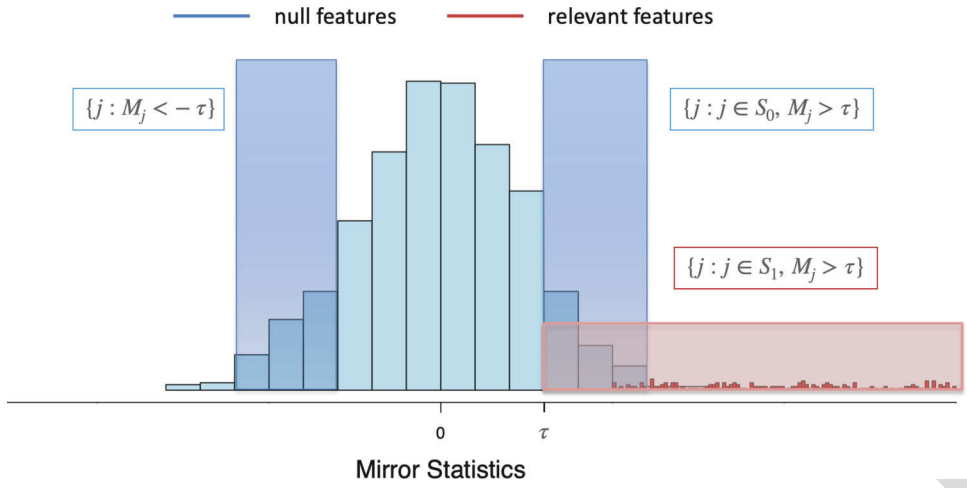

Figure 1. A cartoon illustration of the mirror statistic. M_j denotes the mirror statistic of feature X_j. S_0 and S_1 denote the index set of the null features and the relevant features, respectively. Features with mirror statistics larger than the cutoff τ are selected.

FDR, and does not require two separate steps for selection and inference adjustment.

MDS is built upon multiple independent replications of DS, aiming at reducing the variability of the selection result. Instead of ranking features by their mirror statistics, MDS ranks features by their *inclusion rates*, that is, the selection frequency adjusted by the selection sizes among multiple DS replications. Empirically, we observe that MDS simultaneously reduces the FDR and boosts the power, suggesting that MDS yields better rankings of features than DS. We provide some useful insights on MDS by analyzing the Normal means model and prove that MDS achieves nearly the optimal detection power (see Section 2.3). MDS is conceptually most similar to the stability selection method (Meinshausen and Bühlmann 2010), and a more detailed discussion on them is deferred to Section 2.2. MDS also differs from the recently proposed de-randomizing procedure for the model-X knockoff filter (Ren et al. 2020), whose goal is to control the per family error rate (PFER) and the k family-wise error rate (k-FWER). Methods designed for linear models are also applicable to Gaussian graphical models using the linear representation of its conditional dependence structure (Lauritzen 1996). Given a nominal level q, we apply DS or MDS to each nodewise regression targeting at an FDR control level $q/2$, and then combine the nodewise selection results using the "OR" rule (Meinshausen and Bühlmann 2006). We show that both DS and MDS achieve FDR control for linear and graphical models under standard assumptions including sparsity conditions, regularity conditions on the design matrix, and signal strength conditions.

The rest of the article is organized as follows. Section 2.1 introduces DS with a detailed discussion on the construction of the mirror statistic. Sections 2.2 and 2.3 focus on MDS, in which we show that MDS achieves nearly the optimal detection power for the Normal means model. The desired FDR control properties for DS and MDS in a model-free setting are also proved in Section 2 under certain conditions. Section 3 discusses applications of DS and MDS to linear and Gaussian graphical models. Sections 4.1 and 4.2 demonstrate through extensive simulations that DS and MDS control the FDR properly, and MDS achieves the best or near-best power in almost all cases for linear and graphical models. Section 4.3 applies DS and MDS to the task of identifying mutations associated with drug resistance using an HIV-1 dataset. Section 5 concludes with a few final remarks. Proofs and more simulation details are deferred to supplementary materials. An R implementation of DS and MDS can be found at here.

2. Data Splitting for FDR Control

2.1. Single Data Splitting

Suppose a set of features $(X_1, \ldots, X_p)$ follows a p-dimensional distribution. Denote the n independent observations of these features as $\boldsymbol{X}_{n\times p} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, also known as the *design matrix*, where $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})^\intercal$ is the vector containing n independent realizations of feature X_j. This random-design assumption is nonessential and our methods also apply to the fixed-design scenario (see Remark 3.1). We assume that all features except the intercept (a vector with all 1's) have been normalized to have zero mean and unit variance. For each set of the observed features $(X_{i1}, \ldots, X_{ip})$, there is an associated response variable y_i for $i \in \{1, \ldots, n\}$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ be the vector of n independent responses. We assume that the response variable y only depends on a subset of features $\boldsymbol{X}_{S_1} = \{X_j : j \in S_1\}$, and the task of feature selection is to identify the set S_1. Mathematically, a feature X_j is considered to be irrelevant or null if and only if $y \perp X_j | \boldsymbol{X}_{-j}$ following the definition in Candès et al. (2018). Let S_0 be the index set of the null features, and let $p_0 = |S_0|$, $p_1 = |S_1|$ be the numbers of the null and relevant features, respectively.

Feature selection commonly relies on a set of coefficients $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\intercal$ to measure the importance of each feature. The larger $|\widehat{\beta}_j|$ is, the more likely feature X_j is useful in predicting y (since features have been normalized). For example, in linear regressions, $\widehat{\boldsymbol{\beta}}$ can be the vector of coefficients estimated via OLS or some shrinkage methods. In contrast to common practices that select features based on a single set of estimated coefficients, we construct *two* independent sets of estimates, $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$, potentially with two different statistical procedures, in order to set up an FDR control framework. The independence between

$\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ is ensured by employing a data-splitting strategy. More precisely, we split the n observations into two groups, denoted as $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$, and $(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)})$, and then estimate $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ using each part of the data. The data-splitting procedure is flexible, as long as it is independent of the response vector $\boldsymbol{y}$. The sample sizes for the two groups can also be potentially different. Empirically we find that the half-half sample splitting leads to the highest power. To achieve FDR control, the two sets of coefficients shall satisfy the following assumption besides being independent.

Assumption 2.1 (Symmetry). For each null feature index $j \in S_0$, the sampling distribution of at least one of $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$ is symmetric about zero.

Note that the symmetry assumption is only required for the null features and can be further relaxed to asymptotic symmetry. Furthermore, for $j \in S_0$, it is sufficient that only one of $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$ is symmetric about zero. In Section 3, we propose a Lasso+OLS procedure for linear and Gaussian graphical models so that the symmetry assumption can be satisfied with probability approaching one under certain conditions. Our FDR control framework starts with the construction of a *mirror statistic* that satisfies Properties (P1) and (P2) as discussed in Section 1.2.

A general form of the mirror statistic M_j is

$$M_j = \operatorname{sign}\big(\widehat{\beta}_j^{(1)}\widehat{\beta}_j^{(2)}\big) f\big(|\widehat{\beta}_j^{(1)}|, |\widehat{\beta}_j^{(2)}|\big), \tag{1}$$

where function $f(u, v)$ is nonnegative, symmetric with respect to u and v, and monotonically increasing in both u and v. For a relevant feature, the two coefficient estimates tend to be large in magnitude and have the same sign if the estimation procedures are reasonably efficient. Since $f(u, v)$ is monotonically increasing in both u and v, the corresponding mirror statistic tends to be positive and relatively large, which implies Property (P1). In addition, the independence between the two coefficients, together with the symmetry assumption, imply Property (P2).

Lemma 2.1. Under Assumption 2.1, regardless of the data-splitting procedure, the sampling distribution of M_j is symmetric about zero for $j \in S_0$.

The proof is elementary and thus omitted. Three convenient choices of $f(u, v)$ are:

$$f(u, v) = 2\min(u, v), \quad f(u, v) = uv, \quad f(u, v) = u + v. \tag{2}$$

The first choice equals to the mirror statistic proposed in Xing, Zhao, and Liu (2021), and the third choice corresponds to the "sign-max" of $|\widehat{\beta}_j^{(1)} + \widehat{\beta}_j^{(2)}|$ and $|\widehat{\beta}_j^{(1)} - \widehat{\beta}_j^{(2)}|$, and is optimal in a simplified setting as described in Proposition 2.1.

Proposition 2.1. Consider a prototype model in which (a) for $j \in S_0$, the two coefficients $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$ follow $N(0, 1)$ independently; (b) for $j \in S_1$, the two coefficients $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$ follow $N(\omega, 1)$ independently; (c) for $k \in \{1, 2\}$, the set of coefficients $\{\widehat{\beta}_1^{(k)}, \ldots, \widehat{\beta}_p^{(k)}\}$ are weakly correlated in the sense that $||R^{(k)}||_1 \to 0$ as $p \to \infty$, where $R^{(k)}$ is the correlation matrix of $\widehat{\boldsymbol{\beta}}^{(k)}$ and $||R^{(k)}||_1 = p^{-2}\sum_{i,j}|R_{ij}^{(k)}|$; and (d) $p_1/p_0 \to r > 0$ as $p \to \infty$. Then, $f(u, v) = u + v$ is the optimal choice that yields the highest power.

The proof of Proposition 2.1 (see supplementary materials) might be of independent interest. We rephrase the FDR control problem under the hypothesis testing framework and prove the optimality using the Neyman–Pearson lemma. The form $f(u, v) = u+v$ is derived based on the rejection rule of the corresponding likelihood ratio test. The intention of Proposition 2.1 is to give readers some intuitions and insights on what may be optimal in certain canonical cases. The optimality of the sign-max mirror statistic for the knockoff filter has also been empirically observed by Barber and Candès (2015) and recently proved by Ke, Liu, and Ma (2020) based on a more delicate analysis under the weak-and-rare signal setting (Donoho and Jin 2004). For linear models in more realistic settings (see Section 4.1), we empirically compare the three choices of $f(u, v)$ listed in (2) and observe that the optimality of $f(u, v) = u + v$ holds broadly across various settings.

The symmetry property of the mirror statistics for the null features gives us an upper bound of the number of false positives:

$$\#\{j \in S_0 : M_j > t\} \approx \#\{j \in S_0 : M_j < -t\} \le \#\{j : M_j < -t\}, \quad \forall\, t > 0. \tag{3}$$

The FDP(t) of the selection $\widehat{S}_t = \{j : M_j > t\}$, as well as an "over estimate" of it, referred to as $\widehat{\text{FDP}}(t)$ in the following, are thus given by

$$\text{FDP}(t) = \frac{\#\{j : M_j > t,\, j \in S_0\}}{\#\{j : M_j > t\} \vee 1}, \quad \widehat{\text{FDP}}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1}.$$

For any designated FDR control level $q \in (0, 1)$, we can choose the data-driven cutoff τ_q as follows:

$$\tau_q = \min\{t > 0 : \widehat{\text{FDP}}(t) \le q\},$$

and the final selection is $\widehat{S}_{\tau_q} = \{j : M_j > \tau_q\}$. This data-driven cutoff and the final selection set are motivated by the knockoff filter (Barber and Candès 2015; Candès et al. 2018). The proposed FDR control procedure is summarized in Algorithm 1.

In order to obtain a good estimate of the number of false positives via (3), the mirror statistics of the null features cannot be too correlated. We thus require the following weak dependence assumption, which holds with high probability in certain common settings (see Section 3).

Assumption 2.2 (Weak dependence). The mirror statistics $M_j's$ are continuous random variables, and there exist constants $c > 0$ and $\alpha \in (0, 2)$ such that

$$\operatorname{var}\Big(\sum_{j\in S_0} \mathbb{1}(M_j > t)\Big) \le c p_0^{\alpha}, \quad \forall\, t \in \mathbb{R}, \quad \text{where } p_0 = |S_0|.$$

Assumption 2.2 only restricts the correlations among the null features in consideration, regardless of the correlations associated with the relevant features.

Algorithm 1 False discovery rate control via a single data split.

1. Split the data into two groups $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$ and $(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)})$, independent of the response $\boldsymbol{y}$.
2. Estimate the coefficients $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ using each part of the data. The two estimation procedures can be different.
3. Calculate the mirror statistics following (1).
4. For a nominal FDR level $q \in (0, 1)$, select the features $\{j : M_j > \tau_q\}$ where the cutoff τ_q is

$$\tau_q = \min\left\{t > 0 : \widehat{\mathrm{FDP}}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \le q\right\}. \tag{4}$$

In Section 3.1, we show that for linear models, this assumption holds as long as the covariance matrix of the null features satisfies some regularity condition (e.g., the eigenvalues are doubly bounded). A significant case that Assumption 2.2 is violated is when the null features can be partitioned into a fixed number of groups so that the within-group pairwise correlations of their mirror statistics are a constant. Empirically, however, we observe that even in these cases, DS and MDS still perform well, often outperforming BHq and the knockoff filter (see Figure 4 in supplementary materials).

Recall that $\mathrm{FDP}(t)$ refers to the FDP of the selection $\widehat{S}_t = \{j : M_j > t\}$. The proposition below shows that for any nominal level $q \in (0, 1)$, both $\mathrm{FDP}(\tau_q)$ and $\mathrm{FDR}(\tau_q)$ are under control, where τ_q is the data-dependent cutoff chosen following (4). Note that we require that $p_0 \to \infty$ as $p \to \infty$; otherwise, the FDR control problem becomes trivial as we can just select all.

Proposition 2.2. Suppose $\mathrm{var}(M_j)$ is uniformly upper bounded and also lower bounded away from zero. For any nominal FDR level $q \in (0, 1)$, assume that there exists a constant $t_q > 0$ such that $\mathbb{P}(\mathrm{FDP}(t_q) \le q) \to 1$ as $p \to \infty$. Then, under Assumptions 2.1 and 2.2, the DS procedure in Algorithm 1 satisfies

$$\mathrm{FDP}(\tau_q) \le q + o_p(1) \quad \text{and} \quad \limsup_{p\to\infty} \mathrm{FDR}(\tau_q) \le q.$$

We note that the existence of $t_q > 0$ such that $\mathbb{P}(\mathrm{FDP}(t_q) \le q) \to 1$ essentially guarantees the asymptotic feasibility of FDR control based upon the rankings of features by their mirror statistics. Similar assumptions also appear in Storey, Taylor, and Siegmund (2004) and Wu (2008) in order to achieve a high level of generality. Specifically, it ensures that the data-dependent cutoff τ_q is upper bounded with probability approaching one, which implies that $\liminf p_1/p_0 > 0$ given $\mathrm{var}(M_j)$ being bounded. This may be an undesired assumption for high-dimensional problems as it rules out the sparse case $p_1 \ll p_0$. However, this is only a technical assumption for handling the most general setting without specifying a parameteric model between the response and features. When we apply DS to specific models such as linear or Gaussian graphical models, τ_q is allowed to diverge and this assumption can often be avoided (see Section 3.1).

In Assumption 2.1, the exact symmetry can be relaxed to *uniform asymptotic symmetry*. That is, for $j \in S_0$, if the sampling distribution of either $\widehat{\beta}_j^{(1)}$ or $\widehat{\beta}_j^{(2)}$ is asymptotically symmetric about zero, and the resulting mirror statistics satisfy the uniformity condition:

$$\max_{j\in S_0} \left|\mathbb{P}(M_j > t) - \mathbb{P}(M_j < -t)\right| \to 0, \quad \forall\, t,$$

Proposition 2.2 still holds. For high-dimensional generalized linear models, one way to construct the mirror statistic is to use the debiased Lasso estimator (Van de Geer et al. 2014; Zhang and Zhang 2014; Javanmard and Montanari 2014), which is asymptotically Normal (thus symmetric) under certain conditions. Further, the bias in the debiased Lasso estimator also vanishes uniformly (see Dai et al. 2020 for more details).

Before concluding this section, we remark that DS is inspired by the recently proposed Gaussian mirror method (Xing, Zhao, and Liu 2021), whose main idea is to perturb features one by one and examines the corresponding impact. Compared to the Gaussian mirror method, DS is easier to implement and computationally more efficient especially for large n and p. For linear models, the Gaussian mirror method requires p linear fittings. In contrast, DS perturbs all features simultaneously via data splitting, thus, requiring only two linear fittings. The gain of the computational efficiency can be more significant for graphical models (see Section 3.2). DS requires $2p$ nodewise linear fittings, whereas the Gaussian mirror method would require p^2 nodewise linear fittings, which is generally unacceptable when p is large. In addition, it is more convenient to adapt DS to other statistical models thanks to its conceptual simplicity.

2.2. Multiple Data Splitting

There are two main concerns about DS. First, splitting the data inflates the variances of the estimated regression coefficients, thus, DS can potentially suffer from a power loss in comparison with competing methods that properly use the full data. Second, the selection result of DS may not be stable and can vary substantially across different sample splits.

To remedy these issues, we propose a multiple data-splitting (MDS) procedure to aggregate the selection results obtained from independent replications of DS. For linear and Gaussian graphical models, we prove that MDS achieves asymptotic FDR control under certain conditions. Simulation results in Section 4 confirm this and demonstrate a fairly universal power improvement of MDS over DS. Going beyond these two models, we empirically found that MDS can work competitively for a much wider class of models, and is also generally applicable without requiring p-values or any knowledge regarding the joint distribution of features.

Given $(\boldsymbol{X}, \boldsymbol{y})$, suppose we independently repeat DS m times with random sample splits. Each time the set of the selected

features is denoted as $\widehat{S}^{(k)}$ for $k \in \{1, \ldots, m\}$. For each feature X_j, we define the associated *inclusion rate* I_j and its estimate $\widehat{I}_j$ as

$$I_j = \mathbb{E}\left[\frac{\mathbb{1}(j \in \widehat{S})}{|\widehat{S}| \vee 1} \,\middle|\, \boldsymbol{X}, \boldsymbol{y}\right], \quad \widehat{I}_j = \frac{1}{m}\sum_{k=1}^{m} \frac{\mathbb{1}(j \in \widehat{S}^{(k)})}{|\widehat{S}^{(k)}| \vee 1}, \tag{5}$$

in which the expectation is taken with respect to the randomness in data splitting. Note that this rate is not an estimate of the probability of being selected, but rather an importance measurement of each feature relative to the DS procedure. For example, if feature X_j is always selected by DS, and DS always selects 20 features in each of the m independent replications, the inclusion rate I_j equals to 1/20. MDS ranks the importance of features by their inclusion rates, and is most useful if the following informal statement is approximately true: if a feature is selected less frequently in repeated sample splitting, it is less likely to be a relevant feature. If this holds, we can choose a proper inclusion-rate cutoff to control the FDR, as detailed in Algorithm 2.

Algorithm 2 Aggregating the selection results from multiple DS replications.

1. Sort the estimated inclusion rates (see (5)): $0 \le \widehat{I}_{(1)} \le \widehat{I}_{(2)} \le \cdots \le \widehat{I}_{(p)}$.
2. For a nominal FDR level $q \in (0, 1)$, find the largest $\ell \in \{1, \ldots, p\}$ such that $\widehat{I}_{(1)} + \cdots + \widehat{I}_{(\ell)} \le q$.
3. Select the features $\widehat{S} = \{j : \widehat{I}_j > \widehat{I}_{(\ell)}\}$.

Algorithm 2 suggests a backtracking approach to select the cutoff based on the following argument: if we had m independent datasets with m large enough, and applied DS to all of them for feature selection, the average FDP would be no larger than the designated FDR control level q. Although it is not possible to generate new data, we can consider $\{\widehat{S}^{(k)},\ k = 1, \ldots, m\}$ as an approximation to m independent selection results obtained via data regeneration. We thus, find the largest cutoff such that, if we assume the features with inclusion rates larger/smaller than the cutoff are "true" relevant/null features, respectively, the average FDP among $\{\widehat{S}^{(k)},\ k = 1, \ldots, m\}$ is no larger than q. Empirically, we find that MDS often results in a lower FDR than the nominal level but still enjoys a competitive power. The proposition below gives some intuitions regarding how MDS controls the FDR properly.

Proposition 2.3. Suppose DS asymptotically controls the FDP for any designated level $q \in (0, 1)$. Further, we assume that with probability approaching one, the power of DS is bounded below by some $\kappa > 0$. We consider the following two regimes with $n, p \to \infty$ at a proper rate.

(a) In the non-sparse regime where $\liminf p_1/p > 0$, we assume that the mirror statistics are consistent at ranking features, that is, $\sup_{i \in S_1, j \in S_0} \mathbb{P}(I_i < I_j) \to 0$.
(b) In the sparse regime where $\limsup p_1/p = 0$, we assume that the mirror statistics are strongly consistent at ranking features, that is, $\sup_{i \in S_1} \mathbb{P}(I_i < \max_{j \in S_0} I_j) \to 0$.

Then, for MDS (see Algorithm 2) in both the non-sparse and the sparse regimes, we have

$$\mathrm{FDP} \le q + o_p(1) \quad \text{and} \quad \limsup_{n,p \to \infty} \mathrm{FDR} \le q.$$

Remark 2.1. For any $i \in S_1$, $j \in S_0$, we have $\mathbb{P}(I_i < I_j) \le \mathbb{P}(I_i < \max_{j \in S_0} I_j)$, thus, the condition in Proposition 2.3(b) is stronger than the condition in (a). Besides, in the sparse regime where $p_0 \gg p_1$, we can show that the number of false positives is in the order of $o_p(p_1)$ under the strongly consistent condition.

Although some conditions in Proposition 2.3 are not directly verifiable without invoking further modeling assumptions, the proposition points out a key requirement for MDS to achieve FDR control: The ranking consistency of the baseline algorithm. In Section 3, we show that the ranking consistency condition holds for linear and graphical models under more explicit conditions.

The idea of replicating a procedure multiple times on perturbed data so as to stabilize the inference result and quantify uncertainty is not new. For example, Meinshausen and Bühlmann (2010) proposed a stability selection method, which subsamples the data and runs a feature selection algorithm multiple times using a set of regularization parameters. The final selection set contains only "stable" features, of which the selection frequencies are above some user-specified threshold. While the stability selection method aims at bypassing the difficulty of finding a proper regularization parameter in high-dimensional regressions, MDS is designed to stabilize DS and compensate for the power loss due to sample splitting. Theoretically, the stability selection method guarantees a finite-sample bound on the number of false positives under certain conditions, whereas MDS asymptotically controls the perhaps more delicate FDR. Indeed, MDS requires a careful selection of the inclusion-rate cutoff in order to achieve FDR control. In comparison, the selection-probability cutoff of the stability selection method can be much less stringent. Furthermore, the two methods perturb the data in different ways. For each regularization parameter, the stability selection method obtains a collection of selection sets using different subsamples of the data, whereas MDS always uses the full data, but replicates DS with independent sample splits.

There is also some relevant literature on p-value aggregation in high-dimensional settings. For example, Meinshausen, Meier, and Bühlmann (2009) proposed to obtain a collection of p-values via repeated sample splitting. The multiple p-values of each feature are then aggregated by choosing a proper quantile among them. Based upon that, BHq can be applied to control the FDR. Empirically, we found that the resulting procedure is often too conservative, with a near-zero FDR but also a suboptimal power (see Section 4.1). For other related works, we refer the readers to van de Wiel et al. (2009) and Romano and DiCiccio (2019).

The choice of m. Since our empirical studies suggest that the power of MDS monotonically increases with respect to m, it is always harmless to try a larger m when the computational budget permits. However, we never found it necessary to have a very large m, and a relatively small number of DS replications (say, $m = 50$) is typically good enough, after which the power of MDS no longer improves much. See Figure 1 in supplementary

materials and Figure 4 in Section 4.1 for empirical evidences on the Normal means model and linear models, respectively. For the Normal means model, Figure 2 shows that increasing m from 400 to 10,000 leads to only slightly less noisy feature rankings.

2.3. A Theoretical Study of MDS for the Normal Means Model

We consider the Normal means model to gain some insights on how MDS compensates for the power loss of DS due to sample splitting. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, we assume that X_{ij} follows $N(\mu_j, 1)$ independently. To test whether μ_j is 0, the p-value is given by $p_j = 2\Phi(-|\sqrt{n}\bar{X}_j|)$, where $\bar{X}_j = \sum_{i=1}^{n} X_{ij}/n$, and Φ is the CDF of the standard Normal distribution.

For DS, we construct the mirror statistic M_j using $\bar{X}_j^{(1)}$ and $\bar{X}_j^{(2)}$, and select $\widehat{S}$, that is, reject the null hypotheses that μ_j's are 0, following Algorithm 1. The proposition below holds for any designated FDR control level $q \in (0, 1)$ and for all three choices of mirror statistics detailed in (2). For simplicity, we only prove the case

$$M_j = |\bar{X}_j^{(1)} + \bar{X}_j^{(2)}| - |\bar{X}_j^{(1)} - \bar{X}_j^{(2)}|. \tag{6}$$

Proposition 2.4. For any pair (i, j) and two arbitrary constants $0 < c < c'$, as $n \to \infty$, we have

$$\mathbb{P}\left(M_i < M_j | c \le \sqrt{n}(|\bar{X}_i| - |\bar{X}_j|) \le c'\right) \ge \gamma \text{ and}$$
$$\mathbb{P}(I_i < I_j | c \le \sqrt{n}(|\bar{X}_i| - |\bar{X}_j|) \le c') = o_p(1),$$

in which I_j is defined in (5) and $\gamma > 0$ is a constant depending on c and c'.

Remark 2.2. The p-value p_j is a monotonically decreasing function of the sufficient statistic $|\bar{X}_j|$. Proposition 2.4 shows that for any pair μ_i and μ_j that have a fairly close separation between their p-values p_i and p_j, that is, $|\bar{X}_i| - |\bar{X}_j| = O_p(1/\sqrt{n})$, DS ranks μ_i and μ_j differently from their p-values with a nonvanishing probability, whereas MDS ranks them consistently with their p-values with probability approaching one. Although I_j is not analytically available, MDS approximates it by $\widehat{I}_j$ as in (5). Imagine a perfect knockoff procedure for this Normal means problem, which ranks μ_j's using the knockoff statistic $|\bar{X}_j| - |\bar{X}'_j|$ with $\bar{X}'_j$ being the mean of n independent samples from $N(0, 1)$. Based on the same argument, we can show that the knockoff statistics also rank μ_i and μ_j differently from their p-values with a nonvanishing probability if $|\bar{X}_i| - |\bar{X}_j| = O_p(1/\sqrt{n})$.

To illustrate Proposition 2.4, we fix $p = 800$ and design a small separation between p_1 and p_2 by setting $p_1 = 0.020$, $p_2 = 0.021$. That is, we sample X_{i1}'s conditioning on $\bar{X}_1 = |\Phi^{-1}(0.01)/\sqrt{n}|$ and sample X_{i2}'s conditioning on $\bar{X}_2 = \bar{X}_1 - 0.02/\sqrt{n}$. For $j \ge 3$, we set 20% of μ_j's to be nonzero, and sample them independently from $N(0, 0.5^2)$. We vary the sample size $n \in \{50, 200, 500, 1000, 5000\}$, and estimate the swap probability $\mathbb{P}(M_1 \le M_2)$ for DS and $\mathbb{P}(I_1 \le I_2)$ for MDS over 500 independent runs. For DS, we construct the mirror statistics following (6). For MDS, we set the number of DS replications to be $m = 10n$. The results in Figure 2 empirically validate Proposition 2.4. The left panel shows that for MDS, the swap probability $\mathbb{P}(I_1 \le I_2)$ gets very close to 0 when the sample size is large enough (say, $n \ge 5000$). However, for DS, the swap probability $\mathbb{P}(M_1 \le M_2)$ remains approximately as a constant (slightly below 0.5) as the sample size increases.

Proposition 2.4 implies that for the Normal means model, when the sample size is reasonably large, the estimated inclusion rates $\widehat{I}_j$ and the p-values yield nearly the same rankings of μ_j's with high probability. To illustrate this, we consider a similar simulation setting as above with the sample size $n = 1000$, but without fixing $\bar{X}_1$ or $\bar{X}_2$. In the right panel of Figure 2, we plot both the inclusion rates (blue "$*$" and orange "+") of MDS and the mirror statistics of DS (gray "$\cdot$") against the p-values. For MDS, the blue "$*$" and the orange "+" refer to the estimated inclusion rates based upon $m = 10{,}000$ and $m = 400$ DS replications, respectively. We see that the feature rankings given by the inclusion rates are significantly less noisy compared to that given by the mirror statistics, and the inclusion rate is approximately a monotonically decreasing function of the p-value. Thus, for this simple model, MDS almost recovers the power loss of DS due to sample splitting since the p-values, which are calculated using the full data, summarize all the information relevant to the testing task. Figures 1 and 2 in supplementary materials provide more empirical comparisons between DS, MDS, and BHq across various signal strengths.

3. Specializations for Different Statistical Models

In this section, we discuss how to conduct DS and MDS for linear and Gaussian graphical models, and identify verifiable conditions for each class of models under which both the symmetry assumption and the weak dependence assumption (i.e., Assumptions 2.1 and 2.2) are satisfied. Throughout this section, we split the data into two parts of equal size.

3.1. Linear Models

Suppose the data is generated from the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$. For simplicity, we focus on the random-design scenario, in which each row of the design matrix $\boldsymbol{X}$ follows a p-dimensional distribution with a covariance matrix Σ independently. Remark 3.1 comments on the fixed-design case. In the context of feature selection, the goal is to identify the set of the relevant features S_1, that is, $S_1 = \{j : \beta_j^\star \ne 0\}$ (see Proposition 2.2 in Candès et al. 2018).

We consider the following Lasso+OLS procedure: (a) apply Lasso to the first half of the data $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$ to get the coefficient estimate $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{S}^{(1)} = \{j : \widehat{\beta}_j^{(1)} \ne 0\}$; (b) restricted to the set of features in $\widehat{S}^{(1)}$, apply OLS to get the estimate $\widehat{\boldsymbol{\beta}}^{(2)}$ using the second half of the data $(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)})$. We then construct the mirror statistics by (1) using $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$.

If the *sure screening* property holds for Lasso, that is, all the relevant features are selected in step (ai), then for any selected null feature $j \in S_0 \cap \widehat{S}^{(1)}$, its OLS estimate $\widehat{\beta}_j^{(2)}$ follows a Normal distribution with mean zero conditioning on $\boldsymbol{X}^{(2)}$. Thus, the symmetry assumption is satisfied. Sufficient conditions for the *sure screening* property of Lasso have been well established in

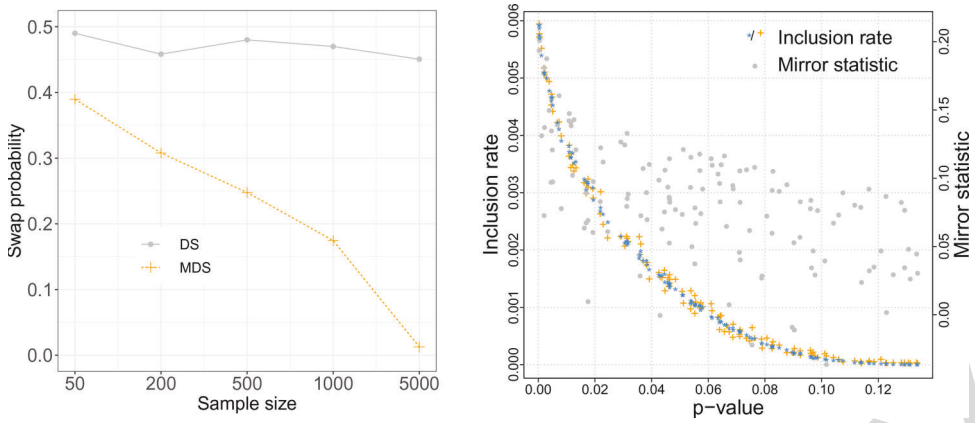

Figure 2. Comparison of DS and MDS on the Normal means model. Given the two p-values p_1, p_2 with $p_1 < p_2$, the left panel plots the estimated swap probability, that is, $\mathbb{P}(M_1 < M_2)$ for DS and $\mathbb{P}(I_1 < I_2)$ for MDS, against the sample size n. The right panel plots the inclusion rates of MDS and the mirror statistics of DS against the p-values. The detailed simulation settings can be found in Section 2.3.

the literature (see Remark 3.1). More generally, we may replace Lasso by any other dimension reduction method as long as the *sure screening* property holds with high probability. Further, using Mehler's identity (Kotz, Balakrishnan, and Johnson 2000), we show that the weak dependence assumption holds with high probability under the regularity condition and the tail condition in Assumption 3.1.

Assumption 3.1.

1. (Signal strength condition) $\min_{j\in S_1} |\beta_j^\star| \gg \sqrt{p_1 \log p/n}$.
2. (Regularity condition) $1/c < \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) < c$ for some $c > 0$.
3. (Tail condition) $\boldsymbol{X}\Sigma^{-1/2}$ has independent sub-Gaussian rows.
4. (Sparsity condition) $p_1 = o(n/\log p)$ and $p_1 \to \infty$.

Proposition 3.1. Consider both DS and MDS, of which the two coefficient estimates $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ are constructed using the Lasso+OLS procedure. For any nominal FDR level $q \in (0, 1)$, under Assumption 3.1, we have

$$\limsup_{n,p\to\infty} \mathrm{FDR} \le q \quad \text{and} \quad \liminf_{n,p\to\infty} \mathrm{Power} = 1$$

in the asymptotic regime where $\log p = o(n^\xi)$ for some $\xi \in (0, 1)$.

This result can be viewed as a consequence of Proposition 2.2 by conditioning on $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$ and restricting ourselves to $\widehat{S}^{(1)}$ as the set of candidate features for consideration in DS. Under Assumption 3.1, we can show that the number of the null features in $\widehat{S}^{(1)}$ is of the same order as p_1 and $\mathrm{var}(M_j)$ is doubly bounded for $j \in \widehat{S}^{(1)}$. Furthermore, the selection cutoff τ_q may diverge and we no longer assume the existence of $t_q > 0$ as in Proposition 2.2.

Remark 3.1. The sure screening property is implied by the signal strength condition and the compatibility condition (Van de Geer and Bühlmann 2009). It is also a crucial condition for high-dimensional knockoff filters (Barber and Candès 2019; Fan et al. 2020) to achieve FDR control and have similar power guarantee. The compatibility condition means that the sample covariance matrix $\widehat{\Sigma}$ of features satisfies $\phi(\widehat{\Sigma}, p_1) \ge \phi_0$ for some $\phi_0 > 0$, in which $\phi(\widehat{\Sigma}, s_0)$ is defined for any integer $s_0 \ge 1$ as

$$\phi^2(\widehat{\Sigma}, s_0) = \min_{|S|\le s_0} \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\boldsymbol{\theta}^\top \widehat{\Sigma}\boldsymbol{\theta}}{||\boldsymbol{\theta}_S||_2^2} : \boldsymbol{\theta} \in \mathbb{R}^p,\ ||\boldsymbol{\theta}_{S^c}||_1 \le 3||\boldsymbol{\theta}_S||_1 \right\}.$$

Unlike the model-X knockoff filter, the randomness of the design matrix is nonessential for DS and MDS, and FDR control of DS in the fixed-design setting can be established similarly as in Xing, Zhao, and Liu (2021). For the random-design case, by Theorem 2.4 in Javanmard and Montanari (2014), if the regularity condition and the tail condition in Assumption 3.1 hold, the compatibility condition holds with high probability for $n \ge cp_1 \log(p/p_1)$. With a properly chosen regularization parameter, the Lasso coefficient estimate $\widehat{\boldsymbol{\beta}}$ satisfies

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star||_2 = O_p(\sqrt{p_1 \log p/n}).$$

Combined with the signal strength condition, we see that the *sure screening* property holds with probability approaching one.

Besides Proposition 3.1, more detailed power analyses of DS and MDS are still unknown. In contrast, some theoretical studies on the power of the knockoff filter have appeared. For example, Fan et al. (2020) showed that, under a similar signal strength condition, that is, $\min_{j\in S_1} |\beta_j^\star| \gg \sqrt{\log p/n}$, and when features follow a multivariate Normal distribution with known covariance matrix, the model-X knockoff filter has asymptotic power one. When the feature covariance matrix is unknown, they proposed a modified knockoff procedure based on data splitting and show that the procedure also has asymptotic power one if the *sure screening* property holds. In a different asymptotic regime where both n/p and p_1/p converge to some fixed constants, the power analysis has been carried out in the setting with iid Gaussian features (e.g., see Weinstein, Barber, and Candès 2017 for the "counting"-knockoffs, and see Weinstein et al. (2020) and Wang and Janson (2020) for the model-X knockoff filter and the conditional randomization test). For correlated designs, Liu and Rigollet (2019) provided some explicit conditions under which the knockoff filter enjoys FDR zero and power one asymptotically. Under the weak-and-rare signal setting, Ke, Liu, and Ma (2020) analyzed both the knockoff filter and the

Gaussian mirror method for some special covariance structures, identifying key components that influence the power of these methods.

Compared to BHq, a main advantage of DS and MDS is that they do not require p-values, which are difficult to obtain for high-dimensional problems. Notable methods for constructing valid p-values include the post-selection inference and the debiased Lasso procedure. Conditioning on the selected model, the post-selection inference derives the exact sampling distribution of the coefficient estimates. Details have been worked out for several popular selection methods such as Lasso (Lee et al. 2016), the forward stepwise regression, and the least angle regression (Tibshirani et al. 2016). However, this type of theory is mostly developed case by case, and cannot be easily generalized to other selection methods. The debiased Lasso procedure removes the biases in the Lasso coefficient estimates so that they enjoy asymptotic Normality under certain conditions (Van de Geer et al. 2014; Zhang and Zhang 2014; Javanmard and Montanari 2014). In particular, Javanmard and Javadi (2019) applies BHq to the p-values obtained via the debiased Lasso procedure for controlling the FDR. However, BHq may still perform poorly using the p-values obtained via these methods. For the post-selection inference, the transformation that converts the coefficient estimates to the p-values may seriously weaken the signal strength. For the debiased-Lasso procedure, the asymptotic null p-values may still be highly nonuniform in finite-sample cases (Dezeure et al. 2015; Candès et al. 2018). To avoid using p-values directly, several authors suggested selecting a proper penalty in penalized regressions based upon the p-value cutoff in order to achieve FDR control. We refer the readers to Benjamini and Gavrilov (2009) and Bogdan et al. (2015) for more details.

We conclude this section by briefly commenting on how to use DS and MDS in the low-dimensional setting with $n/p \to \infty$. For $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$, on a case-by-case basis, we can choose any sensible method (e.g., OLS, Lasso, ridge, or other regularization methods) to obtain the coefficients $\widehat{\boldsymbol{\beta}}^{(1)}$. For $(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)})$, we run OLS using all features to obtain the coefficients $\widehat{\boldsymbol{\beta}}^{(2)}$. The symmetry assumption is automatically satisfied since the model in the OLS step is well specified. The weak dependence assumption still holds under the regularity condition and the tail condition in Assumption 3.1. Therefore, similar to Proposition 3.1, we can show that DS asymptotically controls the FDR without requiring the signal strength and the sparsity conditions. Note that in the low-dimensional setting, the Lasso+OLS procedure is also applicable as long as Assumption 3.1 is satisfied, and may still be a favorable option if both n, p are large and the relevant features are sparse. In particular, in the asymptotic regime where $p/n \to c \in (0, 1/2)$, it can be problematic to directly run OLS on $(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)})$ with all features since the resulting estimate $\widehat{\boldsymbol{\beta}}^{(2)}$ may be unstable and/or its covariance matrix may be ill-conditioned.

3.2. Gaussian Graphical Models

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a p-dimensional multivariate Normal distribution $N(\boldsymbol{\mu}, \Sigma)$. Denote $\Lambda = \Sigma^{-1} = (\lambda_{ij})$ as the precision matrix. Without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$. One can define a corresponding Gaussian graphical model (V, E), in which the set of vertices is $V = (X_1, \ldots, X_p)$, and there is an edge between two different vertices X_i and X_j if X_i and X_j are conditionally dependent given $\{X_k,\ k \neq i, j\}$. The graph estimation can be recast as a nodewise regression problem. To see this, for each vertex X_j, we can write

$$X_j = \boldsymbol{X}_{-j}^{\mathsf{T}} \boldsymbol{\beta}^j + \epsilon_j \quad \text{with} \quad \boldsymbol{\beta}^j = -\lambda_{jj}^{-1} \Lambda_{-j,j},$$

where ϵ_j, independent of $\boldsymbol{X}_{-j}$, follows a Normal distribution with mean zero. Thus, $\lambda_{ij} = 0$ implies that X_i and X_j are conditionally independent. Denote the neighborhood of vertex X_j as $ne_j = \{k :\ k \neq j,\ \beta_k^j \neq 0\}$. Given iid samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $N(\boldsymbol{\mu}, \Sigma)$, it is natural to consider first recovering the support of each $\boldsymbol{\beta}^j$ using a feature selection method such as Lasso (Meinshausen and Bühlmann 2006), and then combining all the nodewise selection results properly to estimate the graph. In view of this, for a nominal FDR level $q \in (0, 1)$, we propose an FDR control procedure as summarized in Algorithm 3.

Algorithm 3 False discovery rate control for Gaussian graphical models via a single data split.

1. Targeting at the level $q/2$, apply the Lasso+OLS procedure (see Section 3.1) to each nodewise regression. Let the nodewise selection results be $\widehat{ne}_j = \{k :\ k \neq j,\ \widehat{\beta}_k^j \neq 0\}$ for $j \in \{1, \ldots, p\}$.
2. Combine the nodewise selection results using the OR rule to estimate the graph:

$$\widehat{E}_{\mathrm{OR}} = \{(i, j) :\ i \in \widehat{ne}_j \text{ or } j \in \widehat{ne}_i\}.$$

A heuristic justification of the proposed method is given below:

$$\begin{aligned} \mathrm{FDP} &= \frac{\#\{(i,j) \in \widehat{E}_{\mathrm{OR}}, (i,j) \notin E\}}{|\widehat{E}_{\mathrm{OR}}| \vee 1} \leq \frac{\sum_{j=1}^p \#\{i \notin ne_j, i \in \widehat{ne}_j\}}{\frac{1}{2}\sum_{j=1}^p \#\{i \in \widehat{ne}_j\} \vee 1} \\ &= \frac{\sum_{j=1}^p \#\{i \notin ne_j, M_{ji} > \tau_{q/2}^j\}}{\frac{1}{2}\sum_{j=1}^p \#\{M_{ji} > \tau_{q/2}^j\} \vee 1} \\ &\approx \frac{\sum_{j=1}^p \#\{i \notin ne_j, M_{ji} < -\tau_{q/2}^j\}}{\frac{1}{2}\sum_{j=1}^p \#\{M_{ji} > \tau_{q/2}^j\} \vee 1} \\ &\leq 2 \max_{1 \leq j \leq p} \frac{\#\{i \notin ne_j, M_{ji} < -\tau_{q/2}^j\}}{\#\{M_{ji} > \tau_{q/2}^j\} \vee 1} \leq q. \end{aligned} \tag{7}$$

For the jth nodewise regression, M_{ji} is the mirror statistic of X_i, $i \neq j$, and $\tau_{q/2}^j$ is the selection cutoff of the mirror statistics. The first inequality in (7) is based on the fact that each edge can be selected at most twice. The approximation in the middle uses the symmetry property of the mirror statistics. The second to last inequality follows from the elementary inequality that $\left(\sum_n a_n\right) / \left(\sum_n b_n\right) \leq \max_n a_n/b_n$ for $a_n \geq 0,\ b_n > 0$.

There are two possible strategies to implement MDS for Gaussian graphical models: (a) apply MDS in each nodewise regression (Step 1 in Algorithm 3) and then aggregate the selection results using the OR rule; (b) replicate the whole procedure in Algorithm 3 (both Steps 1 and 2) multiple times and then aggregate the selection results using MDS. Empirically we found that both strategies achieve FDR control, and the first one tends to have a higher power. Throughout, we focus on the first strategy for MDS.

Let $s = \max_{j\in\{1,\ldots,p\}} |ne_j|$. To theoretically justify our methods, we first show that with probability approaching one, the symmetry assumption is simultaneously satisfied in all nodewise regressions under the following assumptions.

Assumption 3.2.

1. (Regularity condition) $c \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le 1/c$ for some $c > 0$.
2. (Sparsity condition) $s = o(n/\log p)$.
3. (Signal strength condition) $\min\{|\lambda_{ij}| : \lambda_{ij} \neq 0\} \gg \sqrt{s\log p/n}$.

Assumption 3.2 serves the same purpose as Assumption 3.1 for linear models (e.g., ensure that the *sure screening* property holds simultaneously in all nodewise regressions; see Remark 3.1). Similar assumptions also appear in Liu (2013) and Meinshausen and Bühlmann (2006). Under Assumption 3.2, we have the following proposition.

Proposition 3.2. Under Assumption 3.2, as $n, p \to \infty$ satisfying $\log p = o(n)$, the symmetry assumption (Assumption 2.1) is simultaneously satisfied in all nodewise regressions with probability approaching one.

Similar to linear models, the weak dependence assumption is implied by the regularity condition in Assumption 3.2. The following proposition shows the asymptotic FDR control of DS and MDS.

Proposition 3.3. Assume that Assumption 3.2 holds and that $\min_{j\in\{1,\ldots,p\}} |ne_j|/\log p \to \infty$. For any nominal FDR level $q \in (0, 1)$, both DS (see Algorithm 3) and MDS achieve

$$\limsup_{n,p\to\infty} \mathrm{FDR} \le q \quad \text{and} \quad \liminf_{n,p\to\infty} \mathrm{Power} = 1$$

in the asymptotic regime where $\log p = o(n^{\xi})$ for some $\xi \in (0, 1)$.

Assumption $\min_{j\in\{1,\ldots,p\}} |ne_j|/\log p \to \infty$ is a technical one to ensure that a union bound can be applied for all nodewise regressions. Empirically, we find that the data-splitting methods and the GFC method proposed in Liu (2013) are effective in quite different scenarios. GFC tends to work well if the underlying true graph is ultra-sparse, that is, the nodewise sparsity is in the order of $o(\sqrt{n}/(\log p)^{3/2})$. In contrast, DS and MDS perform superbly when the graph is not too sparse. A similar issue also exists in the knockoff-based methods, and we refer the readers to Li and Maathuis (2019) for relevant discussions.

4. Numerical Illustrations

4.1. Linear Model

We simulate the response vector $\boldsymbol{y}$ from the linear model $\boldsymbol{y} = \boldsymbol{X\beta^\star} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, I_n)$, and randomly locate the signal index set S_1. For $j \in S_1$, we sample $\beta_j^\star$ from $N(0, \delta\sqrt{\log p/n})$, and refer to δ as the signal strength. Throughout, the designated FDR control level is set to be $q = 0.1$ and each dot in the figure represents the average from 200 independent runs. The penalization parameter of Lasso is selected based on 10-fold cross-validation.

We first investigate the performance of DS and MDS using different mirror statistics constructed with f_1, f_2, f_3 specified in (2). Each row of the design matrix is independently drawn from $N(0, \Sigma)$. We consider a similar setup as in Ma, Cai, and Li (2020), where Σ is a blockwise diagonal matrix of 10 Toeplitz submatrices whose off-diagonal entries linearly descend from ρ to 0. The detailed formula of Σ is given in (18) in supplementary materials, and we refer to it as the Toeplitz covariance matrix throughout. We vary the correlation ρ and the signal strength δ, and the results are summarized in Figure 3. All three choices of mirror statistics achieve FDR control, and f_3 yields the highest power. Proposition 2.1 shows that f_3 is optimal for orthogonal designs, and the empirical results suggest that f_3 might also be an optimal choice in more realistic settings. For comparisons under other design matrices, please see Figure 6 in supplementary materials. Among all the simulation studies described below, we construct the mirror statistic with f_3. Note that the performance of MDS appears to be more robust to the choice of mirror statistic than that of DS (see Figures 5 and 7 in supplementary materials).

We then examine the impact of the number of DS replications m on the power of MDS. With $n = 500$, $p = 500$, and $p_1 = 50$, we generate features as in the previous simulation. We set the signal strength $\delta = 3$ and test out two scenarios with the correlation $\rho = 0.0$ and $\rho = 0.8$. Figure 4 shows that the power of MDS monotonically increases with the number of DS replications m, and becomes relatively stable after $m \ge 50$ (also see Figure 8 in supplementary materials for results under other design matrices). This suggests that only a small number of DS replications are required to realize the full potential of MDS. Thus, MDS is computationally more feasible for large datasets compared to other methods such as the knockoff filter and the Gaussian mirror method. In the following examples, we set $m = 50$ for MDS.

We proceed to compare DS and MDS with two popular methods in high-dimensional regressions under various design matrices: MBHq (Meinshausen, Meier, and Bühlmann 2009) and the model-X knockoff filter (Candès et al. 2018). For their comparisons in low-dimensional settings, we refer the readers to Figures 9 and 10 in supplementary materials. For MBHq, we obtain 50 p-values for each feature via repeated sample splitting. More precisely, we run Lasso for feature screening on one half of the data, and calculate the p-values for the selected features by running OLS on the other half of the data. We then combine the p-values across different sample splits using the R package *hdi*.[1] For the knockoff filter, we use the equi-correlated knockoffs, in which the covariance matrix of features is estimated using the R package *knockoff*.[2] For all the simulation settings in Section 4.1, we empirically found that the equi-correlated knockoffs yields a higher power compared to the default *asdp* construction.

1. *Normal design matrices.* With $n = 800$, $p \in \{1000, 2000\}$ and $p_1 = 50$, we generate features independently from $N(0, \Sigma)$ with Σ being a Toeplitz covariance matrix. We compare

[1] https://cran.r-project.org/web/packages/hdi/hdi.pdf
[2] https://cran.r-project.org/web/packages/knockoff/index.html

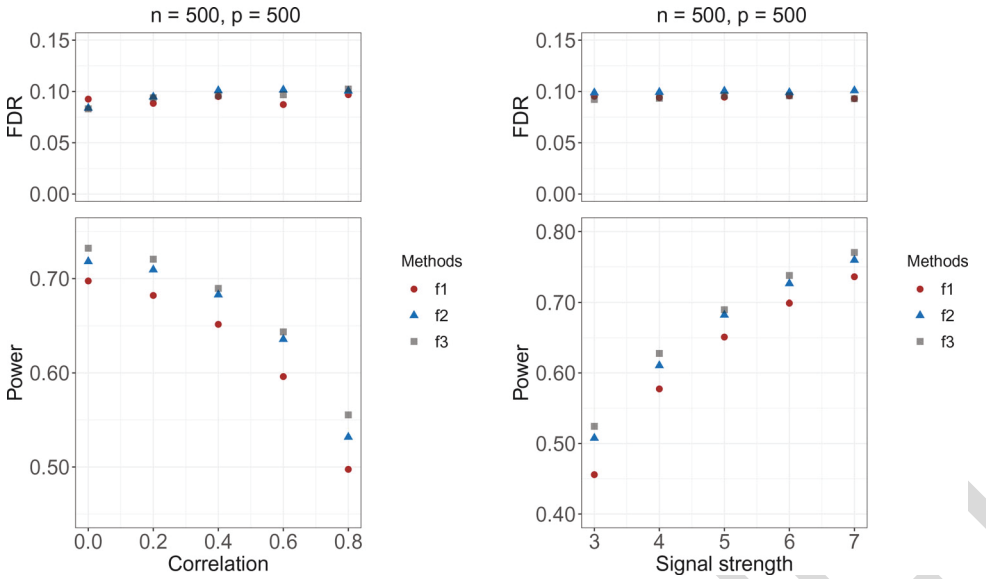

Figure 3. Empirical FDRs and powers of DS under three constructions of the mirror statistics as specified by f_1, f_2, and f_3 in (2). The number of the relevant features is $p_1 = 50$. Left panel: Signal strength $\delta = 5$. Right panel: Correlation $\rho = 0.4$.

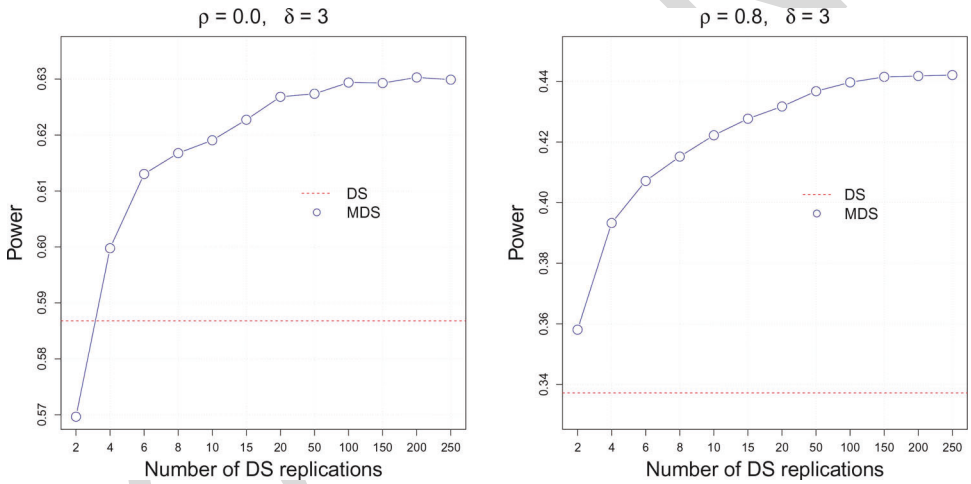

Figure 4. Empirical powers of MDS with different number of DS replications. The sample size is $n = 500$, the number of features is $p = 500$, and the number of the relevant features is $p_1 = 50$. The blue dots and the red lines represent the average powers of MDS and DS over 200 independent runs, respectively.

the performances of the competing methods under different correlations ρ and signal strengths δ. The results for $p = 2000$ are summarized in Figure 5, and the results for $p = 1000$ are summarized in Figure 3 in supplementary materials. The FDRs of all the four methods are under control across different settings. In terms of power, the knockoff filter and MDS are the two leading methods. MDS appears more powerful when features are more correlated, or when the signal strength is relatively weak, whereas the knockoff filter enjoys a higher power in the opposite regimes. We observed that MDS is more robust to highly correlated design matrices compared to the knockoff filter. Figure 4 in Supplementary Materials report the performances of these methods in cases where Σ has a constant pairwise correlation ρ, and the knockoff filter is significantly less powerful than MDS when $\rho \geq 0.4$. The simulation results also suggest that MDS yields better rankings of features compared to DS, thus, enjoys simultaneously a lower FDR and a higher power.

2. *Nonnormal design matrices.* When the joint distribution of the features is unknown and nonnormal, the performance of the knockoff filter is not guaranteed if the knockoffs are generated based on a naive fit of the multivariate Normal distribution to the features. We here illustrate the robustness of DS and MDS with respect to nonnormality by considering the following two design matrices: (a) a mixture of two Gaussians centered at $0.5 \times \mathbb{1}_p$ and $-0.5 \times \mathbb{1}_p$, respectively; (b) a centered multivariate t-distribution with 3 degrees of freedom. Throughout, the covariance matrix Σ is set to be a Toeplitz matrix. Note that in both scenarios, the marginal distribution of each feature is still unimodal, and does not differ much from the Normal distribution in appearance. We fix $n = 800$, $p = 2000$, $p_1 = 70$, and test out different correlations ρ and signal strengths δ. The results are summarized in Figure 6. Because of the model misspecification in the knockoff construction, the knockoff filter appears over conservative when features follow a Gaussian mixture distribution, and loses FDR control when features follow a t-

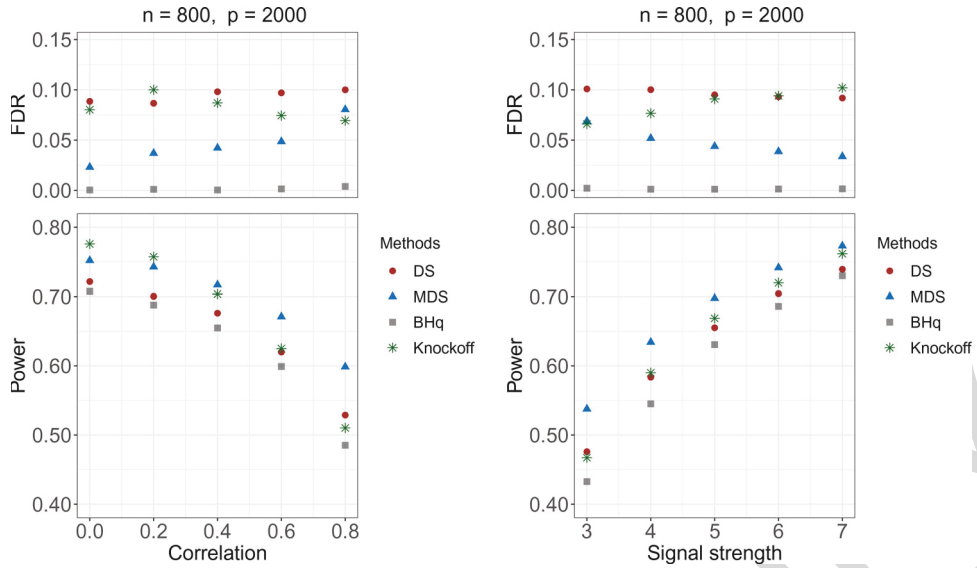

Figure 5. Empirical FDRs and powers for linear models with Normal design matrices. The number of relevant features is $p_1 = 50$. Left panel: Signal strength $\delta = 5$. Right panel: Correlation $\rho = 0.5$.

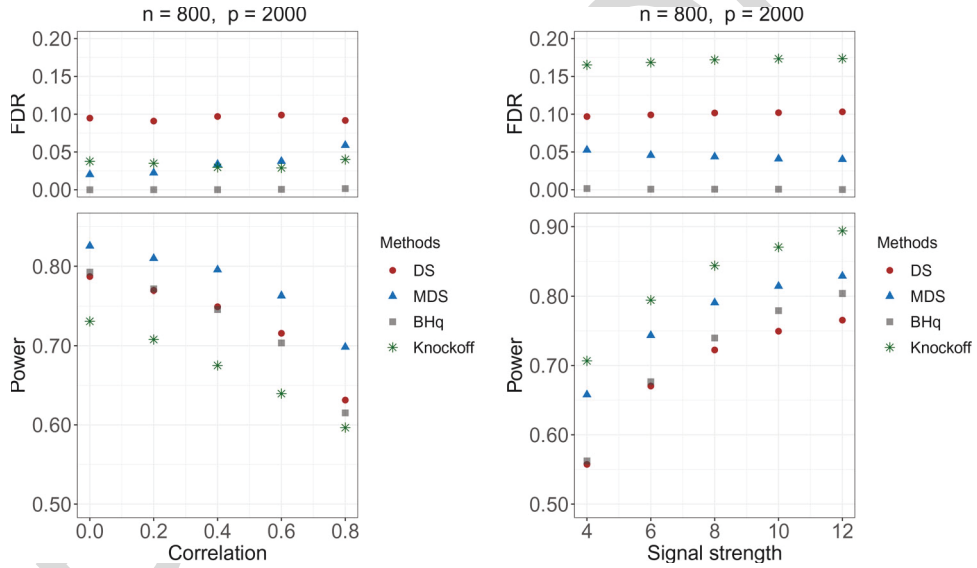

Figure 6. Empirical FDRs and powers for linear models with nonnormal design matrices. The number of relevant features is $p_1 = 70$. Left panel: A two-component Gaussian mixture design with signal strength $\delta = 8$. Right panel: A multivariate t-distribution design with correlation $\rho = 0.5$.

distribution. The latter is perhaps a more concerning issue in the context of controlled feature selection, although the performance of the knockoff filter can be potentially improved by carefully modeling the joint distribution of features based on some structural assumptions (e.g., see Sesia, Sabatti, and Candès 2018). In comparison, MDS maintains FDR control and enjoys a reasonably high power in both scenarios. We also note that, except being overly conservative, MBHq performs quite competitively in all settings.

3. *Real-data design matrices.* We consider using the scRNAseq data in Hoffman et al. (2020) as the design matrix. A total of 400 T47D A1–2 human breast cancer cells were treated with 100 nM synthetic glucocorticoid dexamethasone (Dex). A scRNASeq experiment was performed after 18 hr of the Dex treatment, leading to a total of 400 samples of gene expressions for the treatment group. For the control group, there are 400 vehicle-treated control cells. A scRNAseq experiment was performed at the 18 hr timepoint to obtain the corresponding profile of gene expressions. After proper normalization, the final scRNAseq dataset[3] contains 800 samples, each with 32,049 gene expressions. To further reduce the dimensionality, we first screen out the genes detected in fewer than 10% of cells, and then pick up the top p most variable genes following Hoffman et al. (2020). We fix $p_1 = 70$, and simulate the response vector $\boldsymbol{y}$ with various p and signal strengths. The results are summarized in Figure 7. We see that all the methods achieve FDR control, among which MDS

[3]The data is available at *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE141834*.

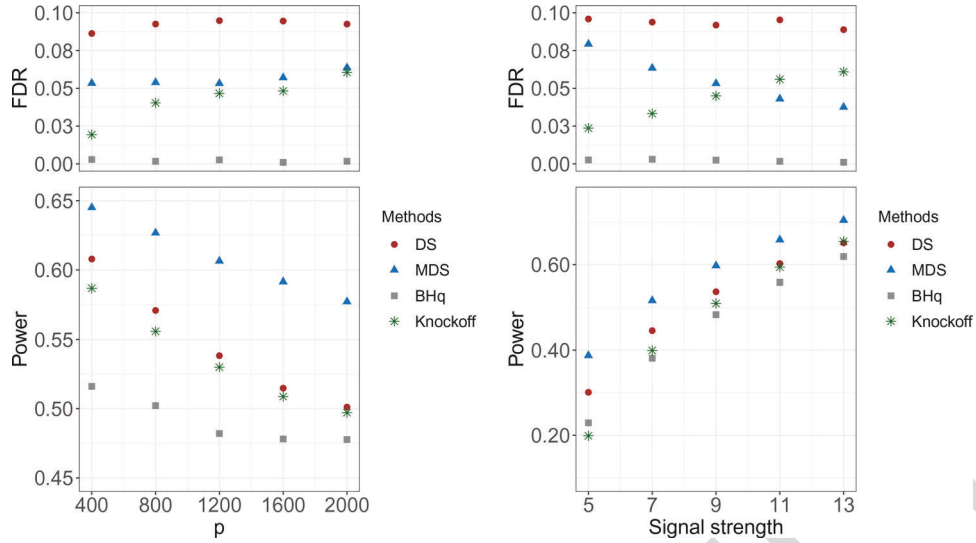

Figure 7. Empirical FDRs and powers for linear models with a design matrix generated based on a scRNAseq dataset. The sample size is $n = 800$, and the number of the relevant features is $p_1 = 70$. The signal strength is scaled by $1/\sqrt{n}$. Left panel: signal strength $\delta = 9$. Right panel: dimension $p = 1200$.

enjoys the highest power. The knockoff filter appears to be conservative (with its FDR significantly below the nominal level 0.1), likely due to the fact that the joint distribution of gene expressions is nonnormal, resulting in a misspecified construction of the knockoffs.

We conclude this section with some remarks on the variance of the FDP. Note that DS and the knockoff filter rank features using the mirror statistics M_j's and the statistics W_j's (see Section 3.2 in Candès et al. 2018), respectively. The statistic W_j enjoys a *flip-sign* property, that is, the signs of W_j's for $j \in S_0$ are independent, thus, the FDP of the knockoff filter fluctuates and concentrates around the FDR. For DS, the signs of the mirror statistics M_j's for $j \in S_0$ are correlated so the variance of the FDP can be a more concerning issue. FDR control becomes less meaningful if the variance is unacceptably large. We empirically check the variances of the FDP for the four competing methods across the aforementioned simulation settings. The results are summarized in Figures 11, 12, 13, and 14 in supplementary materials. We observe that, except for the cases where the knockoff filter appears overly conservative (e.g., Figure 12), the variances of the FDP are comparable for DS and the knockoff filter. More interestingly, perhaps due to its derandomized nature, MDS achieves a lower variance of the FDP than the knockoff filter in a majority of simulation settings.

4.2. Gaussian Graphical Model

We set the designated FDR control level at $q = 0.2$ throughout and each dot in the figure represents the average from 200 independent runs. We consider two types of graphs:

1. *Banded graph.* Precision matrix Λ satisfies $\lambda_{jj} = 1$, $\lambda_{ij} = \text{sign}(a) \cdot |a|^{|i-j|/c}$ if $0 < |i-j| \le s$, and $\lambda_{ij} = 0$ if $|i-j| > s$. Throughout, we set $c = 1.5$ following Li and Maathuis (2019). Other parameters including the sample size n, the dimension p, the partial correlation (signal strength) a, and the nodewise sparsity s will be specified case by case.
2. *Blockwise diagonal graph.* The precision matrix Λ is blockwise diagonal with equally sized squared blocks generated in the same fashion. Throughout, we fix the block size to be 25 $\times$ 25. In each block, all the diagonal elements are set to be 1, and the off-diagonal elements are independently drawn from the uniform distribution $\text{Unif}((-0.8, -0.4) \cup (0.4, 0.8))$.

The precision matrix Λ generated from the aforementioned procedures may not be positive definite. If $\lambda_{\min}(\Lambda) < 0$, we reset $\Lambda \leftarrow \Lambda + (\lambda_{\min}(\Lambda) + 0.005)I_p$ following Liu (2013). Three classes of competing methods are tested out, including (a) DS and MDS; (b) BHq; (c) GFC (Liu 2013). For MDS, nodewisely, we replicate DS 50 times and aggregate the selection results using Algorithm 2. For BHq, the p-values are calculated based on the pairwise partial correlation test using the R package *ppcor* (Kim 2015). For GFC, we use the R package *SILGGM* (Zhang, Ren, and Chen 2018) to implement it.

For the banded graph, we test out the following four scenarios:

(a) fix $p = 100$, $s = 8$, $a = -0.6$, and vary the sample size $n \in \{500, 1000, 1500, 2000, 2500\}$;
(b) fix $n = 1000$, $s = 8$, $a = -0.6$, and vary the dimension $p \in \{50, 100, 150, 200, 250\}$;
(c) fix $n = 1000$, $p = 100$, $a = -0.6$, and vary the nodewise sparsity $s \in \{4, 6, 8, 10, 12\}$;
(d) fix $n = 1000$, $p = 100$, $s = 8$, and vary the signal strength $a \in \{-0.5, -0.6, -0.7, -0.8, -0.9\}$.

For the blockwise diagonal graph, we test out the following two scenarios:

(a) fix $p = 100$, and vary the sample size $n \in \{200, 300, 400, 500, 600\}$;
(b) fix $n = 500$, and vary the dimension $p \in \{50, 100, 150, 200, 250\}$.

Results for the banded graphs and the blockwise diagonal graphs are summarized in Figures 8 and 9, respectively. We see that all the methods achieve FDR control at the designated level across different scenarios. For the banded graphs, DS and MDS

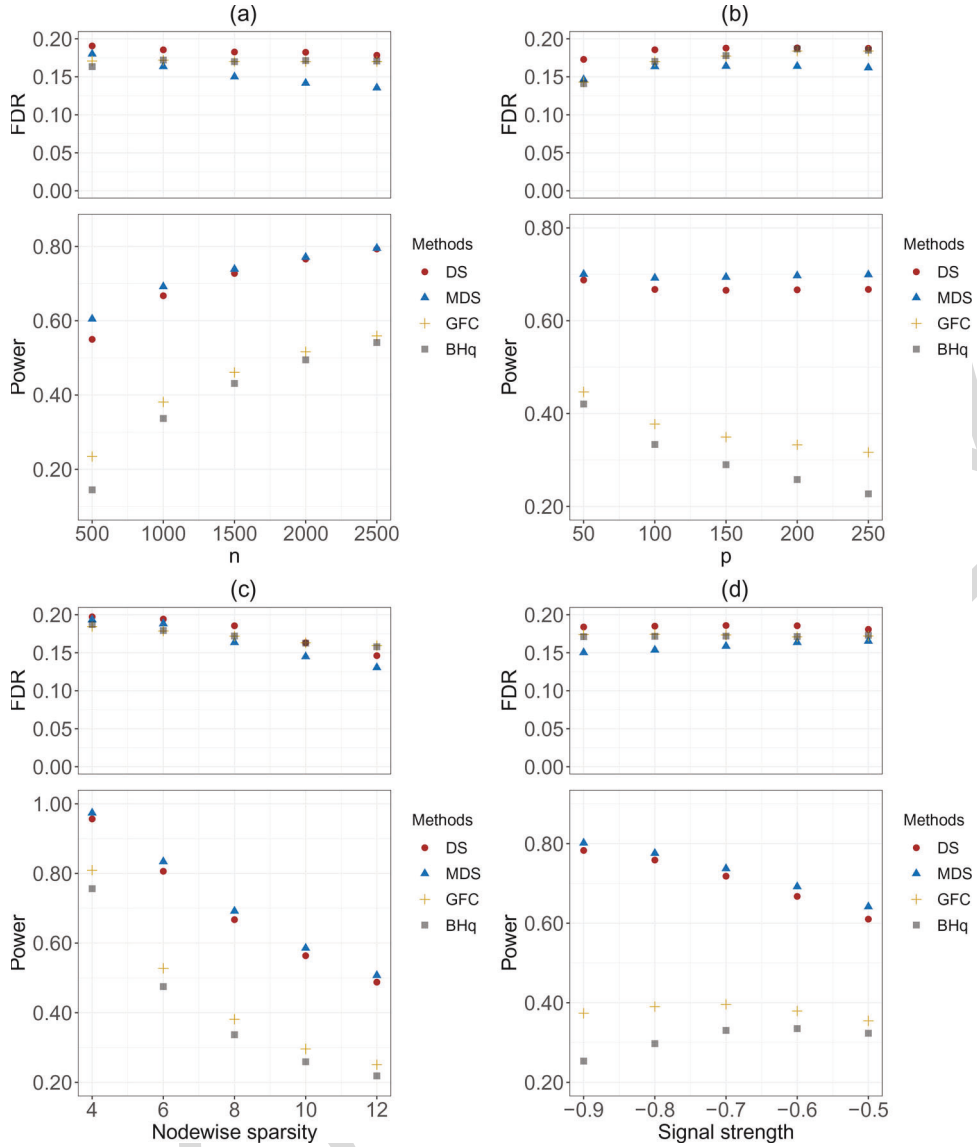

Figure 8. Empirical FDRs and powers for the banded graphs. (a) dimension $p = 100$, sparsity $s = 8$, signal strength $a = -0.6$, and varying sample size n; (b) $n = 1000$, $s = 8$, $a = -0.6$, and varying dimension p; (c) $n = 1000$, $p = 100$, $a = -0.6$, and varying sparsity s; (d) $n = 1000$, $p = 100$, $s = 8$, and varying signal strength a.

are the two leading methods with significantly higher powers and also lower FDRs compared to the other two competing methods. GFC and BHq perform similarly, and GFC has a slightly higher power when p is large or the signal strength is strong. In panel (d) of Figure 8, the power of BHq exhibits an opposite trend compared to the other methods. One possible reason is that the pairwise correlation decreases when we increase a from -0.9 to -0.5. Thus, the power of BHq increases as the p-values become less correlated. For the blockwise diagonal graphs, MDS performs the best across all scenarios, enjoying a higher power and also a lower FDR compared to DS. GFC performs similarly as DS in most scenarios.

4.3. Real Data Application: HIV Drug Resistance

We apply DS and MDS to detect mutations in the Human Immunodeficiency Virus Type 1 (HIV-1) that are associated with drug resistance. The dataset, which has also been analyzed in Rhee et al. (2006), Barber and Candès (2015), and Lu et al. (2018), contains resistance measurements of seven drugs for protease inhibitors (PIs), six drugs for nucleoside reverse-transcriptase inhibitors (NRTIs), and three drugs for nonnucleoside reverse transcriptase inhibitors (NNRTIs). We focus on the first two classes of inhibitors, PI and NRTI.

The response vector $\boldsymbol{y}$ records the log-fold-increase of the lab-tested drug resistance. The design matrix $\boldsymbol{X}$ is binary, in which the jth column indicates the presence or absence of the jth mutation. The task is to select relevant mutations for each inhibitor against different drugs. The data is preprocessed as follows. First, we remove the patients with missing drug resistance information. Second, we exclude those mutations that appear fewer than three times across all patients. The sample size n and the number of mutations p vary from drug to drug, but are all in hundreds with n/p ranging from 1.5 to 4 (see Figures 10 and 11). We assume a linear model between the response and features with no interactions.

Five methods are compared, including DeepPINK with the model-X knockoff (Lu et al. 2018), the fixed-design knockoff

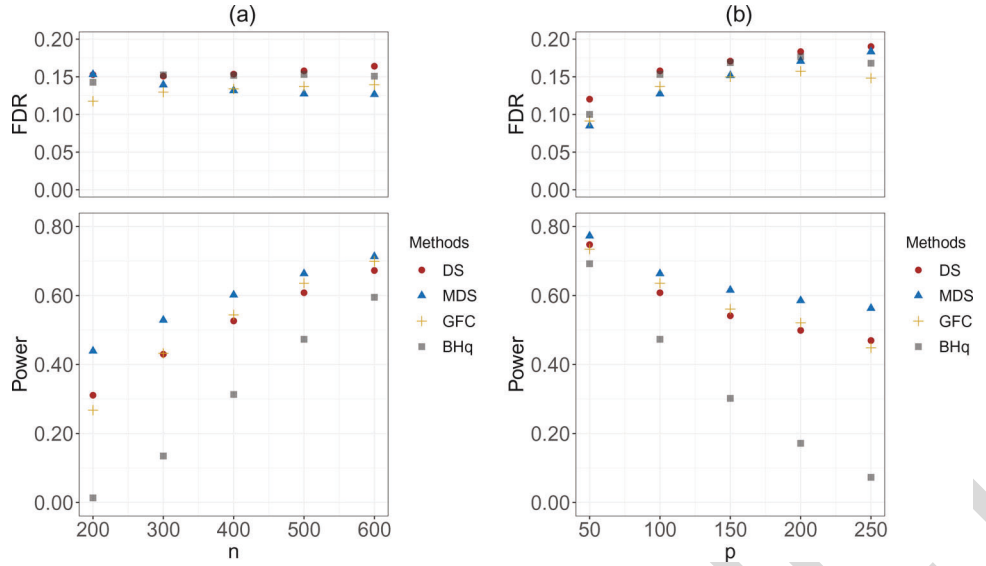

Figure 9. Empirical FDRs and powers for the blockwise diagonal graphs with the block size fixed at 25 × 25. In each block, the diagonal elements are equal to 1, and the off-diagonal elements are independently drawn from Unif$((-0.8,-0.4)\cup(0.4,0.8))$. (a) dimension $p = 100$ and varying sample size n; (b) $n = 500$ and varying p.

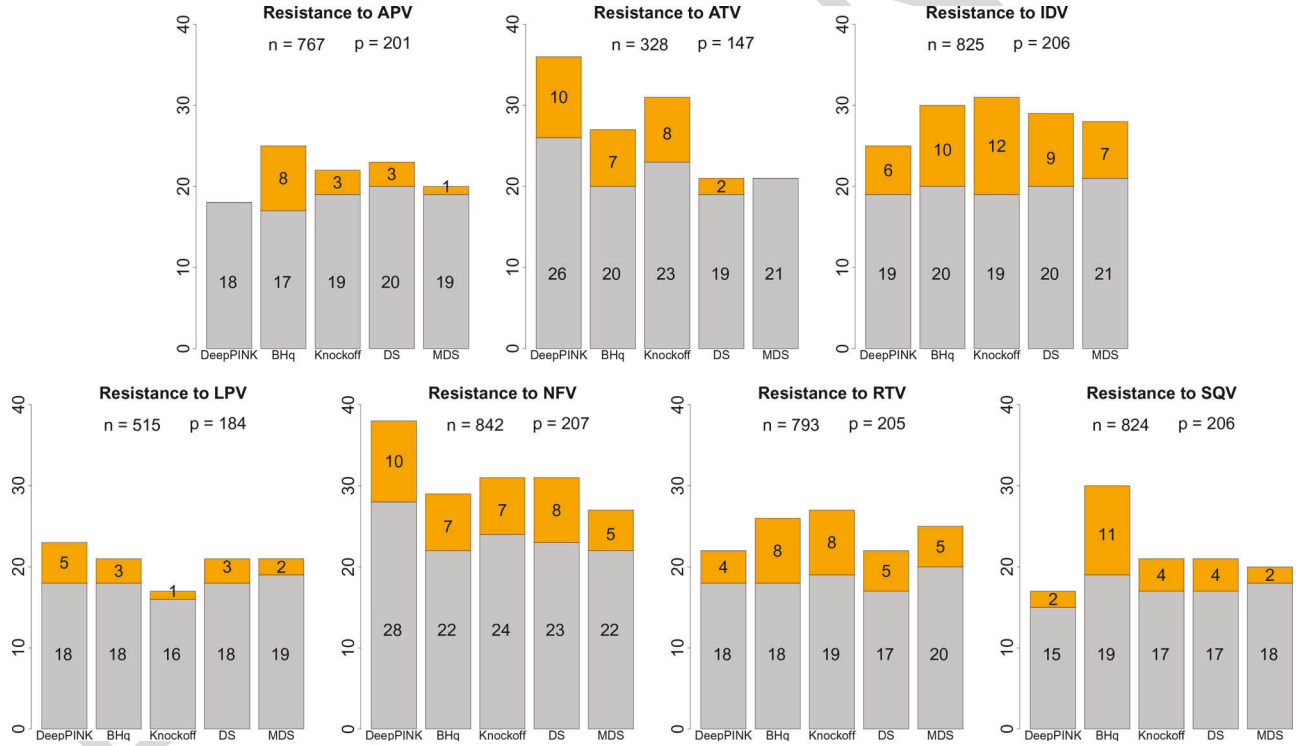

Figure 10. Numbers of the discovered mutations for the seven PI drugs. The gray and orange bars represent the numbers of true and false positives, respectively. The designated FDR control level is $q = 0.2$.

filter (Barber and Candès 2015), BHq, DS, and MDS. For DeepPINK, the knockoff filter, and BHq, we report the selection results obtained in Lu et al. (2018). The designated FDR control level is $q = 0.2$ throughout. As in Barber and Candès (2015), we treat the existing treatment-selected mutation (TSM) panels (Rhee et al. 2005) as the ground truth.

For PI, the number of discovered mutations for each drug, including the number of true and false positives, are summarized in Figure 10. We see that MDS performs the best for three out of seven PI drugs, including ATV, LPV and SQV. For drugs APV, IDV, and RTV, MDS is comparable to DeepPINK, and both perform better than the knockoff filter and BHq. For drug NFV, the knockoff filter and MDS are the two leading methods. Figure 11 shows the corresponding results for the NRTI drugs. Among the six NRTI drugs, MDS performs the best in four, including ABC, D4T, DDI, and X3TC. For drug AZT, the knockoff filter and MDS both perform the best. For drug TDF, MDS is comparable to DeepPINK, and both are much better than BHq and the knockoff filter. In particular, we see that the knockoff filter has no power and does not select any mutation for drugs DDI, TDF, and X3TC.

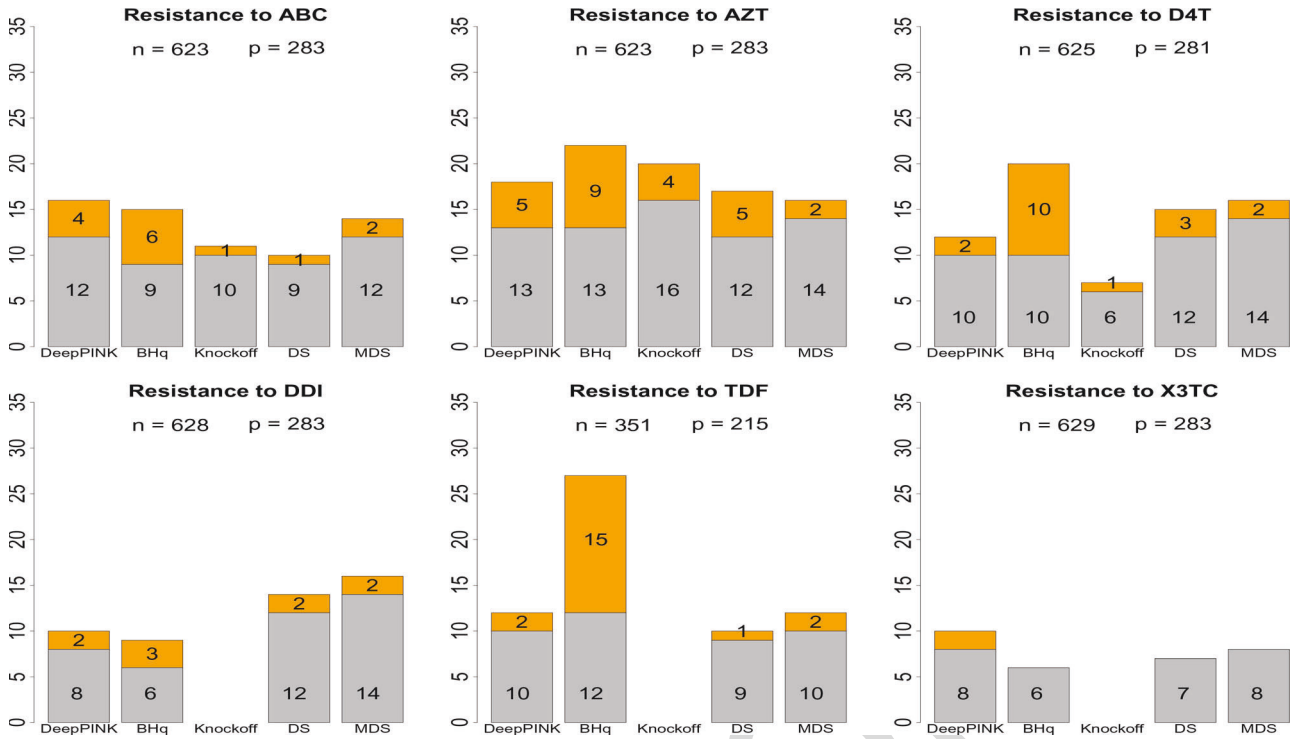

Figure 11. Numbers of the discovered mutations for the six NRTI drugs. The gray and orange bars represent the numbers of true and false positives, respectively. The designated FDR control level is $q = 0.2$.

5. Concluding Remarks

We have described a data-splitting framework for FDR control in high-dimensional regression problems. We demonstrate both theoretically and empirically that the proposed approaches (DS and MDS) allow us to asymptotically control the FDR in linear and Gaussian graphical models under mild conditions. MDS is shown to be particularly attractive as it helps stabilize the selection result and improves the power. Both DS and MDS require no prior knowledge on the joint distribution of features, and are conceptually simple and easy to implement.

Several directions for further developments are worthwhile to consider. For linear models, it is useful to extend the Lasso+OLS procedure for features with certain group structure. A natural strategy is to replace Lasso by the group Lasso (Yuan and Lin 2006). However, the group Lasso can potentially select more than n features (n is the sample size). Thus, the companion OLS step, which guarantees the symmetry assumption, may not be directly applied. It is also of interest to investigate the applicability and theoretical properties of DS and MDS for dealing with neural networks and other nonlinear models in view of more complex data such as images and natural languages. Last but not the least, extensions of the FDR control framework to handle data containing dependent observations (such as time series) or having hierarchical structures are of immediate interest.

Supplementary Materials

Q3

Acknowledgments

We thank Lucas Janson, Wenshuo Wang, Dongming Huang, and two referees for many helpful comments and constructive suggestions.

Funding

This research is supported in part by the National Science Foundation grants DMS-1903139, DMS-2015411 and DMS-2124535.

References

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [2,4,14,15]

——— (2019), "A Knockoff Filter for High-Dimensional Selective Inference," *The Annals of Statistics*, 47, 2504–2537. [2,8]

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020), "Robust Inference with Knockoffs," *The Annals of Statistics*, 48, 1409–1431. [2]

Bates, S., Candés, E. J., Janson, L., and Wang, W. (2020), "Metropolized Knockoff Sampling," *Journal of the American Statistical Association*, 116, 1413–1427. [2]

Benjamini, Y., and Gavrilov, Y. (2009), "A Simple Forward Selection Procedure Based on False Discovery Rate Control," *The Annals of Applied Statistics*, 3, 179–198. [9]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [1,2]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid 'Post-selection Inference," *The Annals of Statistics*, 41, 802–837. [2]

Bogdan, M., Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015), "Slope—Adaptive Variable Selection via Convex Optimization," *The Annals of Applied Statistics*, 9, 1103–1150. [9]

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'model-X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [2,3,4,7,9,10,13]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1]

Cox, D. R. (1975), "A Note on Data-Splitting for the Evaluation of Significance Levels," *Biometrika*, 62, 441–444. [2]

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2020), "A Scale-Free Approach for False Discovery Rate Control in Generalized Linear Models," arXiv preprint: 2007.01237. [5]

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), "High-Dimensional Inference: Confidence Intervals, *p*-values and R-software hdi," *Statistical Science*, 30, 533–558. [9]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [4]

Efron, B. (2005), "Local False Discovery Rates," *Technical report*. [2]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160. [2]

Efroymson, M. (1960), "Multiple Regression Analysis," in *Mathematical Methods for Digital Computers*, eds. A. Ralston and H. S. Wilf, pp. 191–203, New York: Wiley. [1]

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, 362–379. [8]

Hoffman, J. A., Papas, B. N., Trotter, K. W., and Archer, T. K. (2020), "Single-Cell RNA Sequencing Reveals a Heterogeneous Response to Glucocorticoids in Breast Cancer Cells," *Communications Biology*, 3, 1–11. [12]

Huang, D., and Janson, L. (2020), "Relaxing the Assumptions of Knockoffs by Conditioning," *The Annals of Statistics*, 48, 3021–3042. [2]

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016), "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing," *Nature Methods*, 13, 577–580. [2]

Javanmard, A., and Javadi, H. (2019), "False Discovery Rate Control via Debiased Lasso," *Electronic Journal of Statistics*, 13, 1212–1253. [9]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [5,8,9]

Jordon, J., Yoon, J., and Schaar, M. V. D. (2019), "KnockoffGAN: Generating Knockoffs for Feature Selection Using Generative Adversarial Networks," in *The International Conference on Learning Representations*. [2]

Katsevich, E., and Sabatti, C. (2019), "Multilayer Knockoff Filter: Controlled Variable Selection at Multiple Resolutions," *The Annals of Applied Statistics*, 13, 1–33. [2]

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: 2010.08132. [4,8]

Kim, S. (2015), "ppcor: an R Package for a Fast Calculation to Semi-partial Correlation Coefficients," *Communications for Statistical Applications and Methods*, 22, 665–674. [13]

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000), "Bivariate and Trivariate Normal Distributions," *Continuous Multivariate Distributions*, 1, 251–348. [8]

Q4 Lauritzen, S. L. (1996), "Graphical Models." [3]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-selection Inference, with Application to the Lasso," *The Annals of Statistics*, 44, 907–927. [2,9]

Li, J., and Maathuis, M. H. (2019), "Nodewise Knockoffs: False Discovery Rate Control for Gaussian Graphical Models," arXiv preprint: 1908.11611. [10,13]

Liu, J., and Rigollet, P. (2019), "Power Analysis of Knockoff Filters for Correlated Designs," in *Advances in Neural Information Processing Systems 32*, 15446–15455. [8]

Liu, W. (2013), "Gaussian Graphical Model Estimation with False Discovery Rate Control," *The Annals of Statistics*, 41, 2948–2978. [10,13]

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014), "A Significance Test for the Lasso," *The Annals of Statistics*, 42, 413–468. [2]

Lu, Y., Fan, Y., Lv, J., and Noble, W. S. (2018), "DeepPINK: Reproducible Feature Selection in Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 8676–8686. [2,14,15]

Ma, R., Cai, T. T., and Li, H. (2020), "Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models," *Journal of the American Statistical Association*, 116, 984–998. [10]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [3,9,10]

——— (2010), "Stability Selection," *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [3,6]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "p-values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [2,6,10]

Moran, P. A. P. (1973), "Dividing a Sample Into Two Parts a Statistical Dilemma," *Sankhyā: The Indian Journal of Statistics*, Series A, 35, 329–333. [2]

O'Hara, R. B., and Sillanpää, M. J. (2009), "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–117. [1]

Ren, Z., Wei, Y., and Candès, E. (2020), "Derandomizing Knockoffs." [3] Q5

Rhee, S. Y., Fessel, W. J., Zolopa, A. R., Hurley, L., Liu, T., Taylor, J., Nguyen, D. P., Slome, S., Klein, D., and Horberg, M. (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, 456–465. [15]

Rhee, S. Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006), "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance," *Proceedings of the National Academy of Sciences*, 103, 17355–17360. [14]

Romano, J. P., and DiCiccio, C. (2019), "Multiple data splitting for testing." [6] Q6

Romano, Y., Sesia, M., and Candès, E. J. (2019), "Deep Knockoffs," *Journal of the American Statistical Association*, 115, 1861–1872. [2]

Rubin, D., Dudoit, S., and der Laan, M. V. (2006), "A Method to Increase the Power of Multiple Testing Procedures Through Sample Splitting," *Statistical Applications in Genetics and Molecular Biology*, 5. [2] Q7

Sarkar, S. K. (2002), "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures," *The Annals of Statistics*, *30*, 239–257. [1]

Sesia, M., Katsevich, E., Bates, S., Candès, E. J., and Sabatti, C. (2020), "Multi-Resolution Localization of Causal Variants Across the Genome," *Nature Communications*, 11. [2] Q8

Sesia, M., Sabatti, C., and Candès, E. J. (2018), "Gene Hunting with Hidden Markov Model Knockoffs," *Biometrika*, 106, 1–18. [2,12]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–133. [2]

Storey, J. D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *The Annals of Statistics*, 31, 2013–2035. [2]

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, 66, 187–205. [1,5]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620. [9]

Van de Geer, S. A., and Bühlmann, P. (2009), "On the Conditions used to Prove Oracle Results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392. [8]

Van de Geer, S. A., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [5,9]

van de Wiel, M. A., Berkhof, J., and van Wieringen, W. N. (2009), "Testing the Prediction Error Difference Between Two Predictors," *Biostatistics*, 10, 550–560. [6]

Wang, W., and Janson, L. (2020), "A Power Analysis of the Conditional Randomization Test and Knockoffs," arXiv preprint: 2010.02304. [8]

Wasserman, L., and Roeder, K. (2009), "High Dimensional Variable Selection," *The Annals of Statistics*, 37, 2178–2201. [2]

Weinstein, A., Barber, R. F., and Candès, E. J. (2017), "A Power and Prediction Analysis for Knockoffs with Lasso Statistics." [8]

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." [8]

Wu, W. B. (2008), "On False Discovery Control Under Dependence," *The Annals of Statistics*, 36, 364–380. [1,5]

Xing, X., Zhao, Z., and Liu, J. S. (2021), "Controlling False Discovery Rate Using Gaussian Mirrors," *Journal of the American Statistical Association*, 1–45. [4,5,8]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [16]

Zhang, C. H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [5,9]

Zhang, R., Ren, Z., and Chen, W. (2018), "SILGGM: An Extensive R Package for Efficient Statistical Inference in Large-Scale Gene Networks," *PloS Computational Biology*, 14, e1006369. [13]