A General Framework for Powerful Confounder Adjustment in Omics Association Studies

Asmita Roy[1], Jun Chen[2], and Xianyang Zhang*[1]

[1]Texas A&M University
[2]Mayo Clinic

Abstract Genomic data are subject to various sources of confounding, such as batch effects and cell mixtures. To identify genomic features associated with a variable of interest in the presence of confounders, the traditional approach involves fitting a confounder-adjusted regression model to each genomic feature, followed by multiplicity correction. Previously, we showed that this procedure was sub-optimal and proposed a more powerful procedure named the two-dimensional false discovery rate control (2dFDR) procedure, which relied on the test statistics from both confounder-adjusted and unadjusted linear regression models (Yi et al., 2021). Though 2dFDR provides significant power improvement over the traditional method, it is based on the linear model assumption that may be too restrictive for some practical settings. This study proposes a model-free two-dimensional false discovery rate control procedure (MF-2dFDR) to significantly broaden the scope and applicability of 2dFDR. MF-2dFDR uses marginal independence test statistics as auxiliary information to filter out less promising features, and FDR control is performed based on conditional independence test statistics in the remaining features. MF-2dFDR provides (asymptotically) valid inference from samples in settings in which the conditional distribution of the genomic variables given the covariate of interest and the confounders is arbitrary and completely unknown. To achieve this goal, our method requires the conditional distribution of the covariate given the confounders to be known or can be estimated from the data. We develop a conditional randomization procedure to simultaneously select the two cutoff values for the marginal and conditional independence test statistics. Promising finite sample performance is demonstrated via extensive simulations and real data applications.

Keywords: Conditional Independence Testing, Confounding Factor, False Discovery Rate, Genomic Data Analysis, Multiple Testing.

1 Introduction

One central theme of genomic data analysis is identifying genomic features associated with a variable of interest, such as disease status. The associated features are subject to further replication

*Roy and Zhang acknowledge partial support from NSF DMS-1811747 and NSF DMS-2113359. Chen and Zhang acknowledge partial support from NIH 1R01GM144351-01 and NIH 1R21HG011662. Address correspondence to Xianyang Zhang (zhangxiany@stat.tamu.edu).

and validation. The validated features could then be followed up for a more in-depth mechanistic study or be used as biomarkers for disease prevention, diagnosis, and prognosis if they have sufficient predictive power (Majewski and Bernards, 2011; Ziegler et al., 2012). Due to the constraint of clinical sample collection, the variable of interest is often correlated with other variables, which may confound the associations of interest. One example is the identification of microbiome biomarkers for endometrial cancer based on a comparison between benign and malignant tumor samples (Walsh et al., 2019). Patients with benign tumors are usually much younger than those with malignant tumors since the progression to malignancy requires multiple genomic events. Age has also been known to be associated with the female reproduction tract microbiome. Therefore, age is a confounding factor, and we need to control it if the aim is to identify cancer-related microbiome biomarkers reliably. Other common sources of confounding in genomic data analysis include environmental changes (Fare et al., 2003; Gasch et al., 2000), cell mixtures (Liang and Cookson, 2014), technical variation or batch effects (Leek et al., 2010; Lazar et al., 2013) and surgical manipulation (Lin et al., 2006). Controlling the confounders could significantly increase the rate of successful validation, reduce the overall cost and shorten the time from discovery to clinical tests. However, due to a substantial multiple testing burden, confounder adjustment exacerbates the already low statistical power for genome-scale association tests. If no confounder adjustment is performed, we are faced with a severely inflated type I error, with the extent of inflation depending on the number and strength of associations with the confounder. Increasing the statistical power of a confounded association study while controlling for the false positives is a statistical topic of critical importance. Surprisingly, few statistical efforts have been made on this important topic.

The traditional way of confounder adjustment for high-dimensional association tests is to adjust for confounders for each genomic feature and correct the individual association p-values for multiple testing using false discovery rate (FDR) control (Benjamini and Hochberg, 1995a; Storey, 2002). This procedure has been a standard statistical practice for genomic association analysis to maintain the correct type I error rate level. However, in practice, confounders may affect only a subset of omics features (Lu et al., 2004; Glass et al., 2013; Gershoni and Pietrokovski, 2017), and adjusting confounders for every omics feature will be an over-adjustment, leading to substantial power loss. To rescue the power, one naive idea is first to test the dependence between the confounder and each omics feature. If the dependence is not statistically significant, we exclude the confounder in the model. Although this strategy substantially improves the power, it suffers from the so-called selection bias (Efron, 2011; Fithian et al., 2014), which inflates the type I error if the significance cutoff is not properly chosen to reflect the selection effect from the first step.

In a recent work, Yi et al. (2021) made a significant step toward solving this problem. The authors proposed a two-dimensional false discovery rate control procedure (2dFDR) based on linear models with the measurement of the omics feature as the outcome. It depends on the unadjusted and adjusted test statistics from fitting both the unadjusted and the confounder-adjusted linear models to each omics feature. The 2dFDR procedure proceeds in two dimensions. The first dimension uses the unadjusted statistic to screen out a large number of irrelevant features (noises) that are not likely to be associated with the covariate of interest or the confounder. In the second dimension, the

procedure uses the adjusted statistic to identify the true signals within the remaining features and control the FDR at the desired level. Although the unadjusted statistic is biased and captures the effects from both the covariate of interest and the confounder, it can be leveraged to increase the signal density and reduce the multiple testing burden in the second dimension. At a high level, the idea of using the unadjusted statistics is similar to the use of marginal utilities in variable screening (Fan and Lv, 2008). However, the 2dFDR procedure takes into account the selection effect from the first dimension and thus provides asymptotically valid inference.

Built upon the 2dFDR procedure in Yi et al. (2021), we propose a general framework for integrating confounder adjustment into multiple testing. The new framework significantly extends the scope and applicability of the original 2dFDR in the following aspects:

1. The new framework relaxes the linear model assumption in Yi et al. (2021). Indeed, the conditional distribution of the omics variables given the variable of interest and the confounders can be arbitrary and completely unknown. Thus, the new framework can be applied to different types of outcomes such as continuous/binary/count outcomes.

2. We allow the joint use of any conditional and marginal independence test statistics in the new framework. Both parametric and nonparametric dependence tests could be applied.

3. Yi et al. (2021) focuses on the case of univariate covariate while the covariate of interest can be multivariate within our framework.

4. The new framework allows different ways of estimating the conditional distribution of the covariate given the confounders. Examples include residual permutation/bootstrap, model-based simulation, Markov Chain Monte Carlo (MCMC), and conditional generative adversarial networks.

5. Due to explicit modeling of the relationship between the variable of interest and confounders, the new method provides more reliable FDR control, especially when the confounding effect is strong.

Finally, it is worth mentioning another line of research on estimating latent confounding factors. Principal component analysis was first suggested by Alter et al. (2000) to estimate the latent confounding factors. More recently, a variety of methods have been proposed for confounder adjustment in similar statistical settings, see, e.g., Friguet et al. (2009); Gagnon-Bartsch and Speed (2012); Leek and Storey (2007, 2008). The theoretical properties of some of these approaches were recently studied by Wang et al. (2017). Although we assume that the confounders are fully observed in our framework, conceptually, our method can also be coupled with the above techniques for confounder adjustment when the confounders are unobserved.

The rest of the article is organized as follows. Section 2 describes the problem setups and a two-dimensional (2d) rejection region based on a primary statistic for testing the conditional independence between the omics feature and the covariate of interest given the confounders and an auxiliary statistic for testing the marginal independence between the omics feature and covariate.

Section 3 introduces an oracle FDR-controlling procedure, where the conditional distribution of the covariate given the confounders is assumed to be known. We prove asymptotic FDR control for the oracle procedure in Section 4. We discuss several ways of estimating the conditional distribution in Section 5. We review some nonparametric conditional independence tests and discuss their use in our framework in Section 6. Sections 7 and 8 are devoted to numerical studies and real data analysis respectively. Section 9 concludes.

2 Problem Statement and 2d Rejection Region

We formulate the feature selection problem by allowing the omics variables to depend on the covariate of interest and confounders arbitrarily. To state the problem and the procedure carefully, suppose we have n i.i.d. samples $\{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)\}_{i=1}^n$ with $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})^\top$ from a population, each of the form $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, where $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top \in \mathbb{R}^m$ and $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$. Here $\mathbf{Y}$ represents a vector of omics features, $\mathbf{X}$ is the covariate of interest, and $\mathbf{Z}$ denotes the set of confounders. We aim to discover as many as possible omics features Y_i that are dependent of $\mathbf{X}$ conditionally on the confounders $\mathbf{Z}$. We formulate this as the problem of testing

$$H_{0,j} : Y_j \perp\!\!\!\perp \mathbf{X}|\mathbf{Z} \quad \text{against} \quad H_{1,j} : Y_j \not\perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$$

for $1 \leq j \leq m$. To tackle this problem, one must adjust for the confounders and the multiplicity in testing. The burden from both adjustments could lead to potential power loss, especially when the confounding effect is strong.

Our idea to resolve this issue is to use two statistics jointly, namely a primary statistic for testing the conditional independence specified in $H_{0,j}$ and an auxiliary statistic for testing the marginal independence $Y_j \perp\!\!\!\perp \mathbf{X}$, for deciding whether or not to reject $H_{0,j}$. The purpose of using the auxiliary statistic is to enrich signals, reduce the multiple testing burden and thus enhance the multiple testing power. As marginal dependence does not necessarily imply conditional dependence (e.g., Y_j and $\mathbf{X}$ are both functions of $\mathbf{Z}$), the use of auxiliary statistics could lead to selection bias and requires proper adjustment in the selection of cut-off values. One of our goals is to carefully design a way to simultaneously select the cut-off values for the primary statistic and the auxiliary statistic to control the FDR at the desired level.

As a motivation, we consider m independent generalized linear models:

$$\begin{aligned} f(Y_j|\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \phi_j) &= \exp\left\{\frac{\theta_j Y_j - b(\theta_j)}{\phi_j} + c(Y_j, \phi_j)\right\}, \\ g(\mathbb{E}[Y_j]) &= g(b'(\theta_j)) = \mathbf{X}^\top \boldsymbol{\alpha}_j + \mathbf{Z}^\top \boldsymbol{\beta}_j, \end{aligned}$$

for $1 \leq j \leq m$, where g is a known link function, θ_j is the canonical parameter, ϕ_j is the dispersion parameter and $\boldsymbol{\alpha}_j \in \mathbb{R}^p, \boldsymbol{\beta}_j \in \mathbb{R}^d$ are the coefficients associated with the covariate of interest and confounders respectively. Under the above model, there are four different categories to consider

A. Associated with both the covariate of interest and confounders: $\boldsymbol{\alpha}_j \neq \mathbf{0}, \boldsymbol{\beta}_j \neq \mathbf{0}$;

B. Solely associated with the covariate of interest: $\boldsymbol{\alpha}_j \neq \mathbf{0}, \boldsymbol{\beta}_j = \mathbf{0}$;

C. Solely associated with the confounders: $\boldsymbol{\alpha}_j = \mathbf{0}, \boldsymbol{\beta}_j \neq \mathbf{0}$;

D. Not associated with either the covariate of interest or confounders: $\boldsymbol{\alpha}_j = \mathbf{0}, \boldsymbol{\beta}_j = \mathbf{0}$.

We note that (i) $\boldsymbol{\alpha}_j = \mathbf{0}$ if and only if $Y_j \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$; (ii) when $\boldsymbol{\beta}_j = \mathbf{0}$, testing the conditional independence boils down to testing the marginal independence $Y_j \perp\!\!\!\perp \mathbf{X}$. As a way to enrich signals, we use a marginal independence test to screen out the omics features in Category D and further use a conditional independence test to pick out the true signals from Categories A and B. More precisely, we let T_j^C and T_j^M be two test statistics computed based on the samples $\{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)\}_{i=1}^n$ for testing the conditional independence $Y_j \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$ and the marginal independence $Y_j \perp\!\!\!\perp \mathbf{X}$, respectively. Throughout the discussions below, we assume that a large positive value of T_j^M (T_j^C) provides evidence against marginal (conditional) independence. The readers are referred to Section 6 for some examples of conditional and unconditional independence tests. Given the thresholds, $t_1, t_2 \geq 0$, the two-dimensional (2d) procedure can be described as follows.

Dimension 1. Use the marginal independence test statistics to determine a preliminary set of features $\mathcal{D}_1 = \left\{1 \leq j \leq m : T_j^M \geq t_1\right\}$.

Dimension 2. Reject $H_{0,j}$ for $T_j^C \geq t_2$ and $j \in \mathcal{D}_1$. As a result, the final set of discoveries is given by $\mathcal{D}_2 = \left\{1 \leq j \leq m : T_j^M \geq t_1, T_j^C \geq t_2\right\}$.

Although marginal dependence does not imply conditional dependence, it can be leveraged to increase the signal density and reduce multiple testing burden in the second dimension. We illustrate the rational behind MF-2dFDR through the following example. A detailed description of the method is provided in the next section.

Example 1. Consider the following data generating process:

$$Y_j \sim \text{Bernoulli}(p_j), \quad \log\left(\frac{p_j}{1-p_j}\right) = \alpha_j X + \beta_j Z, \tag{1}$$

where $X = (\rho Z + \epsilon)/\sqrt{\rho^2 + 1}$ with Z and ϵ being independently generated from $N(0,1)$. Here $\rho \in \{0.1, 0.5, 1\}$ represents weak (+), medium (++) and strong (+++) confounding level. We generate α_j and β_j independently from the mixture distribution:

$$0.15 \times \text{Unif}(-0.7, -0.5) + 0.15 \times \text{Unif}(0.5, 0.7) + 0.7 \times \delta_0$$

where δ_0 denotes a point mass at 0. We let T_j^C be the t-statistic for testing $\tilde{H}_{0,j} : \alpha_j = 0$ under the logistic model (1), and let T_j^M be the t-statistic for testing $\tilde{H}_{0,j}$ under the reduced model by forcing $\beta_j = 0$ in model (1). In Figure 1, we plot the marginal (T^M) statistic against the conditional (T^C) statistic for various confounded scenarios. The standard approach performs (one-dimensional) FDR control based on the conditional statistic (T^C) only (we refer it as 1dFDR). When the correlation between the variable of interest and the confounder (denoted as $\text{cor}(X, Z) = \rho/\sqrt{\rho^2 + 1} \in$

$\{0.1, 0.45, 0.71\}$) is high, the signals (green) and noises (red) overlap much on T^C. To achieve the desired FDR level, 1dFDR requires a high cutoff (black line). For MF-2dFDR, it first uses T^M to exclude a large number of irrelevant features (horizontal blue line). Next, a lower cutoff (vertical blue line) is used to achieve the same FDR level. As a result, it achieves significant power improvement, and the improvement increases with the correlation between the variable of interest and the confounder.

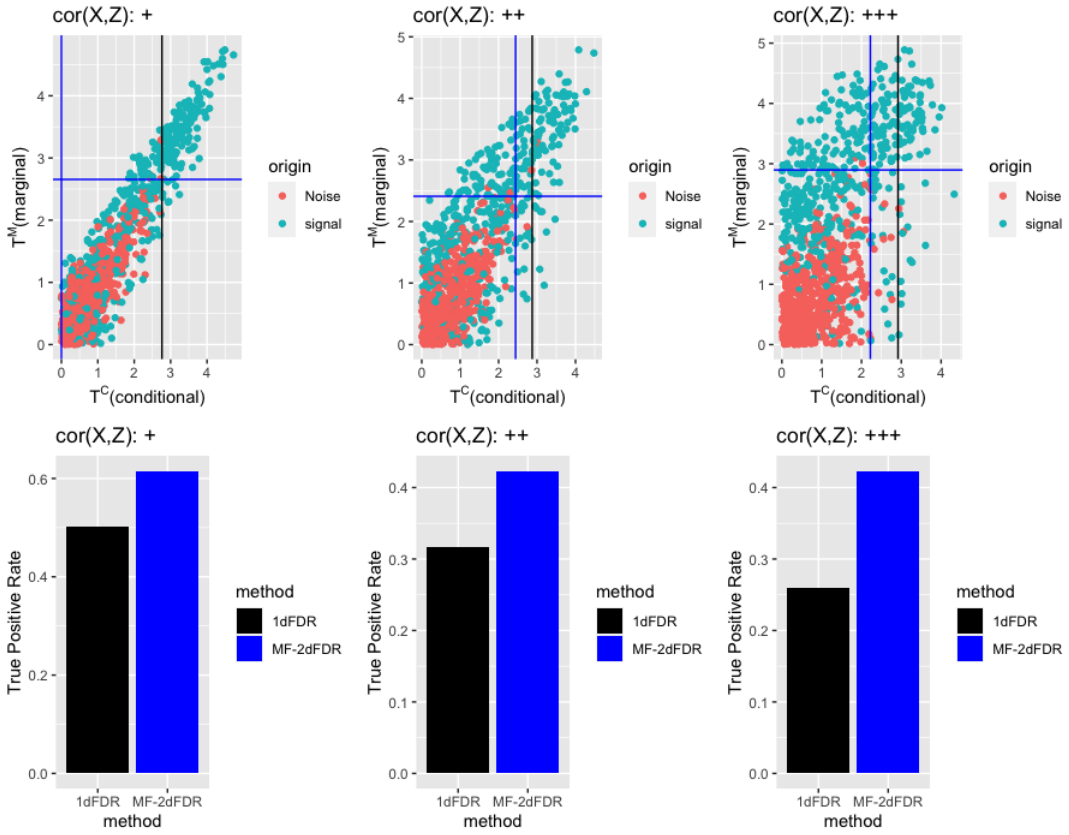


Figure 1: Illustration of MF-2dFDR using simulated datasets. The three panels in the first row denote the decision boundaries for 1dFDR (black line) and MF-2dFDR (blue lines) at the 5% FDR level for three degrees of confounding. 1dFDR relies on the conditional statistic (T^C) only (one dimension) while MF-2dFDR is based on both the marginal and conditional statistics (T^C and T^M), i.e., it uses two dimensions. T^M is used to screen out a large number of irrelevant features (blue horizontal line), followed by a less stringent cutoff of T^C that achieves higher power while keeping the FDR controlled. The figures in the second row illustrate the power difference between the two methods. When the correlation is low ($\rho = 0.1$) using MF-2dFDR provides little improvement over 1dFDR. When the correlation is higher ("++," "+++," $\rho = 0.5, 1$), the signals (green) and noises (red) are more difficult to separate on T^C. By using T^M, MF-2dFDR excludes a large number of noises without losing many signals. The signal density on T^C is enriched, leading to significant power gain.

3 Oracle Procedure

We introduce an oracle FDR-controlling procedure, where we assume that the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$, denoted by $P_{\mathbf{X}|\mathbf{Z}}$ below, is known. Section 5 introduces several ways of estimating this conditional distribution from the observations.

3.1 Estimating the false discovery proportion

Our goal here is to develop a principled way of finding the cutoff values (t_1, t_2) such that the FDR is controlled at a desired level while the number of rejections is as large as possible. Let $\mathcal{M}_0 = \{1 \le j \le m : H_{0,j} \text{ is true}\}$ and $m_0 = |\mathcal{M}_0|$ be the set and the number of true null hypotheses respectively. Write $\widetilde{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, $\widetilde{\mathbf{Y}}_j = (Y_{1,j}, \ldots, Y_{n,j})^\top$ and $\widetilde{\mathbf{Z}} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^\top$. Based on the 2d rejection region, the false discovery proportion (FDP) is given by

$$\mathrm{FDP}(t_1, t_2) = \frac{\sum_{j \in \mathcal{M}_0} \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}}{1 \vee \sum_{j=1}^m \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}}, \tag{2}$$

where $a \vee b = \max\{a, b\}$ for $a, b \in \mathbb{R}$. Note that the FDP is zero when no rejection is made. We replace the numerator in the definition of $\mathrm{FDP}(t_1, t_2)$ by its conditional expectation with respect to $\widetilde{\mathbf{X}}$ given $\widetilde{\mathbf{Y}}_j$ and $\widetilde{\mathbf{Z}}$, which leads to the following approximate upper bound on the FDP:

$$\begin{aligned}\mathrm{FDP}(t_1, t_2) \approx & \frac{\sum_{j \in \mathcal{M}_0} \mathbb{P}_0(T_j^M \ge t_1, T_j^C \ge t_2 | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})}{1 \vee \sum_{j=1}^m \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}} \\ \le & \frac{\sum_{j=1}^m \mathbb{P}_0(T_j^M \ge t_1, T_j^C \ge t_2 | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})}{1 \vee \sum_{j=1}^m \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}} := \mathrm{FDP}_{\mathrm{oracle}}(t_1, t_2),\end{aligned} \tag{3}$$

where $\mathbb{P}_0(\cdot | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})$ denotes the conditional probability under the null hypothesis $H_{0,j}$. The upper bound $\mathrm{FDP}_{\mathrm{oracle}}(t_1, t_2)$ relies on the conditional distribution $P_{\mathbf{X}|\mathbf{Z}}$. To find a feasible conservative estimator of the FDP, it remains to estimate the conditional probabilities in the numerator of $\mathrm{FDP}_{\mathrm{oracle}}(t_1, t_2)$. To this end, we write $T_j^M = T_j^M(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j)$ and $T_j^C = T_j^C(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})$ to emphasize their dependence on the samples. As $\mathbb{P}(\widetilde{\mathbf{X}} | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}) = \mathbb{P}(\widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}})$ under $H_{0,j}$ and $\mathbb{P}(\widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}) = \prod_{i=1}^n \mathbb{P}_{\mathbf{X}|\mathbf{Z}}(\mathbf{X}_i | \mathbf{Z}_i)$, we have under $H_{0,j}$ that

$$\begin{aligned}&\mathbb{P}_0(T_j^M(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j) \ge t_1, T_j^C(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}) \ge t_2 | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}) \\ =&\mathbb{E}\left[\mathbf{1}\{T_j^M(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j) \ge t_1, T_j^C(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}) \ge t_2\} | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}\right] \\ =&\int \mathbf{1}\{T_j^M(\widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}}_j) \ge t_1, T_j^C(\widetilde{\mathbf{x}}, \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}) \ge t_2\} d \prod_{i=1}^n \mathbb{P}_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i | \mathbf{Z}_i),\end{aligned}$$

where $\widetilde{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ with $\mathbf{x}_i \in \mathbb{R}^p$, which can be calculated once we know the conditional distribution $P_{\mathbf{X}|\mathbf{Z}}$. One way to approximate $P_0(T_j^M \ge t_1, T_j^C \ge t_2 | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})$ is via Monte Carlo simulation. Specifically, we generate

$$\mathbf{X}_{i,b} \sim^{\mathrm{ind}} P_{\mathbf{X}|\mathbf{Z}}(\cdot | \mathbf{Z}_i), \quad i = 1, 2, \ldots, n, \quad b = 1, 2, \ldots, B.$$

Denote by $T_{j,b}^M$ and $T_{j,b}^C$ the marginal and conditional independence test statistics computed based on $(\widetilde{\mathbf{X}}_b, \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})$ with $\widetilde{\mathbf{X}}_b = (\mathbf{X}_{1,b}, \ldots, \mathbf{X}_{n,b})^\top$ respectively. We propose to estimate $P_0(T_j^M \ge t_1, T_j^C \ge$

$t_2|\widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}})$ by

$$\bar{F}_{j,B}(t_1, t_2) := \frac{1}{B+1} \sum_{b=0}^{B} \mathbf{1}\{T_{j,b}^M \geq t_1, T_{j,b}^C \geq t_2\}$$

with $(T_{j,0}^M, T_{j,0}^C) = (T_j^M, T_j^C)$. Hence a conservative estimate for the FDP is given by

$$\widetilde{\mathrm{FDP}}(t_1, t_2) = \frac{\sum_{j=1}^m \bar{F}_{j,B}(t_1, t_2)}{1 \vee \sum_{j=1}^m \mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}}.$$

3.2 Finding the optimal cut-off

We now introduce a greedy approach to select the cut-offs. For a desired FDR level q, we first define

$$\mathcal{F}_q = \{(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+ : \widetilde{\mathrm{FDP}}(t_1, t_2) \leq q\}$$

as the feasible set that contains all the cut-off values controlling the FDP estimate at the level q. We then select the optimal cut-off as the one delivering the most number of rejections from the feasible set:

$$(\tilde{t}_1, \tilde{t}_2) = \underset{(t_1,t_2)\in\mathcal{F}_q}{\arg\max} \sum_{j=1}^m \mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}.$$

Finally, we reject all the hypotheses $H_{0,j}$ such that

$$T_j^M \geq \tilde{t}_1 \quad \text{and} \quad T_j^C \geq \tilde{t}_2.$$

3.3 Family-wise error rate control

Family-wise error rate (FWER), referring to the probability of making one false discovery, provides more stringent type I error rate control. It is preferable to the FDR if the overall conclusion from various individual inferences is likely to be erroneous when at least one of them is, or the existence of a single false claim would cause significant loss. It is natural to ask whether our method can be modified to control other error measures such as FWER. Here we describe such a procedure to control the FWER. Given the rejection rule $\mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}$, we let $\widetilde{\mathrm{FWER}}(t_1, t_2) := \sum_{j=1}^m \bar{F}_{j,B}(t_1, t_2)$ be an estimate of the FWER. We choose the optimal cut-off value as the one that maximizes the number of rejections while controls the FWER estimate at a prespecified level q:

$$(\breve{t}_1, \breve{t}_2) = \underset{(t_1,t_2)\in\mathcal{G}_q}{\arg\max} \sum_{j=1}^m \mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\},$$

where $\mathcal{G}_q = \{(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+ : \widetilde{\mathrm{FWER}}(t_1, t_2) \leq q\}$. Then we reject $H_{0,j}$ whenever $T_j^M \geq \breve{t}_1$ and $T_j^C \geq \breve{t}_2$. We name the above procedure MF-2dFWER. In Section 10.4, we investigate its finite sample performance and report the empirical FWER and power for MF-2dFWER and its corresponding 1d version (1dFWER) in Figures 25 and 26.

3.4 Estimating the null proportion

Following the idea in Storey (2002), we can further improve the power of our method by estimating the proportion of null hypotheses. As a motivation, we suppose T_j^C follows the mixture distribution

$$\pi_0 \mathbb{P}_0 + (1-\pi_0)\mathbb{P}_1,$$

where π_0 represents the null proportion, $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the distributions under the null and alternative, respectively. Under this two-group mixture model, we have

$$P(T_j^C \le \lambda) = \pi_0 \mathbb{P}_0(T_j^C \le \lambda) + (1-\pi_0)\mathbb{P}_1(T_j^C \le \lambda) \ge \pi_0 \mathbb{P}_0(T_j^C \le \lambda),$$

which implies that

$$\frac{\sum_{j=1}^m \mathbf{1}\{T_j^C \le \lambda\}}{\sum_{j=1}^m \mathbb{P}_0(T_j^C \le \lambda)} \approx \frac{\sum_{j=1}^m P(T_j^C \le \lambda)}{\sum_{j=1}^m \mathbb{P}_0(T_j^C \le \lambda)} \ge \pi_0,$$

where the approximation is due to the law of large numbers. Therefore, we propose to estimate the null proportion π_0 by

$$\widehat{\pi}_0(\lambda) = 1 \wedge \frac{\sum_{j=1}^m \mathbf{1}\{T_j^C \le \lambda\}}{\sum_{j=1}^m F_{j,B}(\lambda)}, \quad \text{where } F_{j,B}(\lambda) := \frac{1}{B+1}\sum_{b=0}^B \mathbf{1}\{T_{j,b}^C \le \lambda\}.$$

We can then implement the MF-2dFDR based on the following estimate of the FDP:

$$\widetilde{\mathrm{FDP}}_\lambda(t_1,t_2) = \frac{\widehat{\pi}_0(\lambda)\sum_{j=1}^m \bar{F}_{j,B}(t_1,t_2)}{1 \vee \sum_{j=1}^m \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}},$$

which can be regarded as John Storey's version of the MF-2dFDR procedure.

4 FDR Control

We first show that under the global null, a version of the MF-2dFDR procedure provides finite sample FDR control (or equivalently FWER control). The key to the proof relies on the symmetry of the statistics $\{(T_{j,b}^M, T_{j,b}^C) : j = 1,2,\ldots,m\}$ across the index b. Let $\{(t_1(s), t_2(s)) \in \mathbb{R}^+ \times \mathbb{R}^+ : 1 \le s \le \mathcal{S}\}$ be a sequence of thresholds such that $t_1(s) \le t_1(s')$ and $t_2(s) \le t_2(s')$ for $1 \le s < s' \le \mathcal{S}$. Let $V^b(s) = \sum_{j=1}^m \mathbf{1}\{T_{j,b}^M \ge t_1(s), T_{j,b}^C \ge t_2(s)\}$ and $\widetilde{V}^b(s) = \sum_{i \in \mathcal{M}_0} \mathbf{1}\{T_{j,b}^M \ge t_1(s), T_{j,b}^C \ge t_2(s)\}$ for $0 \le b \le B$. Define

$$s^* = \min\left\{1 \le s \le \mathcal{S} : \frac{(B+1)^{-1}\sum_{b=0}^B V^b(s)}{1 \vee V^0(s)} \le q\right\}.$$

Then we reject any hypothesis such that $T_{j,0}^M \ge t_1(s^*)$ and $T_{j,0}^C \ge t_2(s^*)$.

Theorem 1. *Under the global null, the MF-2dFDR procedure provides finite sample FDR control or equivalently FWER control.*

Under general setting, the symmetry among $(T^M_{j,b}, T^C_{j,b})_{j=1}^m$ no longer holds and the finite sample FDR control is not guaranteed. Fortunately, we manage to show that the MF-2dFDR provides asymptotic FDR control as n, m both diverge to infinity. To achieve this goal, we impose the following assumptions.

Assumption 1. *Conditional on* $(\mathbf{X}, \mathbf{Z})$*,* Y_j*'s are independent across* $1 \le j \le m$*. Moreover, for* $j \in \mathcal{M}_0$*,* Y_j*'s are independent conditional on* $\mathbf{Z}$*.*

Assumption 1 requires that the marginal models of Y_j conditional on $\mathbf{X}$ and $\mathbf{Z}$ are independent across $1 \le j \le m$. For instance, consider the model

$$\begin{aligned} Y_j &= u_j(\mathbf{X}) + v_j(\mathbf{Z}) + \epsilon_j, \quad j \notin \mathcal{M}_0, \\ Y_j &= v_j(\mathbf{Z}) + \epsilon_j, \quad j \in \mathcal{M}_0, \end{aligned} \tag{4}$$

where $u_j(\cdot)$ and $v_j(\cdot)$ are some functions defined on $\mathbb{R}^p$ and $\mathbb{R}^d$ respectively. In this case, Assumption 1 is fulfilled provided that ϵ_j's are independent across j.

Assumption 2. *Recall that* m_0 *denotes the number of true null hypotheses. Suppose* $m_0/m \to \pi_0 \in (0,1)$ *and there exist two bivariate functions* $\widetilde{V}(\cdot,\cdot)$ *and* $\widetilde{S}(\cdot,\cdot)$ *defined on* $\mathbb{R}^+ \times \mathbb{R}^+$ *such that*

$$\left| \frac{1}{m_0} \sum_{j \in \mathcal{M}_0} P(T_j^M \ge t_1, T_j^C \ge t_2 | \widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}) - \widetilde{V}(t_1, t_2) \right| \to^p 0,$$
$$\left| \frac{1}{m} \sum_{j=1}^{m} P(T_j^M \ge t_1, T_j^C \ge t_2 | \widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}) - \widetilde{S}(t_1, t_2) \right| \to^p 0,$$

for any fixed $t_1, t_2 \ge 0$*.*

Assumption 2 is a high-level condition. We justify this assumption under the linear models in Section 10.1. Our next assumption is similar to the requirement in Theorem 4 of Storey et al. (2004), which ensures the existence of cut-off values to control the FDR at level q. It reduces the search region for the optimal cut-offs to a rectangle of the form $[0, t_{0,1}] \times [0, t_{0,2}]$.

Assumption 3. *Assume that there exist* $t_{0,1}$ *and* $t_{0,2}$ *such that,*

$$\frac{\pi_0 \widetilde{V}(t_{0,1}, 0)}{\widetilde{S}(t_{0,1}, 0)} \le q' < q, \quad \frac{\pi_0 \widetilde{V}(0, t_{0,2})}{\widetilde{S}(0, t_{0,2})} \le q'' < q,$$

and $\widetilde{S}(t_{0,1}, t_{0,2}) > c > 0$*, where* π_0 *is defined in Assumption 2.*

To state the main theorem, we define the optimal threshold value based on the oracle FDP

estimate as

$$(t_1^*, t_2^*) = \underset{(t_1,t_2)\in\mathbb{R}^+\times\mathbb{R}^+:\mathrm{FDP}_{\mathrm{oracle}}(t_1,t_2)\le q}{\arg\max} \sum_{j=1}^{m} \mathbf{1}\{T_j^M \ge t_1, T_j^C \ge t_2\}.$$

The theorem below establishes the asymptotic FDR control of the MF-2dFDR procedure.

Theorem 2. *Under Assumptions 1-3,*

$$\limsup_{n,m\to+\infty} \mathbb{E}\left[FDP(t_1^*, t_2^*)\right] \le q,$$

where $FDP(t_1, t_2)$ is defined in (2).

5 Estimating the Conditional Distributions

As the conditional distribution $P_{\mathbf{X}|\mathbf{Z}}$ is seldom known, we need to estimate it from the data. There are indeed several ways of generating samples from $P_{\mathbf{X}|\mathbf{Z}}$. Examples include classical methods such as residual permutation (Winkler et al., 2014) and parametric bootstrap (Davison and Hinkley, 1997) as well as modern approaches such as conditional generative adversarial network (conditional GAN) (Mirza and Osindero, 2014; Zhou et al., 2022). In the following subsections, we shall describe the residual permutation, residual bootstrap, and parametric bootstrap in more detail. Compared to the conditional GAN, these procedures are more suitable for omics applications, given the limited sample sizes in many omics association studies.

5.1 Residual permutation and residual bootstrap

When $\mathbf{X}$ is a continuous random vector, we can model the relationship between $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ and $\widetilde{\mathbf{Z}} \in \mathbb{R}^{n\times d}$ through a multivariate linear regression model given by

$$\widetilde{\mathbf{X}} = \widetilde{\mathbf{Z}}\mathbf{B} + \mathbf{E}, \tag{5}$$

where $\mathbf{B} \in \mathbb{R}^{d\times p}$ is the matrix of coefficients and $\mathbf{E} \in \mathbb{R}^{n\times p}$ is the error matrix. Consider the following strategy to generate samples from $P_{\mathbf{X}|\mathbf{Z}}$.

Step 1: Fitting regression model. Fit the multivariate linear regression model in (5). Let $\widehat{\mathbf{E}} = \widetilde{\mathbf{X}} - \widetilde{\mathbf{Z}}\widehat{\mathbf{B}}$ be the residuals from the fitted model, where $\widehat{\mathbf{B}}$ is the least squares estimate of $\mathbf{B}$.

Step 2: Residual permutation. Permute the rows of the residual matrix $\widehat{\mathbf{E}}$ and denote the resulting matrix by $\widehat{\mathbf{E}}^*$. Let $\widetilde{\mathbf{X}}_b = (\mathbf{X}_{1,b}, \ldots, \mathbf{X}_{n,b})^\top = \widetilde{\mathbf{Z}}\widehat{\mathbf{B}} + \widehat{\mathbf{E}}^*$.

Step 2′: Residual bootstrap. Let $\widehat{\mathbf{E}}^{**}$ be a $n \times p$ matrix whose rows are sampled with replacement from those of $\widehat{\mathbf{E}}$. Let $\widetilde{\mathbf{X}}_b = (\mathbf{X}_{1,b}, \ldots, \mathbf{X}_{n,b})^\top = \widetilde{\mathbf{Z}}\widehat{\mathbf{B}} + \widehat{\mathbf{E}}^{**}$.

Remark 1. To allow nonlinearity, we can replace Z_i by $(g_1(Z_i), \ldots, g_{d'}(Z_i)) \in \mathbb{R}^{d'}$ for some transformations $(g_1, \ldots, g_{d'})$ in the multivariate regression model.

5.2 Parametric bootstrap

Suppose the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ takes the parametric form of $P_{\mathbf{X}|\mathbf{Z}}(\mathbf{X}_i|\mathbf{Z}_i;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^r$ is an unknown parameter. It is natural to estimate the parameter by maximizing the conditional log-likelihood

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log P_{\mathbf{X}|\mathbf{Z}}(\mathbf{X}_i|\mathbf{Z}_i;\theta).$$

Then we can generate $\mathbf{X}_{i,b}$ from the estimated likelihood $P_{\mathbf{X}|\mathbf{Z}}(\mathbf{X}_i|\mathbf{Z}_i;\widehat{\theta})$. For example, suppose $\mathbf{X}$ is a Bernoulli random variable with the conditional success probability given by $\{1+\exp(-\mathbf{Z}_i^\top\theta)\}^{-1}$. Then we can sample $\mathbf{X}_{i,b}$ from the Bernoulli distribution with success probability $\{1+\exp(-\mathbf{Z}_i^\top\widehat{\theta})\}^{-1}$, where $\widehat{\theta}$ is an estimate of θ by fitting a logistic model to the data with $\widetilde{\mathbf{X}}$ being the binary response and $\widetilde{\mathbf{Z}}$ being the covariates.

6 Independence Tests

We review some parametric and nonparametric unconditional/conditional independence tests and discuss their use within our framework. In Section 6.1, we focus on the model-based (parametric) independence tests. In Sections 6.2.1-6.2.2, we consider two types of nonparametric independence tests targeting linear and nonlinear dependence respectively. These three types of independence tests will all be implemented in our numerical studies.

6.1 Model-based statistics

Suppose the conditional likelihood of Y_j given $\mathbf{X}$ and $\mathbf{Z}$ has the form of

$$P_{Y_j|\mathbf{X},\mathbf{Z}}(Y_j|\mathbf{X}^\top\boldsymbol{\alpha}_j + \mathbf{Z}^\top\boldsymbol{\beta}_j). \tag{6}$$

The log-likelihood function based on the observations is given by

$$L_{n,j}(\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j) = \sum_{i=1}^{n} \log P_{Y_j|\mathbf{X},\mathbf{Z}}(Y_{i,j}|\mathbf{X}_i^\top\boldsymbol{\alpha}_j + \mathbf{Z}_i^\top\boldsymbol{\beta}_j).$$

In this case, testing $H_{0,j}$ is equivalent to testing whether $\boldsymbol{\alpha}_j$ is zero. Thus we let T_j^C be a statistic for testing $\boldsymbol{\alpha}_j = 0$ under the model (6). Examples include the Wald test and the likelihood-ratio test. To test the marginal independence, we consider the reduced model $P_{Y_j|\mathbf{X},\mathbf{Z}}(Y_j|\mathbf{X}^\top\boldsymbol{\alpha}_j)$ by forcing $\boldsymbol{\beta}_j = 0$ in (6). Under the reduced model, we let T_j^M be a statistic for testing $\boldsymbol{\alpha}_j = 0$, which can be viewed as testing the marginal independence $Y_j \perp\!\!\!\perp \mathbf{X}$. When $P_{Y_j|\mathbf{X},\mathbf{Z}}$ is the likelihood function associated with a linear model with Gaussian error, we can let T_j^C and T_j^M be the adjusted and unadjusted z-statistics considered in Yi et al. (2021). In this sense, the statistics in Yi et al. (2021) fall into our framework.

6.2 Nonparametric dependence metrics

Nonparametric dependence testing, aiming to determine whether two random vectors are dependent without specifying the exact parametric forms of the distributions, is one of the fundamental problems in statistics. Classical metrics or test statistics for dependence testing include the RV coefficient, rank correlation coefficient, and nonparametric Cramér-von Mises type statistics. Modern approaches are built on distance and kernel embedding, which can detect non-linear and non-monotone dependence. Notable examples include the distance covariance (Székely et al., 2007), Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005, 2007) and the sign distance covariance (Bergsma and Dassios, 2014). Below we shall review the RV coefficient and HSIC and discuss their conditional versions for testing the conditional independence.

6.2.1 RV coefficients

Pearson correlation and partial correlation coefficients are perhaps the most commonly used nonparametric dependence metrics for measuring marginal and conditional dependence. Here we describe the RV coefficient and its conditional version as multivariate generalizations of the squared Pearson correlation coefficient and the squared partial correlation coefficient for detecting linear and conditional linear dependence.

For two random vectors $\mathbf{U}$ and $\mathbf{V}$, we let $\Sigma_{\mathbf{U},\mathbf{V}}$ be the covariance matrix between $\mathbf{U}$ and $\mathbf{V}$. The RV coefficient between $\mathbf{X}$ and Y_j is defined as

$$\mathrm{RV}(\mathbf{X}, Y_j) = \frac{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X},Y_j}\mathbf{\Sigma}_{Y_j,\mathbf{X}})}{\sqrt{\mathrm{tr}(\Sigma^2_{\mathbf{X},\mathbf{X}})\mathrm{tr}(\Sigma^2_{Y_j,Y_j})}}.$$

To estimate the RV coefficient, we simply replace the covariance matrices in the definition above with the sample covariance matrices.

To introduce the conditional version of the RV coefficient, we let $\mathbf{e}_{\mathbf{X}}$ and e_{Y_j} be the residuals by regressing $\mathbf{X}$ and Y_j on $\mathbf{Z}$ respectively. The conditional RV coefficient is defined as

$$\mathrm{cRV}(\mathbf{X}, Y_j|\mathbf{Z}) = \mathrm{cRV}(\mathbf{e}_{\mathbf{X}}, e_{Y_j}).$$

Similar to Remark 1, to account for the nonlinear dependence of $\mathbf{X}$ and Y_j on $\mathbf{Z}$, we can replace $\mathbf{Z}$ by certain basis function transform on it, e.g., spline transformation.

6.2.2 Hilbert-Schmidt independence criterion

Hilbert-Schmidt Independence Criterion (HSIC) was introduced as a kernel-based independence measure by Gretton et al. (2005, 2007). Let $k_p(\cdot,\cdot)$ be a reproducing kernel Hilbert space (RKHS) kernel defined on $\mathbb{R}^p \times \mathbb{R}^p$. Commonly used kernels in this context include the Gaussian kernel and the Laplacian kernel. The HSIC for quantifying the strength of dependence between $\mathbf{X}$ and Y_j can

be defined as

$$\mathrm{HSIC}(\mathbf{X}, Y_j) = \mathbb{E}[k_p(\mathbf{X}, \mathbf{X}')k_1(Y_j, Y_j')] + \mathbb{E}[k_p(\mathbf{X}, \mathbf{X}')]\mathbb{E}[k_1(Y_j, Y_j')] - 2\mathbb{E}[k_p(\mathbf{X}, \mathbf{X}')k_1(Y_j, Y_j'')]$$

where $(\mathbf{X}', Y_j')$ and $(\mathbf{X}'', Y_j'')$ are independent copies of $(\mathbf{X}, Y_j)$. When k_p and k_1 are characteristic kernels (Sriperumbudur et al., 2011), HSIC completely characterizes the dependence in the sense that $\mathbf{X}$ and Y_j are independent if and only if $\mathrm{HSIC}(\mathbf{X}, Y_j) = 0$. To estimate the HSIC, define $\mathbf{K}_{\mathbf{X}} = (k_{\mathbf{X},ab})_{a,b=1}^n$ with $k_{\mathbf{X},ab} = k_p(\mathbf{X}_a, \mathbf{X}_b)$ and $\mathbf{K}_{Y_j} = (k_{Y_j,ab})_{a,b=1}^n$ with $k_{Y_j,ab} = k_1(Y_{a,j}, Y_{b,j})$. Let $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ with $\mathbf{1}$ being the n-dimensional vector of all ones. Set $\widetilde{\mathbf{K}}_{\mathbf{X}} = \mathbf{H}\mathbf{K}_{\mathbf{X}}\mathbf{H}$ and $\widetilde{\mathbf{K}}_{Y_j} = \mathbf{H}\mathbf{K}_{Y_j}\mathbf{H}$. The sample HSIC is defined as

$$\widehat{\mathrm{HSIC}}(\mathbf{X}, Y_j) = \frac{1}{n}\mathrm{Tr}(\widetilde{\mathbf{K}}_{\mathbf{X}}\widetilde{\mathbf{K}}_{Y_j}),$$

which has been shown to be a consistent estimator, see Gretton et al. (2005).

A conditional version of the HSIC (cHSIC) for measuring and testing conditional dependence was proposed in Zhang et al. (2012). Here we describe the construction of their statistic. Let $\mathbf{K}_{\mathbf{X},\mathbf{Z}} = (k_{\mathbf{X},\mathbf{Z},ab})_{a,b=1}^n$ with $k_{\mathbf{X},\mathbf{Z},ab} = k_{p+d}((\mathbf{X}_a, \mathbf{Z}_a), (\mathbf{X}_b, \mathbf{Z}_b))$ and define $\mathbf{K}_{Y_j,\mathbf{Z}}$ in a similar way. Denote by $\widetilde{\mathbf{K}}_{\mathbf{X},\mathbf{Z}} = \mathbf{H}\mathbf{K}_{\mathbf{X},\mathbf{Z}}\mathbf{H}$ and $\widetilde{\mathbf{K}}_{Y_j,\mathbf{Z}} = \mathbf{H}\mathbf{K}_{Y_j,\mathbf{Z}}\mathbf{H}$ the centered versions of $\mathbf{K}_{\mathbf{X},\mathbf{Z}}$ and $\mathbf{K}_{Y_j,\mathbf{Z}}$ respectively. Further define

$$\widetilde{\mathbf{K}}_{\mathbf{X}\mathbf{Z}|\mathbf{Z}} = \epsilon^2(\widetilde{\mathbf{K}}_{\mathbf{X}\mathbf{Z}} + \epsilon\mathbf{I})^{-1}\widetilde{\mathbf{K}}_{\mathbf{X}\mathbf{Z}}(\widetilde{\mathbf{K}}_{\mathbf{X}\mathbf{Z}} + \epsilon\mathbf{I})^{-1},$$
$$\widetilde{\mathbf{K}}_{Y_j\mathbf{Z}|\mathbf{Z}} = \epsilon^2(\widetilde{\mathbf{K}}_{Y_j\mathbf{Z}} + \epsilon\mathbf{I})^{-1}\widetilde{\mathbf{K}}_{Y_j\mathbf{Z}}(\widetilde{\mathbf{K}}_{Y_j\mathbf{Z}} + \epsilon\mathbf{I})^{-1},$$

for some small positive constant ϵ. The sample cHSIC is given by

$$\widehat{\mathrm{cHSIC}}(\mathbf{X}, Y_j|\mathbf{Z}) = \frac{1}{n}\mathrm{Tr}(\widetilde{\mathbf{K}}_{\mathbf{X}\mathbf{Z}|\mathbf{Z}}\widetilde{\mathbf{K}}_{Y_j\mathbf{Z}|\mathbf{Z}}).$$

We refer the readers to Zhang et al. (2012) for more detailed properties about the cHSIC.

7 Numerical Studies

7.1 Simulation setting

We conduct comprehensive simulations to evaluate the performance of MF-2dFDR and compare it to competing methods. Throughout the simulations, we control the following three factors, namely the degree of confounding (ρ, which determines the strength of association between $\mathbf{X}$ and $\mathbf{Z}$), the signal strength (distributions of α_j and β_j) and the signal density (proportion of non-zero elements in $\{\alpha_j\}$ and $\{\beta_j\}$). Specifically, we generate α_j and β_j independently over j from the mixture distribution

$$\frac{\pi}{2} \times U(-l - 0.2, -l) + \frac{\pi}{2} \times U(l, l + 0.2) + (1 - \pi) \times \delta_0$$

where $\pi \in (0,1)$ and δ_0 denotes a point mass at 0. For each factor, we consider three different scenarios:

1. Degree of confounding: $\rho \in \{0.1, 1, 1.5\}$ roughly correspond to weak (+), medium (++) and strong(+++) confounding respectively. See Section 10.3 for the role of ρ in each simulated model.

2. Signal density: $\pi \in \{5\%, 10\%, 20\%\}$ represents low, medium and high signal density respectively.

3. Signal effect: $l \in \{0.2, 0.3, 0.4\}$ represents weak, moderate and strong effect respectively.

We report the empirical FDR and power averaged over 100 simulation runs for all possible combinations of the three factors.

7.2 Competing methods

We compare the finite sample performance of the following seven methods.

1. 1dFDR-MS: The 1d procedure based on the t-statistics for testing $\boldsymbol{\alpha}_j = 0$ under the full model (see the detailed descriptions of each data generating model in Section 10.3). The 1d procedure is essentially the same as the MF-2dFDR procedure, except that instead of a two-dimensional rejection region, we are searching for a cutoff along a single dimension, namely that of the conditional statistic. The FDP is estimated using the resampled $\mathbf{X}_{i,b}$ (from the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$), but only the conditional statistic is used for estimating the number of false rejections, as opposed to both in the oracle procedure. The statistics used in this 1d procedure is the model-based statistic, i.e., the z-statistic (or t-statistic, depending on the model) corresponding to the coefficient of $\mathbf{X}$ for a full model fit.

2. 1dFDR-RV: The 1d procedure based on the conditional RV coefficient. To account for the potential non-linearity in the underlying relationship between $\mathbf{X}$ and $\mathbf{Z}$ (and similarly, $\mathbf{Y}$ and $\mathbf{Z}$), the residuals obtained from a cubic spline regression of $\mathbf{X}$ on $\mathbf{Z}$ (and similarly, $\mathbf{Y}$ on $\mathbf{Z}$) have been used in the calculation of the conditional RV coefficient.

3. 1dFDR-HSIC: The 1d procedure based on the cHSIC described in Section 6.2.2.

4. 2dFDR: The 2dFDR procedure proposed in Yi et al. (2021), which is based on linear models with the measurement of the omics feature as the outcome and the covariate of interest and confounders as the predictors.

5. MF-2dFDR-MS: The proposed MF-2dFDR procedure with T_j^C and T_j^M being the t-statistics for testing $\boldsymbol{\alpha}_j = 0$ under the full model and reduced model as described in Section 6.1.

6. MF-2dFDR-RV: The proposed MF-2dFDR procedure with $T_j^C = \widehat{\text{cRV}}(\mathbf{X}, Y_j|\mathbf{Z})$ and $T_j^M = \widehat{\text{RV}}(\mathbf{X}, Y_j)$ which denote the sample estimates of the conditional and the unconditional RV

coefficients respectively. As before, to account for the potential non-linearity in the underlying relationship between $\mathbf{X}$ and $\mathbf{Z}$ (and similarly, $\mathbf{Y}$ and $\mathbf{Z}$), the residuals obtained from a cubic spline regression of $\mathbf{X}$ on $\mathbf{Z}$ (and similarly, $\mathbf{Y}$ on $\mathbf{Z}$) have been used in the calculation of the conditional RV coefficient.

7. MF-2dFDR-HSIC: The proposed MF-2dFDR procedure with $T_j^C = \widehat{\text{cHSIC}}(\mathbf{X}, Y_j|\mathbf{Z})$ and $T_j^M = \widehat{\text{HSIC}}(\mathbf{X}, Y_j)$, where we set $\epsilon = 0.001$ and used the Gaussian kernel with the bandwidth parameter chosen using the median heuristic (Garreau et al., 2017).

The 1d procedure can be viewed as a special case of the corresponding 2d procedure by forcing the cutoff of the auxiliary statistic to be zero. As the 2d procedure is searching over a larger rejection region (by allowing the cutoff of the auxiliary statistic to be greater than zero and meanwhile lowering the cutoff for the primary statistic), the proposed 2d procedure is guaranteed to make more rejections in finite sample.

7.3 Data generating processes

To examine the performance of the above methods under different settings, we consider the following data generation scenarios. As $\mathbf{X}$ and $\mathbf{Z}$ are univariate in all cases, we denote them by X and Z. The detailed models are provided in Section 10.3. Throughout the simulations, we set the sample size n to be 100 and the number of hypotheses m (i.e., the number of features) to be 1000.

A. *Linear/nonlinear models with continuous X and Z.* Consider the additive model

$$Y_j = \alpha_j f(X) + \beta_j g(Z) + \epsilon_j, \quad \epsilon_j \sim N(0,1), \quad j = 1, \ldots, m. \tag{7}$$

Here X and Z are associated with each other through the following model:

$$X \sim N(\rho h(Z), 1), \quad Z \sim N(0,1), \tag{8}$$

where ρ controls the degree of confounding and $h : \mathbb{R} \to \mathbb{R}$ is a possibility nonlinear function. We rescale X to dissociate any possible entanglement between signal strength and the degree of confounding. This type of simulation setup has been used in Models 1-4 to explore the effect of the relations among X, Y_j, and Z on the FDR and power. The empirical FDR and power of 1dFDR-RV, 1dFDR-HSIC, 2dFDR, MF-2dFDR-RV and MF-2dFDR-HSIC are summarized in Figures 2-3 and again in 14-15. 1dFDR-MS and MF-2dFDR-MS have not been included under this scenario because the statistics associated with these procedures are directly proportional to the statistics in 1dFDR-RV and MF-2dFDR-RV, respectively.

B. *Linear/nonlinear models with discrete X and continuous Z.* In particular, we consider the functional form in (7) and generate

$$X \sim \text{Bernoulli}\left(\frac{e^{\rho Z}}{1 + e^{\rho Z}}\right),$$

where $Z \sim N(0,1)$. Models 5-7 explore this setup. In this case, we generate $X_{i,b}$ through a fitted logistic regression model using Z as the predictor. We report the FDR and power for 1dFDR-MS, 1dFDR-RV, 2dFDR, MF-2dFDR-MS and MF-2dFDR-RV as described in Section 7.2 in Figures 16-18. MF-2dFDR-HSIC and 1dFDR-HSIC are not used in this data generating setup because for binary variables, HSIC is not efficient and the bandwidth parameter is not well-defined.

C. *Linear models with discrete X and Z.* We consider the linear model

$$Y_j = \alpha_j X + \beta_j Z + \epsilon_j,$$

where

$$X \sim \text{Bernoulli}\left(\frac{e^{\rho Z}}{1+e^{\rho Z}}\right) \quad \text{and} \quad Z \sim \text{Bernoulli}(0.7).$$

The results for 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS and MF-2dFDR-RV are reported in Figure 19.

D. *Binary response.* The following logistic regression model has been considered:

$$Y_j \sim \text{Bernoulli}(p_j), \quad \log\left(\frac{p_j}{1-p_j}\right) = \alpha_j X + \beta_j Z, \tag{9}$$

with $X \sim N(\rho Z^2, 1)$ and $Z \sim N(0,1)$. We implement the 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS and MF-2dFDR-RV, and report the results in Figure 4. In MF-2dFDR-MS, T_j^C is the statistic for testing $\alpha_j = 0$ under the full model (9) and T_j^M is the statistic for testing $\alpha_j = 0$ by forcing $\beta_j = 0$ in (9).

E. *Count response.* We consider the Poisson model

$$Y_j \sim \text{Poisson}(\lambda_j), \quad \log \lambda_j = \alpha_j X + \beta_j Z,$$

with $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$. We implement the 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS and MF-2dFDR-RV, and report the results in Figure 5. Additionally, we consider the negative binomial regression model

$$Y_j \sim \text{Negative Binomial}(\text{size} = 3, \mu_j = e^{f_j(X,Z)})$$

where $f_j(X,Z) = \alpha_j X + \beta_j Z$ for $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$. We implement the 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS and MF-2dFDR-RV, and report the results in Figure 6.

In Section 10.4, we report some additional numerical results under the following scenarios: (1) FWER control, (2) global null, (3) dependent errors, and (4) separating the effects of the densities of the signal of interest and the confounder signal.

7.4 Simulation results

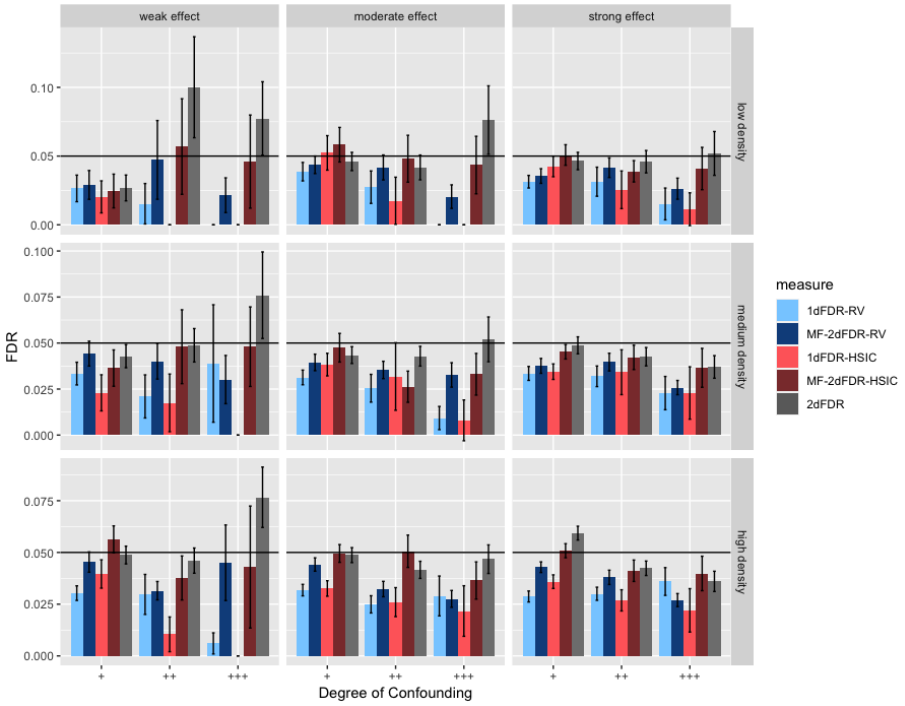
We now discuss the major simulation findings under each scenario. Full simulation results are summarized in the Figures 2-6 and 14-26. Under Scenario A (X and Z are both continuous), when the underlying models between Y and (X, Z), and X and Z are both linear (see Figure 2), all the methods provide tight FDR control except for the 2dFDR which has slight FDR inflation in some instances when the confounding effect is strong. In contrast, the proposed MF-2dFDR-RV, which is equivalent to MF-2dFDR-MS, controls the FDR at the target level across all cases, indicating more robustness of the proposed method than the original 2dFDR. In terms of power, we observe that the power decreases as the confounding effect becomes stronger for all procedures. The 2d procedure is comparable to the 1d counterpart when the confounding effect is weak but is substantially more powerful when the confounding effect is strong. We also observe that MF-2dFDR-RV is comparable to 2dFDR and is more powerful than MF-2dFDR-HSIC. When the underlying model is nonlinear, 2dFDR suffers from severe FDR inflation (see, e.g., Figure 3a). In contrast, MF-2dFDR controls the FDR at the target level across different cases. Among the MF-2dFDR variants, MF-2dFDR-RV delivers the highest power in most cases.

Under Scenario B, where X is discrete while Z is continuous (see, e.g., Figure 16), the empirical FDR is well controlled for MF-2dFDR even when the confounding effect is strong. 2dFDR suffers from moderate FDR inflation (e.g Figure 17a) in some instances, e.g., in the case of strong confounding. Not surprisingly, the 2d procedure is significantly more powerful than the corresponding 1d version. Moreover, the RV-based methods generally make more true rejections compared to the HSIC-based methods.

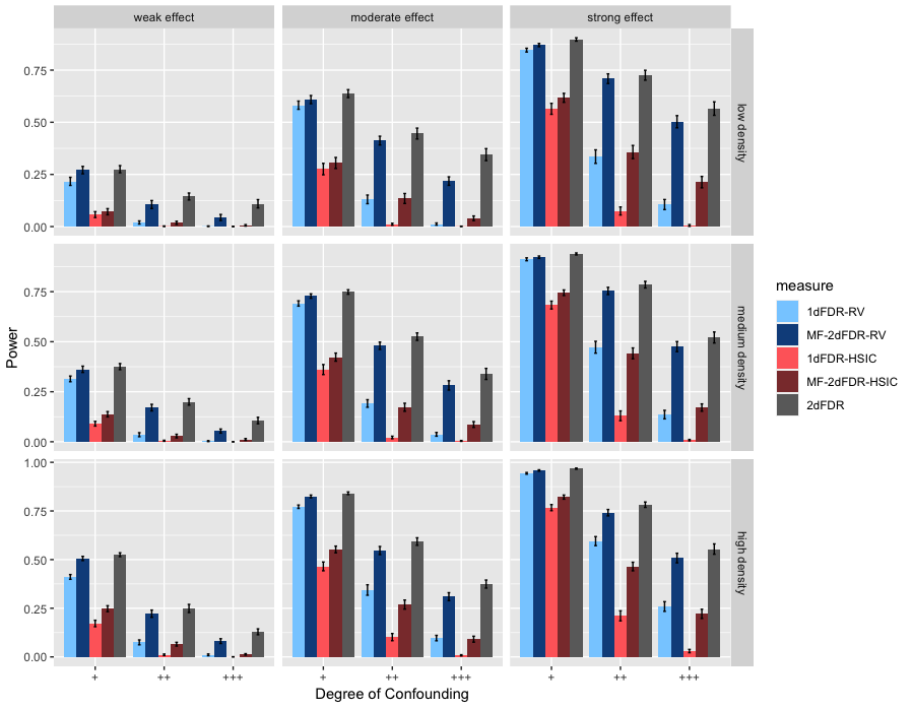
Under Scenario C, where X and Z are both Bernoulli, as seen from Figure 19, all the approaches have empirical FDR under control. When the degree of confounding is high, MF-2dFDR delivers higher power than 1dFDR does.

Under Scenarios D and E, the original 2dFDR is not applicable, and hence only 1dFDR and MF-2dFDR have been compared in the simulations. As seen from Figures 4-6, for Scenarios D-E (binary and count response), all the methods provide reliable FDR control. MF-2dFDR produces significant power improvement over the 1dFDR methods.

To sum up, the proposed MF-2dFDR provides reliable FDR control for all the simulation settings even when the degree of confounding is strong because MF-2dFDR explicitly models the relationship between X and Z. The 2d procedure delivers more rejections compared to the 1d counterpart, and the larger number of rejections typically translates into a higher detection power for the 2d methods. We also see that MF-2dFDR-RV provides the best power in many simulation settings. As the (conditional) RV coefficients are calculated based on spline transformed covariates and confounding factors, MF-2dFDR-RV can capture the nonlinearity between Y and (X, Z) and X and Z in many cases.

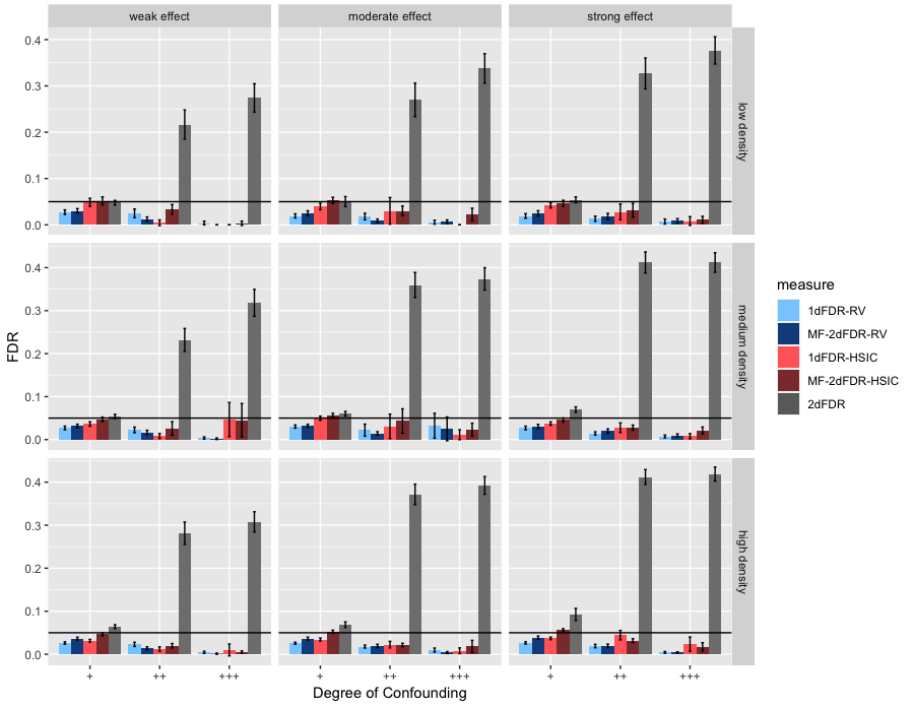


(a) FDR

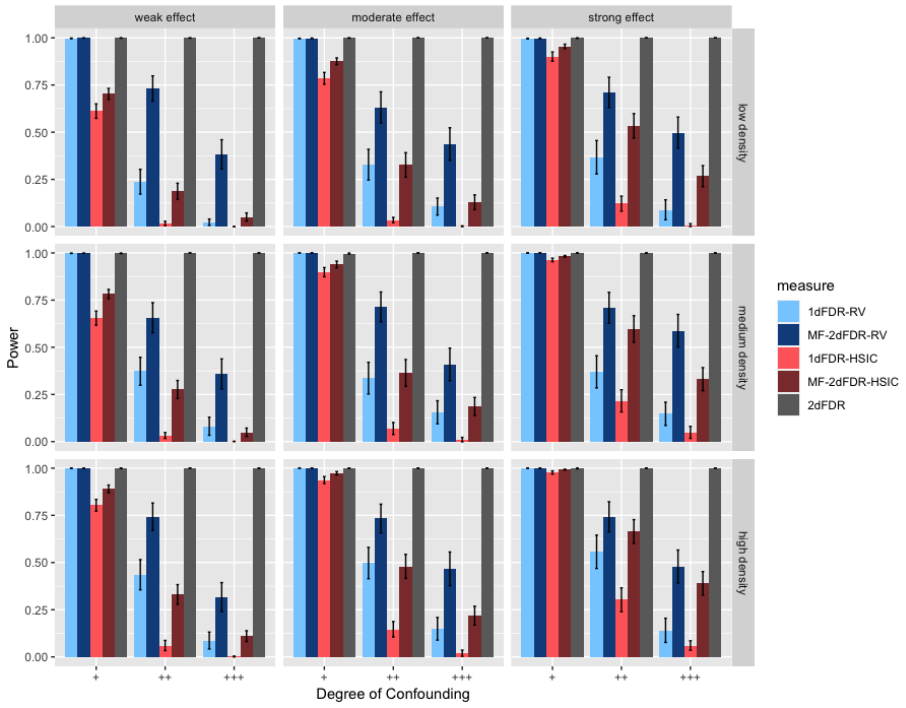


(b) Power

Figure 2: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ and $X \sim N(\rho Z, 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

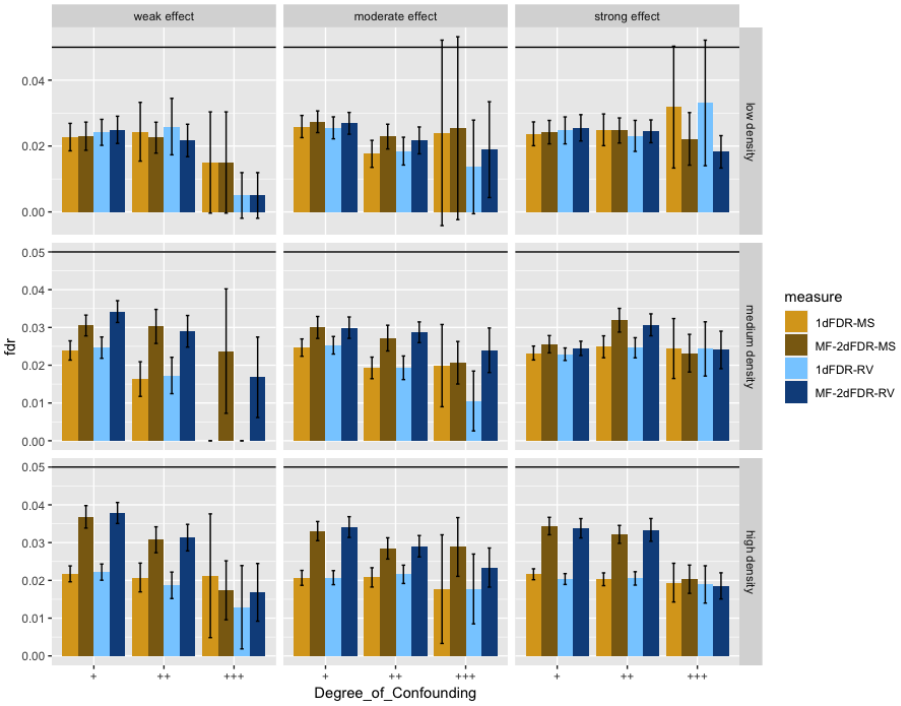


(a) FDR

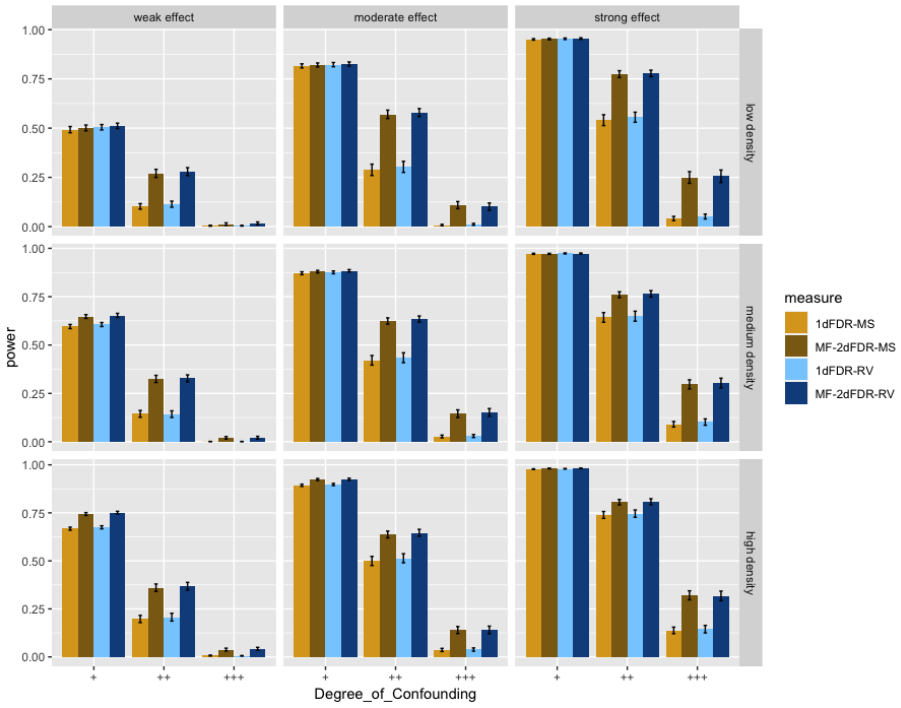


(b) Power

Figure 3: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X^3 + \beta_j e^Z + \epsilon_j$ and $X \sim N(\rho Z^2, 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

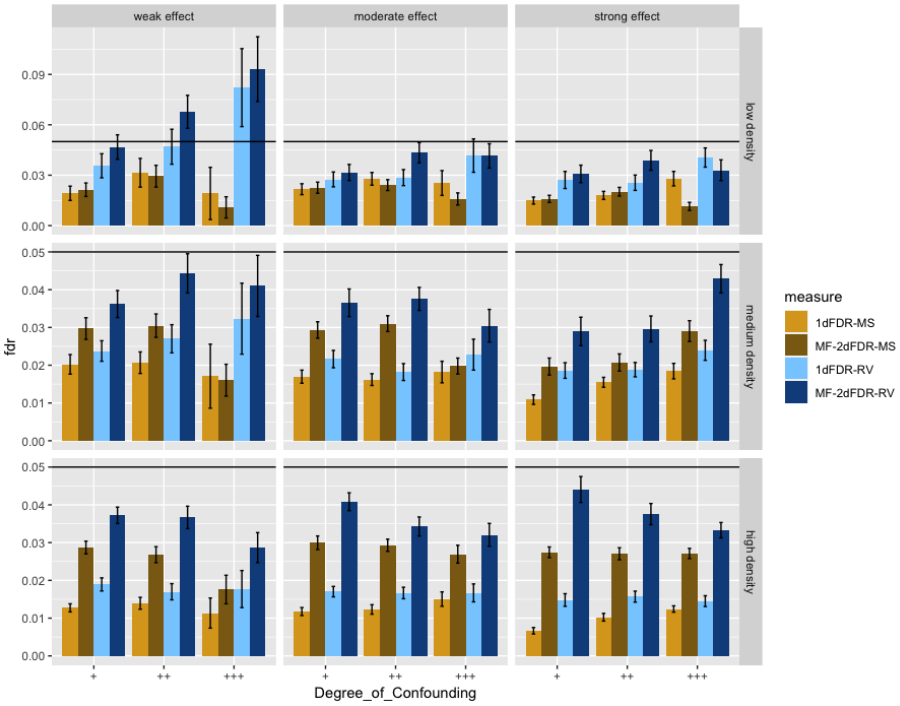


(a) FDR

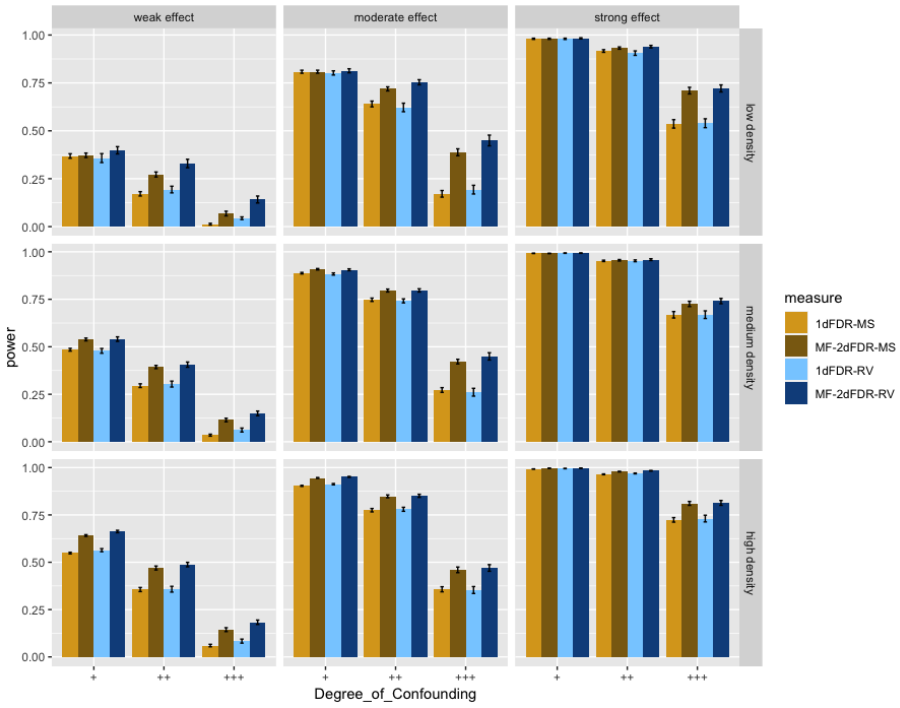


(b) Power

Figure 4: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j \sim \text{Bernoulli}((1+e^{-f_j(X,Z)})^{-1})$, where $f_j(X,Z) = \alpha_j X + \beta_j Z$, $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

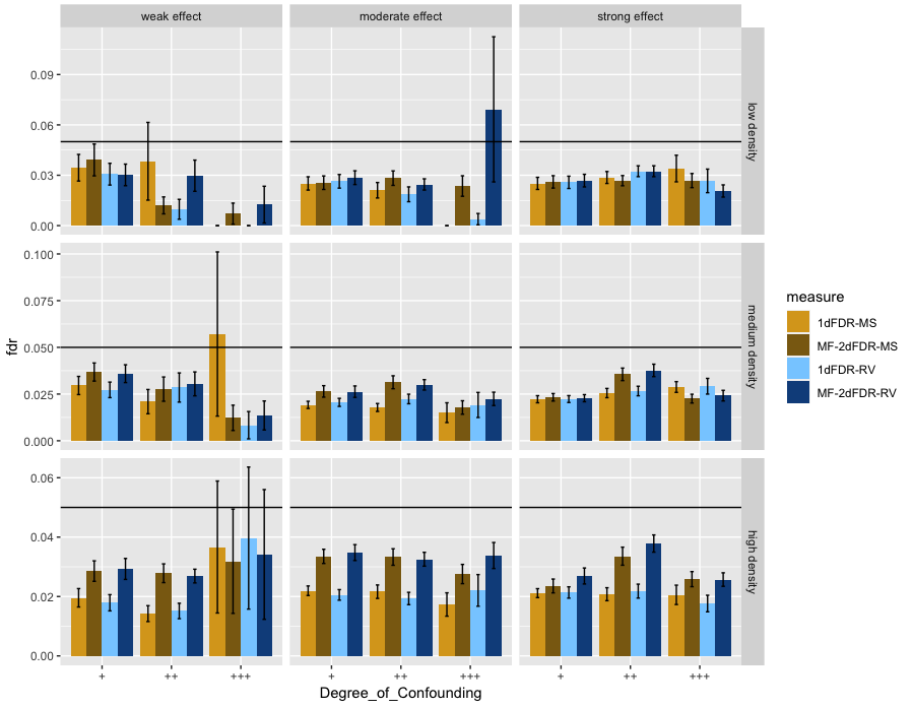


(a) FDR

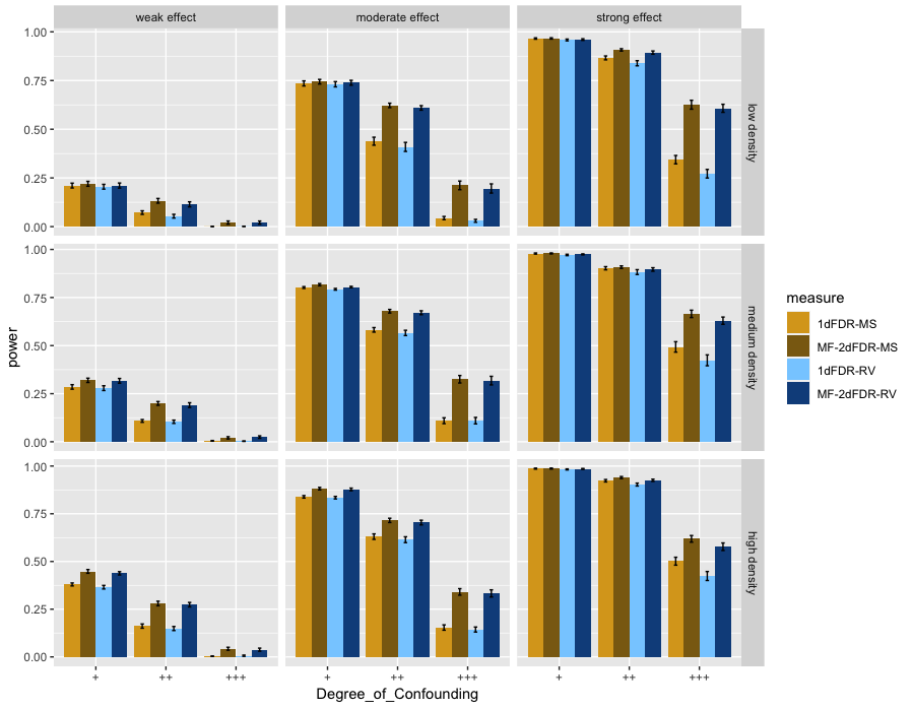


(b) Power

Figure 5: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j \sim \text{Poisson}(\lambda_j)$, where $\log \lambda_j = \alpha_j X + \beta_j Z$ with $X \sim N(\rho Z, 1)$ and $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.



(a) FDR



(b) Power

Figure 6: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j \sim$ Negative Binomial(size $= 3, \mu_j = e^{f_j(X,Z)}$), where $f_j(X,Z) = \alpha_j X + \beta_j Z$, $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

8 Real Data Analysis

8.1 Microbiome data

In the first example, we analyze a microbiome dataset in the R package *GUniFrac*, which is part of a microbiome data set for studying the smoking effect on the upper respiratory tract microbiome (Chen et al., 2021; Charlson et al., 2010). The original data set contains samples from the right and left nasopharynx and oropharynx. Here we use the data from the left oropharynx of 32 nonsmokers and 28 smokers ($n = 60$). The microbiome composition was profiled using 16S rRNA gene-targeted sequencing and processed using the QIIME bioinformatics pipeline (D'Argenio et al., 2014), resulting in a count table recording the frequencies of 856 detected OTUs (operational taxonomic units). Sex is a confounding factor in this data set, with more smokers in males (odds ratio equals 2.3). The aim here is to identify smoking-associated OTUs while adjusting sex. For illustration purposes, the OTU abundances were treated as both continuous and binary outcomes.

A. *Continuous Outcome:* We first filtered out the OTUs occurring in less than 10% of the subjects, which resulted in a total of 174 OTUs. The OTU abundance data were then transformed using a center log-ratio transformation, adding a pseudo-count of 0.5. The numbers of rejections for varying levels of FDR (ranging from 0 to 0.2) were calculated for the following methods: Benjamini-Hochberg (BH, Benjamini and Hochberg (1995b)) procedure, 2dFDR, MF-2dFDR-MS, MF-2dFDR-RV, 1dFDR-MS, 1dFDR-RV. The BH procedure was applied to the p-values corresponding to the tests of significance of the coefficients of IR in a linear regression model with the IR and BMI being the predictors. The numbers of rejections at different FDR levels are shown in Figure 7.

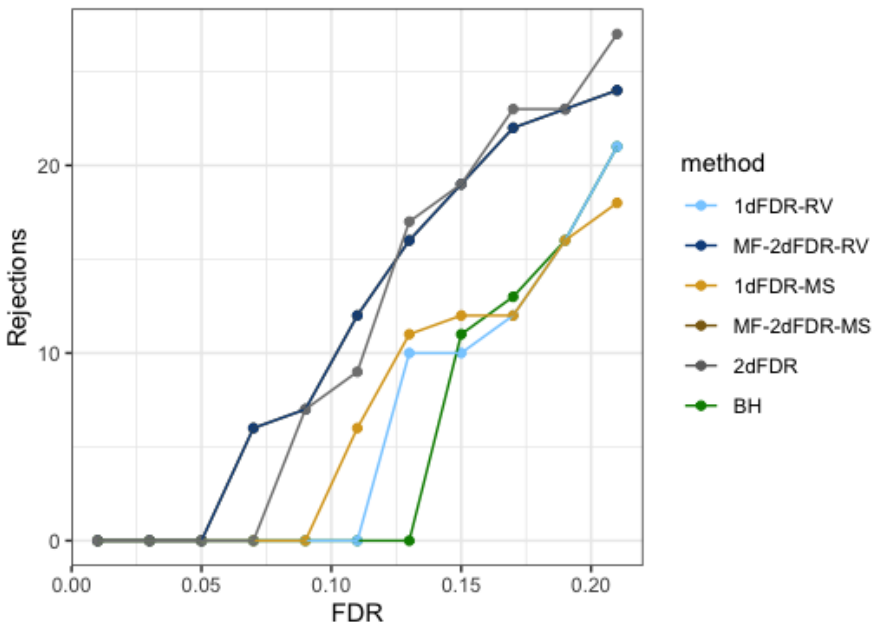


Figure 7: Number of Rejections versus FDR levels for different methods for the Smoking data with continuous outcomes.

The trend is consistent with the simulations, where we have observed that the MF-2dFDR procedure is more powerful than the corresponding 1dFDR procedure and MF-2dFDR-RV makes the highest number of true rejections in most simulation setups. In addition, we produced a Venn diagram (Figure 8) of the rejected features for each method at the FDR level 0.10 to visualize the degree to which the rejected features in various methods overlap. We find that at level 0.1, MF-2dFDR-MS successfully identifies all the seven features identified by the 2dFDR procedure and five additional features.

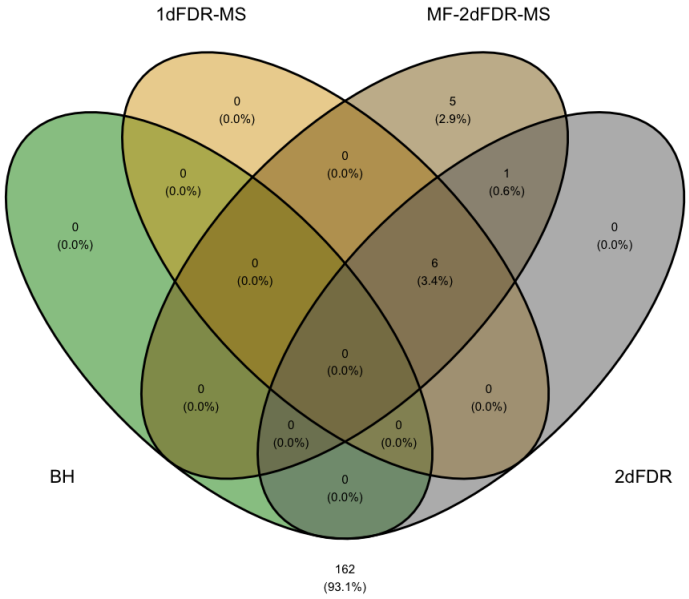


Figure 8: Venn diagram of features identified by different methods for smoking microbiome data

B. *Binary Outcome:* The abundance data of the 174 OTUs tested in Case A were converted into presence/absence data after rarefaction to the minimal sequencing depth (since presence/absence depends on the sequencing depth strongly, rarefaction removes the confounding effect due to sequence depth). Because X, Y, and Z are all categorical (specifically, binary) in this case, for the conditional statistic, i.e., T^C, the Mantel Haenszel statistic was used. For the marginal statistic, i.e., T^M, the Pearson's chi-square statistic was used. Note that the original 2dFDR in Yi et al. (2021) is not applicable in this case as the outcomes are binary. As shown in Figure 9, for all levels of FDR, the BH procedure makes no rejections, and overall, the MF-2dFDR procedure makes a higher number of rejections compared to the corresponding 1dFDR procedure.

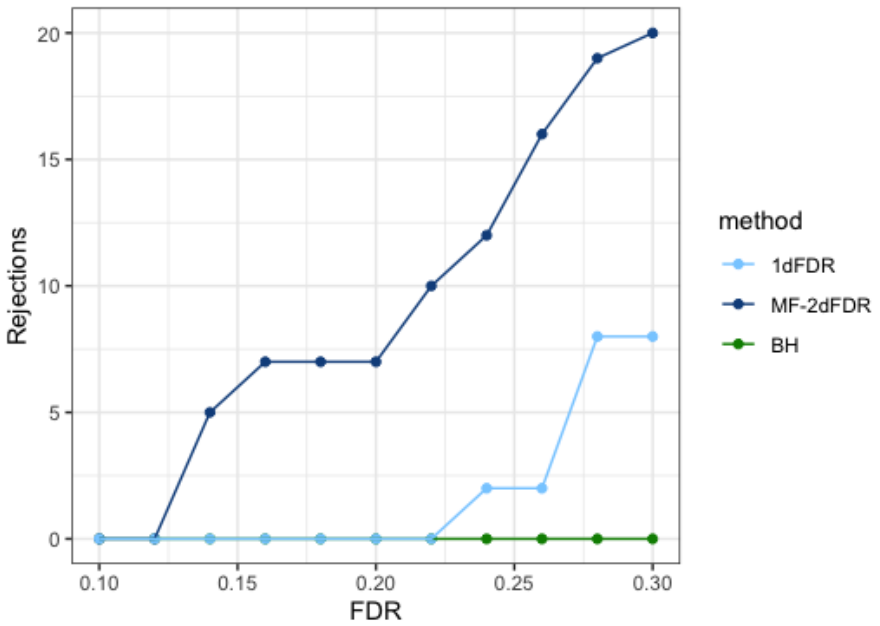


Figure 9: Number of Rejections versus FDR for different methods for smoking microbiome data, where the continuous abundance data were transformed into presence/absence (binary) data.

8.2 Metabolomics data

Next, we consider an Insulin Resistance dataset (Pedersen et al., 2016, 2018) where the goal is to identify serum metabolites associated with insulin resistance (IR) while controlling the effect of the Body Mass Index (BMI) of the individual. 289 non-diabetic Danish adults were recruited for the study, where their IR was estimated by homeostatic model assessment (HOMA-IR) (Matthews et al., 1985). Untargeted metabolome profiles were generated on fasting serum samples, producing measurements on 325 polar metabolites and 876 molecular lipids (collectively called serum metabolites, $m = 1201$). The BMI of a subject is a confounding factor as the IR of a subject is largely influenced by the BMI (correlation coefficient = 0.57). In this example, 2dFDR discovers the largest number of metabolites (481 at 5% FDR), followed by MF-2dFDR-RV (432 metabolites at 5% FDR). Both are a significant improvement over 1dFDR-RV (333), 1dFDR-HSIC (323), and the BH procedure (377). The comparison of the number of rejections versus FDR level for all methods is displayed in Figure 10.

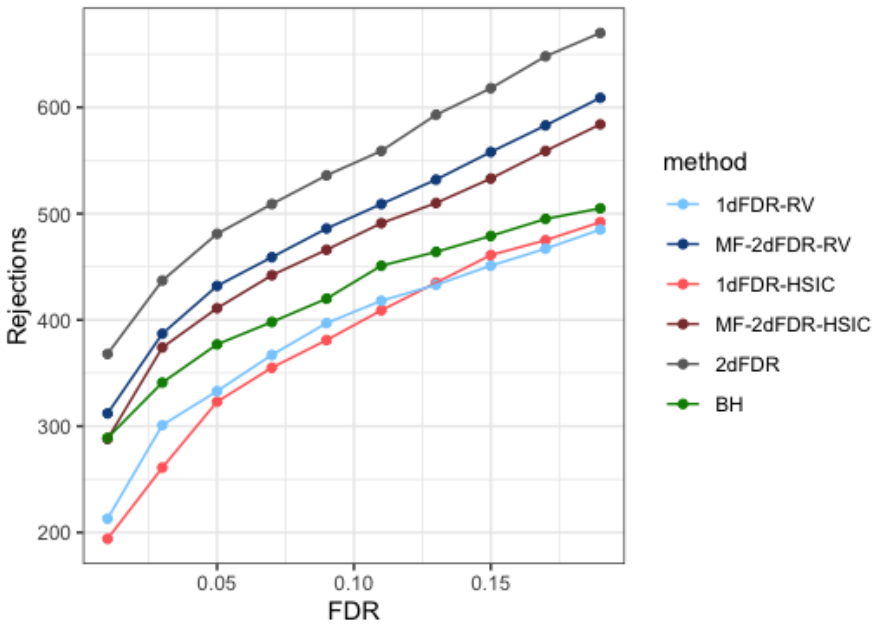


Figure 10: Number of Rejections versus FDR for different methods for the Insulin Resistance dataset

Again, the result generally agrees with the findings from the simulation studies. While 2dFDR is the most powerful in this example, its inflated type I error rate observed in many non-linear simulation setups raises some concern about the reliability of the rejections solely found by itself.

Figure 11 shows the Venn diagram of the serum metabolites detected by the different methods and their degree of overlap at FDR = 0.05 is provided. It is interesting to note that while MF-2dFDR-RV and 2dFDR have detected 403 metabolites in common, the BH procedure has significantly fewer overlapping metabolites with either of these methods.

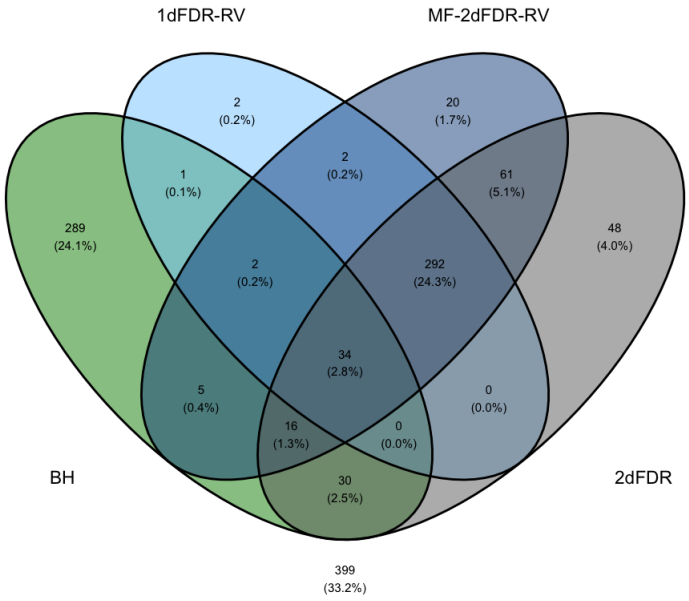


Figure 11: Venn diagram of features identified by different methods for metabolomics data

An additional challenge we faced while analyzing the metabolomics data was generating samples from the conditional distribution $P_{\mathbf{X}|\mathbf{Z}}$. As observed in Figure 12, there is distinct heteroscedasticity in the conditional distribution of IR given BMI. Traditional resampling methods such as residual permutation (Winkler et al., 2014) may fail as homoscedasticity is one of the underlying assumptions. To combat this, the data set was binned into two parts, namely BMI ≤ 26 and BMI > 26, and two separate regressions were fitted to these two subsamples, and the residuals were permuted within each segment. The right panel of Figure 12, which plots the resampled IR versus BMI using binned residual permutation, shows that the heteroscedastic structure has been preserved in the resampled data. The middle panel shows resampling using the traditional residual permutation and we can see that the original shape of the data has not been maintained in this case.

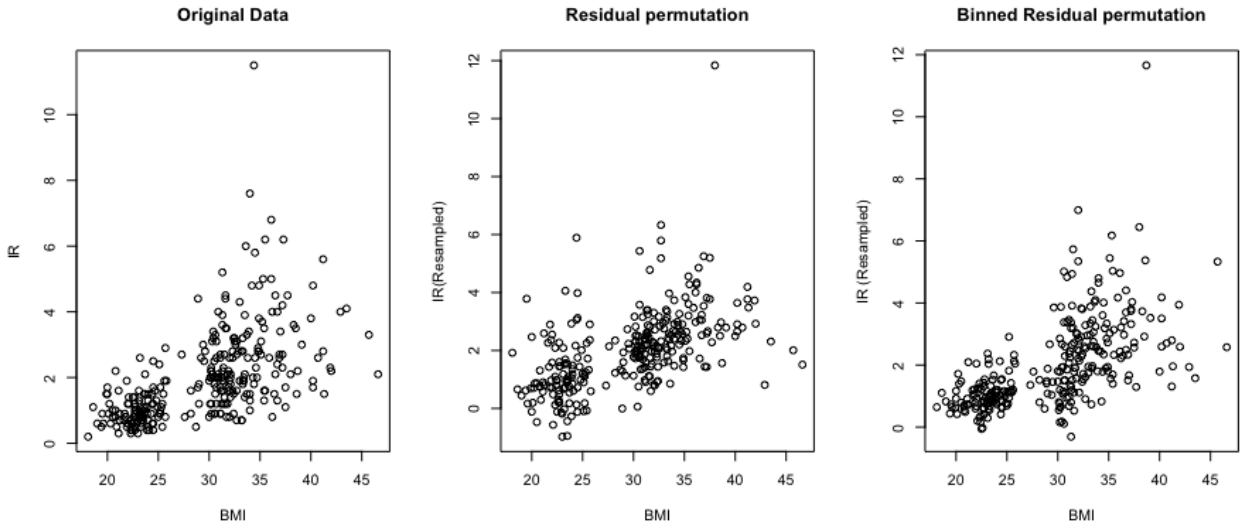


Figure 12: Scatterplots of IR versus BMI for 289 subjects. Left panel: the original data; Middle panel: Resampled data using the traditional residual permutation; Right panel: Resampled data using binned residual permutation

8.3 Gene expression data

Finally, we consider a Pouchitis dataset (Morgan et al., 2015), where the goal is to identify gene expressions associated with patient outcomes in a cohort with ileal pouch-anal anastomosis (IPAA) surgery in the past one year, adjusting for potential confounders such as antibiotics use and sex. This dataset considered a large population of patients having undergone IPAA at Mount Sinai Hospital, Canada. The expression levels of 19,908 genes were measured in two sites, the J-pouch and the pre-pouch ileum (PPI), using the procedure described in (Morgan et al., 2015). We considered the biopsies collected only from the pouch ($n = 74$) in this example. The conditioning variables were sex, smoking status, and antibiotic use in the previous month. The variable of interest is the disease outcome, including FAP (Familial Adenomatous Polyposis), No Pouchitis, Acute Pouchitis, Chronic Pouchitis, and Crohn's Disease like Inflammation. Figure 13 shows the number of genes identified as associated with the disease outcome conditioning on sex, smoking status, and antibiotic usage. At the FDR level of 0.05, the MF-2DFDR identifies the maximum number of genes (2345), followed by

1DFDR-MS (1811) and BH procedure (1640), respectively.

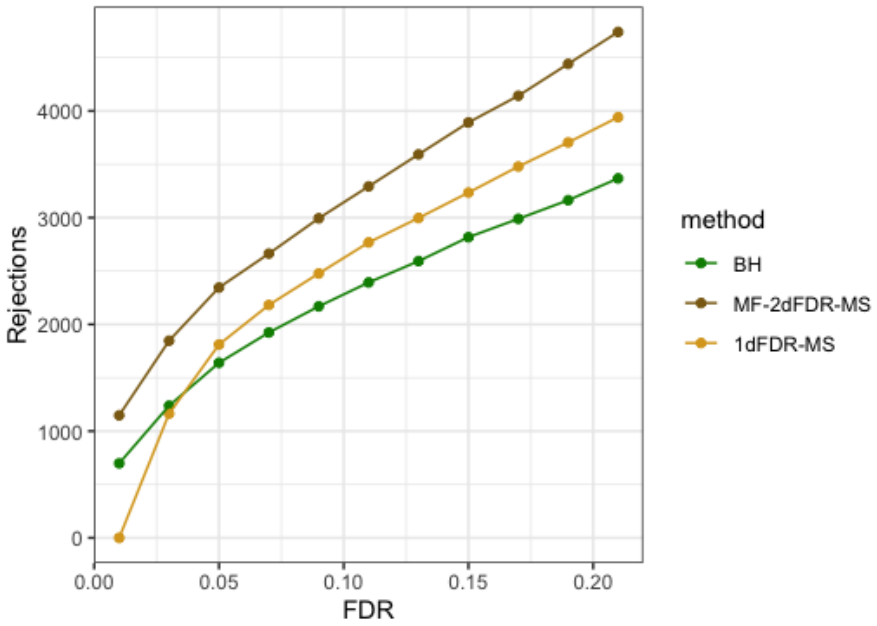


Figure 13: Number of Rejections versus FDR for different methods in the pouchitis gene expression dataset

9 Conclusion

We have proposed a general framework (MF-2dFDR) for performing multiple hypothesis testing while adjusting for confounding effects. Within this new framework, the conditional distribution of the omics features given the variable of interest and confounders can be arbitrary and completely unknown. The framework is flexible by allowing the joint use of any conditional and marginal independence tests, continuous/binary/count/multivariate responses, and various ways of modeling the conditional distribution of the variable of interest given the confounders. As a general methodology, MF-2dFDR can be applied to multiple types of omics data. In view of the numerical results, we recommend the use of MF-2dFDR-RV (based on the spline-transformed variables) under most scenarios due to its robustness and efficiency. When all the three variables (X, Y, Z) are categorical, as in Case B for the smoking microbiome data, the Pearson's chi-square statistic and the Mantel-Haenszel statistic are recommended for testing the marginal and the conditional dependence, respectively.

References

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97,** 10101–10106.

Benjamini, Y. and Hochberg, Y. (1995a). Controlling the false discovery rate: a practical and power-

ful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300.

Benjamini, Y. and Hochberg, Y. (1995b). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300.

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli* **20,** 1006–1028.

Cao, H., Chen, J., and Zhang, X. (2020). Optimal false discovery rate control for large scale multiple testing with auxiliary information. *Unpublished Manuscript. https://web.stat.tamu.edu/ zhangxiany/Order-FDR.pdf* .

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., and Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLOS ONE* **5,** e15216. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0015216&type=printable.

Chen, J., Zhang, X., and Zhou, H. (2021). Gunifrac r package version 1.4.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application.* Number 1. Cambridge university press.

D'Argenio, V., Casaburi, G., Precone, V., and Salvatore, F. (2014). Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *BioMed Research International* **2014,** 1–10.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106,** 1602–1614.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., et al. (2003). Effects of atmospheric ozone on microarray data quality. *Analytical chemistry* **75,** 4672–4675.

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

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* **104,** 1406–1415.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13,** 539–552.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* **11,** 4241–4257.

Gershoni, M. and Pietrokovski, S. (2017). The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC biology* **15,** 1–15.

Glass, D., Viñuela, A., Davies, M. N., Ramasamy, A., Parts, L., Knowles, D., Brown, A. A., Hedman, Å. K., Small, K. S., Buil, A., et al. (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology* **14,** 1–12.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., et al. (2007). A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., and Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics* **14,** 469–490.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11,** 733–739.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3,** e161.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105,** 18718–18723.

Liang, L. and Cookson, W. O. (2014). Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. *Human molecular genetics* **23,** R83–R88.

Lin, D. W., Coleman, I. M., Hawley, S., Huang, C. Y., Dumpit, R., Gifford, D., Kezele, P., Hung, H., Knudsen, B. S., Kristal, A. R., et al. (2006). Influence of surgical manipulation on prostate gene expression: implications for molecular correlates of treatment effects and disease prognosis. *Journal of clinical oncology* **24,** 3763–3770.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B. A. (2004). Gene regulation and dna damage in the ageing human brain. *Nature* **429,** 883–891.

Majewski, I. J. and Bernards, R. (2011). Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nature medicine* **17,** 304–312.

Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F., and Turner, R. C. (1985). Homeostasis model assessment: insulin resistance and ?-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28,** 412–419.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., Stempak, J. M., Gevers, D., Xavier, R. J., Silverberg, M. S., and Huttenhower, C. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology* **16,** 67.

Naaman, M. (2021). On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters* **173,** 109088.

Pedersen, H. K., Forslund, S. K., Gudmundsdottir, V., Østergaard Petersen, A., Hildebrand, F., Hyötyläinen, T., Nielsen, T., Hansen, T., Bork, P., Ehrlich, S. D., Brunak, S., Oresic, M., Pedersen, O., and Nielsen, H. B. (2018). A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. *Nature Protocols* **13,** 2781–2800.

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A. H., Forslund, K., Hildebrand, F., Prifti, E., Falony, G., Chatelier, E. L., Levenez, F., Doré, J., Mattila, I., Plichta, D. R., Pöhö, P., Hellgren, L. I., Arumugam, M., Sunagawa, S., Vieira-Silva, S., Jørgensen, T., Holm, J. B., Trošt, K., Consortium, M., Kristiansen, K., Brix, S., Raes, J., Wang, J., Hansen, T., Bork, P., Brunak, S., Oresic, M., Ehrlich, S. D., and Pedersen, O. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535,** 376–381.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research* **12,**.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 479–498.

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

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics* **35,** 2769–2794.

Walsh, D. M., Hokenstad, A. N., Chen, J., Sung, J., Jenkins, G. D., Chia, N., Nelson, H., Mariani, A., and Walther-Antonio, M. R. (2019). Postmenopause as a key factor in the composition of the endometrial cancer microbiome (ecbiome). *Scientific reports* **9,** 1–16.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics* **45,** 1863.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage* **92,** 381–397.

Yi, S., Zhang, X., Yang, L., Huang, J., Liu, Y., Wang, C., Schaid, D. J., and Chen, J. (2021). 2dfdr: a new approach to confounder adjustment substantially increases detection power in omics association studies. *Genome biology* **22,** 1–18.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775* .

Zhou, X., Jiao, Y., Liu, J., and Huang, J. (2022). A deep generative approach to conditional sampling. *Journal of the American Statistical Association* pages 1–12.

Ziegler, A., Koch, A., Krockenberger, K., and Großhennig, A. (2012). Personalized medicine using dna biomarkers: a review. *Human genetics* **131,** 1627–1638.

10 Appendix

10.1 More on Assumption 2

We provide further discussions on Assumption 2 and justify it under model (4) with

$$u_j(\mathbf{X}) = \sum_{k=1}^{J_1} \alpha_{k,j} B_{k,\mathbf{X}}(\mathbf{X}), \quad v_j(\mathbf{Z}) = \sum_{k=1}^{J_2} \beta_{k,j} B_{k,\mathbf{Z}}(\mathbf{Z}), \quad \boldsymbol{\epsilon}_j = (\epsilon_{1,j}, \ldots, \epsilon_{n,j})^\top \sim N(0, \sigma_j^2 \mathbf{I}),$$

where $B_{k,\mathbf{X}}(\cdot) : \mathbb{R}^p \to \mathbb{R}$ and $B_{k,\mathbf{Z}}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ are some known basis functions. Define $\mathbf{B}_\mathbf{X} = (B_{k,\mathbf{X}}(\mathbf{X}_i))_{1\le i\le n, 1\le k\le J_1} \in \mathbb{R}^{n\times J_1}$ and $\mathbf{B}_\mathbf{Z} = (B_{k,\mathbf{Z}}(\mathbf{Z}_i))_{1\le i\le n, 1\le k\le J_2} \in \mathbb{R}^{n\times J_2}$. Let $\mathbf{P}_\mathbf{Z}^\perp$ be the orthogonal projection onto the column space of $\mathbf{B}_\mathbf{Z}$. We consider the statistics

$$\begin{aligned} T_j^M &= \hat{\sigma}_j^{-2} \|(\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} \mathbf{B}_\mathbf{X}^\top \widetilde{\mathbf{Y}}_j\|^2, \\ T_j^C &= \hat{\sigma}_j^{-2} \|(\mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X})^{-1/2} \mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \widetilde{\mathbf{Y}}_j\|^2, \end{aligned}$$

where $\hat{\sigma}_j^2$ is a consistent variance estimator of σ_j^2 such that $\hat{\sigma}_j^2 \to^p \sigma_j^2$. Conditional on $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})$, $(\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} \mathbf{B}_\mathbf{X}^\top \widetilde{\mathbf{Y}}_j$ and $(\mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X})^{-1/2} \mathbf{B}_\mathbf{X} \mathbf{P}_\mathbf{Z}^\perp \widetilde{\mathbf{Y}}_j$ jointly follow the multivariate normal distribution with the mean

$$\begin{pmatrix} (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{1/2} \boldsymbol{\alpha}_j + (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{Z}) \boldsymbol{\beta}_j \\ (\mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X})^{1/2} \boldsymbol{\alpha}_j \end{pmatrix}$$

and the covariance matrix

$$\sigma_j^2 \begin{pmatrix} \mathbf{I} & (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} (\mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X})^{1/2} \\ (\mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X})^{1/2} (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} & \mathbf{I} \end{pmatrix}.$$

Define $\mathbf{\Sigma}_\mathbf{X} = \mathrm{cov}(\widetilde{\mathbf{B}}_\mathbf{X})$, $\mathbf{\Sigma}_{\mathbf{XZ}} = \mathrm{cov}(\widetilde{\mathbf{B}}_\mathbf{X}, \widetilde{\mathbf{B}}_\mathbf{Z})$ and $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Z}} = \mathbf{\Sigma}_\mathbf{X} - \mathbf{\Sigma}_{\mathbf{XZ}} \mathbf{\Sigma}_\mathbf{Z}^{-1} \mathbf{\Sigma}_{\mathbf{ZX}}$, where $\widetilde{\mathbf{B}}_\mathbf{X} = (B_{1,\mathbf{X}}(\mathbf{X}), \ldots, B_{J_1,\mathbf{X}}(\mathbf{X}))^\top$ and $\widetilde{\mathbf{B}}_\mathbf{Z} = (B_{1,\mathbf{Z}}(\mathbf{Z}), \ldots, B_{J_2,\mathbf{X}}(\mathbf{Z}))^\top$. By the law of large numbers, we have

$$\begin{aligned} & n^{-1} \mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X} \to^p \mathbf{\Sigma}_\mathbf{X}, \\ & n^{-1/2} (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{X})^{-1/2} (\mathbf{B}_\mathbf{X}^\top \mathbf{B}_\mathbf{Z}) \to^p \mathbf{\Sigma}_\mathbf{X}^{-1/2} \mathbf{\Sigma}_{\mathbf{XZ}}, \\ & n^{-1} \mathbf{B}_\mathbf{X}^\top \mathbf{P}_\mathbf{Z}^\perp \mathbf{B}_\mathbf{X} \to^p \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Z}}. \end{aligned}$$

In this case, we have

$$\begin{aligned} \widetilde{V}(t_1, t_2) &= \lim_{m,n \to +\infty} \frac{1}{m_0} \sum_{j \in \mathcal{M}_0} F(t_1, t_2; 0, \sqrt{n}\boldsymbol{\beta}_j / \sigma_j), \\ \widetilde{S}(t_1, t_2) &= \lim_{m,n \to +\infty} \frac{1}{m} \sum_{j=1}^{m} F(t_1, t_2; \sqrt{n}\boldsymbol{\alpha}_j / \sigma_j, \sqrt{n}\boldsymbol{\beta}_j / \sigma_j), \end{aligned}$$

with $F(t_1, t_2; \mathbf{a}, \mathbf{b}) = P(\|\mathbf{V}_{1,j}\|^2 > t_1, \|\mathbf{V}_{2,j}\|^2 > t_2)$, where $(\mathbf{V}_{1,j}, \mathbf{V}_{2,j})$ follow the multivariate normal distribution with the mean

$$\begin{pmatrix} \mathbf{\Sigma}_\mathbf{X}^{1/2} \mathbf{a} + \mathbf{\Sigma}_\mathbf{X}^{-1/2} \mathbf{\Sigma}_{\mathbf{XZ}} \mathbf{b} \\ \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Z}}^{1/2} \mathbf{a} \end{pmatrix}$$

and the covariance matrix

$$\begin{pmatrix} \mathbf{I} & \mathbf{\Sigma}_\mathbf{X}^{-1/2} \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Z}}^{1/2} \\ \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Z}}^{1/2} \mathbf{\Sigma}_\mathbf{X}^{-1/2} & \mathbf{I} \end{pmatrix}.$$

If $(\sqrt{n}\boldsymbol{\alpha}_j / \sigma_j, \sqrt{n}\boldsymbol{\beta}_j / \sigma_j)$ follows some distribution $\mathcal{F}$ independently across j and conditional on $\boldsymbol{\alpha}_j = 0$, $\sqrt{n}\boldsymbol{\beta}_j / \sigma_j$ follows the distribution $\mathcal{F}_0$ independently for $j \in \mathcal{M}_0$, then we have

$$\begin{aligned} \widetilde{V}(t_1, t_2) &= \int F(t_1, t_2, (0, \boldsymbol{\beta})) d\mathcal{F}_0(\boldsymbol{\beta}), \\ \widetilde{S}(t_1, t_2) &= \int F(t_1, t_2, (\boldsymbol{\alpha}, \boldsymbol{\beta})) d\mathcal{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned}$$

10.2 Technical details

We first present the following result which was recently proved in Naaman (2021).

Lemma 1 (Dvoretzky–Kiefer–Wolfowitz inequality)**.** *Let $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n$ be independent d-dimensional random vectors with the distribution function $F(\mathbf{t}) = P(\boldsymbol{\xi}_i \leq \mathbf{t})$, where $\boldsymbol{\xi}_i \leq \mathbf{t}$ means that $\xi_{ij} \leq t_j$*

for $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{id})$ *and* $1 \leq j \leq d$. *Denote the standard empirical distribution function by* $F_n(\mathbf{t}) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{\boldsymbol{\xi}_i \leq \mathbf{t}\}$. *Then we have*

$$P\left(\sup_{\mathbf{t}\in\mathbb{R}^d} |F_n(\mathbf{t}) - F(\mathbf{t})| > \epsilon\right) \leq d(n+1)\exp\left(-2n\epsilon^2\right).$$

We now present the proof of the main theoretical result.

Proof of Theorem 1. Define the filtration

$$\mathcal{F}_s = \sigma\left(\left\{\mathbf{1}\{T_{j,b}^M \geq t_1(a)\}, \mathbf{1}\{T_{j,b}^C \geq t_2(a)\}\right\}_{1\leq j\leq m, 0\leq b\leq B} : 1 \leq a \leq s\right)$$

for $1 \leq s \leq \mathcal{S}$ and the process $U(s) = \widetilde{V}^0(s)/\{\sum_{b=0}^{B} \widetilde{V}^b(s)\}$, which is adapted to the filtration $\mathcal{F}_s$. The conditional distribution of $\widetilde{V}^b(t)$ given the sigma-field $\sigma(\{\mathbf{1}\{T_{j,b}^M \geq t_1(a)\}, \mathbf{1}\{T_{j,b}^M \geq t_2(a)\}\}_{1\leq j\leq m} : 1 \leq a \leq s)$ with $s < t$ are the same across all $b = 0, 1, \ldots, B$. Thus by the symmetry, we must have for $s < t$, $E[U(t)|\mathcal{F}_s] = (B+1)^{-1}$. Thus $U(t) - 1/(B+1)$ is a martingale difference sequence. Also, we have $\{s^* \leq t\} \in \mathcal{F}_t$. Therefore, s^* is a stopping time. By the optional stopping time theorem,

$$E[U(s^*)] = \frac{1}{B+1}. \tag{10}$$

Recall from the definition of s^* that

$$\frac{(B+1)^{-1}\sum_{b=0}^{B} V^b(s^*)}{1 \vee V^0(s^*)} \leq q. \tag{11}$$

Using (10) and (11), we obtain

$$E\left[\frac{\widetilde{V}^0(s^*)}{1 \vee V^0(s^*)}\right] \leq (B+1)qE\left[\frac{\widetilde{V}^0(s^*)}{\sum_{b=0}^{B} V^b(s^*)}\right] = (B+1)qE\left[U(s^*)\right] = q.$$

□

Proof of Theorem 2. For $(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+$, define the following processes

$$\begin{aligned}
S_{n,m}(t_1, t_2) &= m^{-1} \sum_{j=1}^{m} \mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}, \\
V_{n,m}(t_1, t_2) &= m_0^{-1} \sum_{j\in\mathcal{M}_0} \mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}, \\
Q_{n,m}(t_1, t_2) &= m_0^{-1} \sum_{j\in\mathcal{M}_0} P_0(T_j^M \geq t_1, T_j^C \geq t_2 | \widetilde{\mathbf{Y}}_j, \widetilde{\mathbf{Z}}).
\end{aligned}$$

We divide the proof into two steps. In Step 1, we obtain some uniform convergence results while in Step 2, we apply these results to show the FDR control.

Step 1. Conditional on $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})$, $\mathbf{1}\{T_j^M \geq t_1, T_j^C \geq t_2\}$ are independent across $j \in \mathcal{M}_0$. By Lemma 1, we have

$$\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}\left[\mathbf{1}\{T_j^M\ge t_1,T_j^C\ge t_2\}-P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{X},\widetilde{\mathbf{Z}})\right]\right|\to^p 0. \tag{12}$$

By Assumption 1, conditional on $\widetilde{\mathbf{Z}}$ and for any fixed t_1 and t_2, $P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Y}}_j,\widetilde{\mathbf{Z}})$ are independent across $j\in\mathcal{M}_0$. Therefore, by the law of large numbers,

$$\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}\left\{P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Y}}_j,\widetilde{\mathbf{Z}})-P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Z}})\right\}\to^p 0.$$

Following the proof of the Glivenko-Cantelli Theorem, we can strengthen the point-wise convergence to the uniform convergence, i.e.,

$$\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}\left\{P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Y}}_j,\widetilde{\mathbf{Z}})-P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Z}})\right\}\right|\to^p 0. \tag{13}$$

Similarly, the result in Assumption 2 can also be strengthened to the uniform convergence, i.e.,

$$\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{X},\widetilde{\mathbf{Z}})-\widetilde{V}(t_1,t_2)\right|\to^p 0. \tag{14}$$

It implies that

$$\begin{aligned}
&\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Z}})-\widetilde{V}(t_1,t_2)\right|\\
=&\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\mathbb{E}\left[\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}})-\widetilde{V}(t_1,t_2)\middle|\widetilde{\mathbf{Z}}\right]\right| \qquad (15)\\
\le&\mathbb{E}\left[\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|\frac{1}{m_0}\sum_{j\in\mathcal{M}_0}P(T_j^M\ge t_1,T_j^C\ge t_2|\widetilde{\mathbf{Z}})-\widetilde{V}(t_1,t_2)\right|\middle|\widetilde{\mathbf{Z}}\right]\to^p 0,
\end{aligned}$$

by Lebesgue's dominated convergence theorem. Combining (12), (13), (14) and (15) together, we get

$$\begin{aligned}
\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|V_{n,m}(t_1,t_2)-\widetilde{V}(t_1,t_2)\right|\to^p 0,\\
\sup_{t_1\le t_{0,1},t_2\le t_{0,2}}\left|Q_{n,m}(t_1,t_2)-\widetilde{V}(t_1,t_2)\right|\to^p 0.
\end{aligned}$$

Using similar arguments by conditioning on $(\widetilde{\mathbf{X}}, \widetilde{Z})$, we can show that

$$\sup_{t_1 \le t_{0,1}, t_2 \le t_{0,2}} \left| S_{n,m}(t_1, t_2) - \widetilde{S}(t_1, t_2) \right| \to^p 0.$$

Following the arguments in the proof of Lemma 8.2 of Cao et al. (2020), we have under Assumptions 1-3 that

$$\sup_{t_1 \le t_{0,1}, t_2 \le t_{0,2}} \left| \mathrm{FDP}(t_1, t_2) - \frac{\pi_0 \widetilde{V}(t_1, t_2)}{\widetilde{S}(t_1, t_2)} \right| \to^p 0,$$
$$\sup_{t_1 \le t_{0,1}, t_2 \le t_{0,2}} \left| \frac{m_0 Q_{n,m}(t_1, t_2)}{1 \vee m S_{n,m}(t_1, t_2)} - \frac{\pi_0 \widetilde{V}(t_1, t_2)}{\widetilde{S}(t_1, t_2)} \right| \to^p 0.$$

In view of Assumption 3, the above convergence implies that

$$\begin{aligned} & P\left(\mathrm{FDP}_{\mathrm{oracle}}(t_{0,1}, 0) < q, \mathrm{FDP}_{\mathrm{oracle}}(0, t_{0,2}) < q\right) \\ \ge & P\left(\frac{m_0 Q_{n,m}(t_{0,1}, 0)}{1 \vee m S_{n,m}(t_{0,1}, 0)} < q, \frac{m_0 Q_{n,m}(0, t_{0,2})}{1 \vee m S_{n,m}(0, t_{0,2})} < q \right) \to 1, \end{aligned}$$

and thus we must have

$$P(t_1^* \le t_{0,1}, t_2^* \le t_{0,2}) \to 1.$$

Step 2. On the event $t_1^* \le t_{0,1}$ and $t_2^* \le t_{0,2}$ which has probability converging to one, we have

$$\begin{aligned} & \mathrm{FDP}(t_1^*, t_2^*) - \mathrm{FDP}_{\mathrm{oracle}}(t_1^*, t_2^*) \\ \le & \mathrm{FDP}(t_1^*, t_2^*) - \frac{m_0 Q_{n,m}(t_1^*, t_2^*)}{1 \vee m S_{n,m}(t_1^*, t_2^*)} \\ = & \mathrm{FDP}(t_1^*, t_2^*) - \frac{\pi_0 \widetilde{V}(t_1^*, t_2^*)}{\widetilde{S}(t_1^*, t_2^*)} + \frac{\pi_0 \widetilde{V}(t_1^*, t_2^*)}{\widetilde{S}(t_1^*, t_2^*)} - \frac{m_0 Q_{n,m}(t_1^*, t_2^*)}{1 \vee m S_{n,m}(t_1^*, t_2^*)} \\ \le & \sup_{t_1 \le t_{0,1}, t_2 \le t_{0,2}} \left| \mathrm{FDP}(t_1, t_2) - \frac{\pi_0 \widetilde{V}(t_1, t_2)}{\widetilde{S}(t_1, t_2)} \right| \\ & + \sup_{t_1 \le t_{0,1}, t_2 \le t_{0,2}} \left| \frac{m_0 Q_{n,m}(t_1, t_2)}{1 \vee m S_{n,m}(t_1, t_2)} - \frac{\pi_0 \widetilde{V}(t_1, t_2)}{\widetilde{S}(t_1, t_2)} \right| = o_p(1). \end{aligned}$$

Thus we have

$$\mathrm{FDP}(t_1^*, t_2^*) \le \mathrm{FDP}_{\mathrm{oracle}}(t_1^*, t_2^*) + o_p(1) = q + o_p(1).$$

By Lemma 8.3 of Cao et al. (2020), we have

$$\limsup_{n,m \to +\infty} \mathbb{E}\left[\mathrm{FDP}(t_1^*, t_2^*)\right] \le q.$$

□

10.3 DGPs in the simulation studies

We provide the specific data generating processes (DGPs) considered in Section 7.3.

1. $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ and $X \sim N(\rho Z, 1)$, where $Z \sim N(0,1)$;
2. $Y_j = \alpha_j X^3 + \beta_j e^Z + \epsilon_j$ and $X \sim N(\rho Z^2, 1)$, where $Z \sim N(0,1)$;
3. $Y_j = \alpha_j X^3 + \beta_j Z^3 + \epsilon_j$ and $X \sim N(\rho(Z + Z^2), 1)$, where $Z \sim N(0,1)$;
4. $Y_j = \alpha_j(X + |X^3|) + \beta_j e^Z + \epsilon$ and $X \sim N(\rho(Z + Z^2), 1)$, where $Z \sim N(0,1)$;
5. $Y_j = \alpha_j e^X + \beta_j Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim N(0,1)$;
6. $Y_j = \alpha_j e^X + \beta_j e^Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim N(0,1)$;
7. $Y_j = \alpha_j e^X + \beta_j Z^2 + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim N(0,1)$;
8. $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim \text{Bernoulli}(0.7)$;
9. $Y_j \sim \text{Bernoulli}((1+e^{-f_j(X,Z)})^{-1})$, where $f_j(X,Z) = \alpha_j X + \beta_j Z$, $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$;
10. $Y_j \sim \text{Poisson}(\lambda_j)$, where $\log \lambda_j = \alpha_j X + \beta_j Z$ with $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$;
11. $Y_j \sim \text{Negative Binomial}(\text{size} = 3, \mu_j = e^{f_j(X,Z)})$, where $f_j(X,Z) = \alpha_j X + \beta_j Z$, $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$;

10.4 Additional simulation results

- FWER control: We investigate the finite sample performance of MF-2dFWER and its corresponding 1d version. In Figures 25-26, we report the empirical FWER and power of MF-2dFWER and 1dFWER for both the linear and nonlinear models. In either case, the empirical FWER is well controlled for both methods. The 2d procedure again produces higher power than the 1d version, especially for stronger confounders.

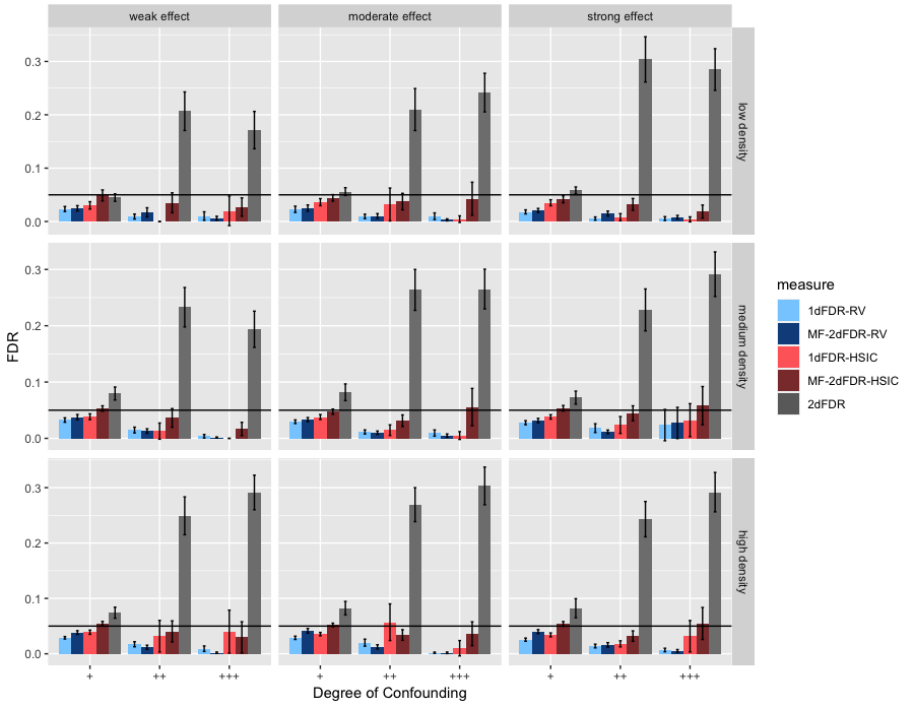
- Global null: We examine the performance of 2dFDR, MF-2dFDR-RV, MF-2dFDR-HSIC, 1dFDR-RV and 1dFDR-HSIC under the global null. Specifically, we consider the model $Y_j = \beta_j Z$, where $X \sim N(\rho Z, 1)$ and $Z \sim N(0,1)$. None of the methods produced any rejections for all degrees of confounding.

- Dependent errors: To evaluate the impact of dependence on the methods' performance, we consider the model: $Y_j = \alpha_j e^X + \beta_j e^Z + \epsilon_j$ where $\epsilon_j = 0.7\epsilon_{j-1} + e_j$ and $X \sim N(\rho(Z + Z^2), 1)$ with $Z \sim N(0,1)$ and $\{e_j\}_{j=1}^m$ being a white noise process. The results are summarized in Figure 20. Overall, MF-2dFDR is robust to the AR(1) type dependence with reliable FDR control and reasonable power.

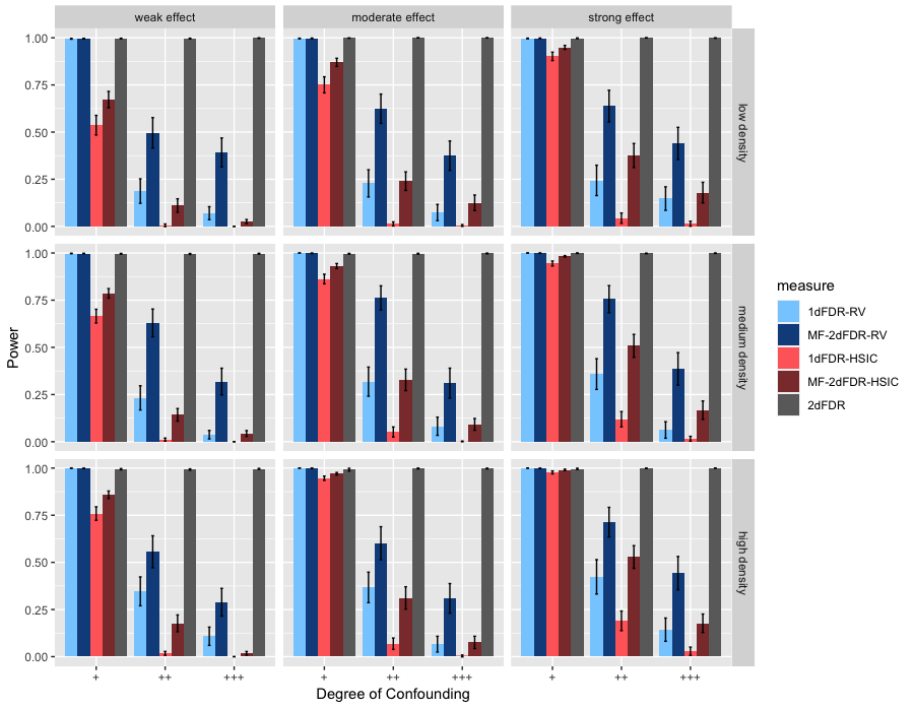
- Separating the effects of densities of the signal of interest and the confounder signal: In all preceding simulations, the density of the signal of interest and the confounding signal had been kept at the same level—weak, moderate or strong. In this simulation setup, we attempt to tease apart the effects of the two types of signals.

    1. First, we fix the density of the signal of interest at the 10% level and vary the density of the confounding signal through weak, moderate, and strong. The associated plots are given in Figure 21 and Figure 23, corresponding to a linear and non-linear DGPs respectively.
    2. Next, we fix the density of the confounding signal to 10% and vary the density of the signal of interest through weak, moderate, and strong. The associated plots are in Figure 22 and Figure 24, corresponding to a linear and non-linear DGPs respectively.

    In both the linear and the non-linear DGPs, we find that varying the density of the signal of interest while keeping the density of the confounding variable constant is displaying a starker difference (increase) in the power as the densities are increased.

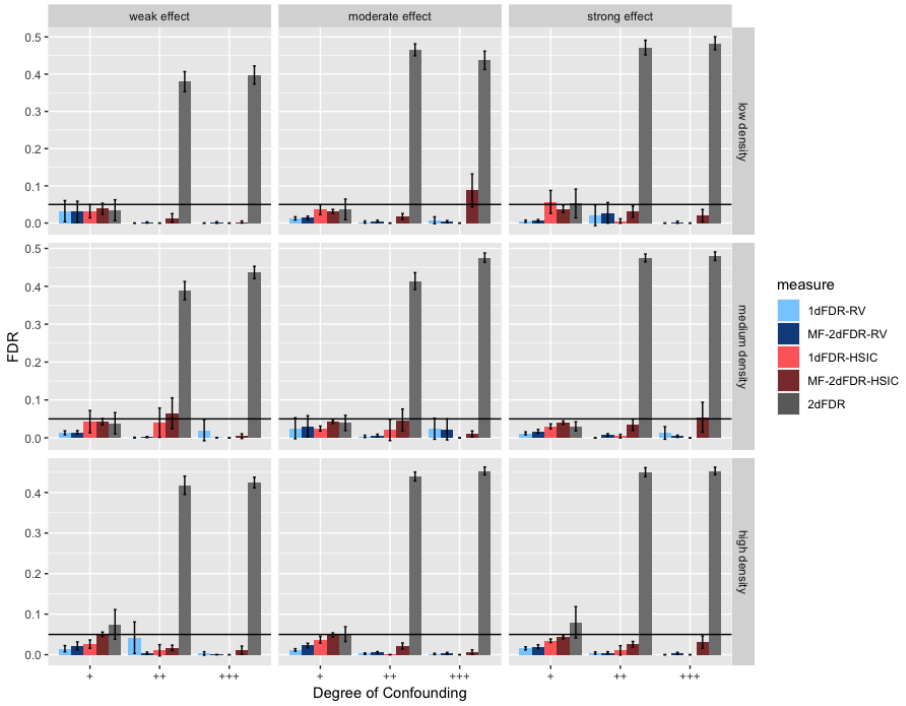


(a) FDR

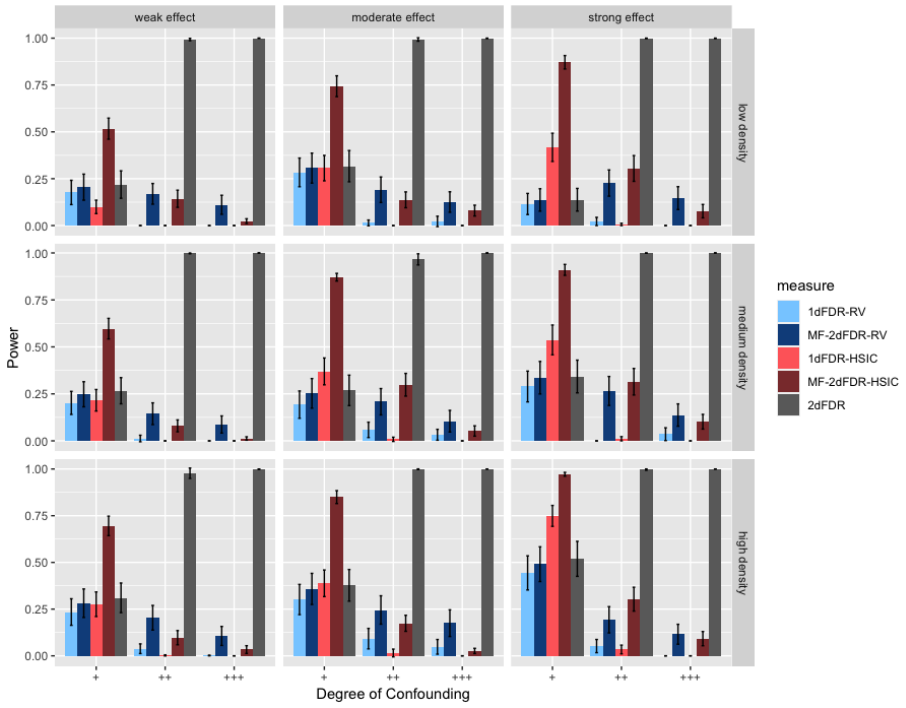


(b) Power

Figure 14: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X^3 + \beta_j Z^3 + \epsilon_j$ and $X \sim N(\rho(Z + Z^2), 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

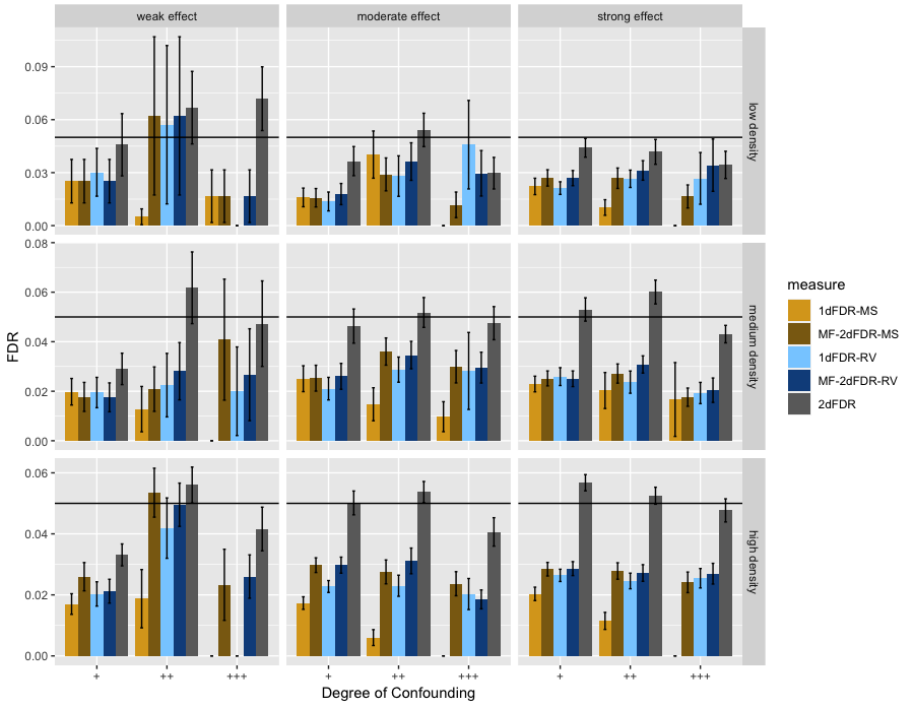


(a) FDR

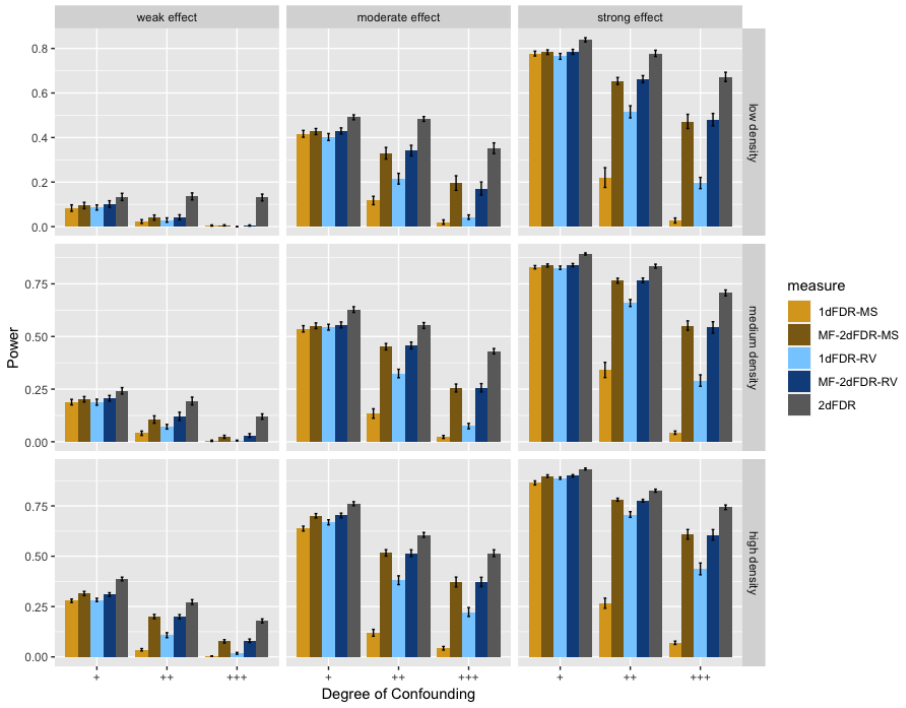


(b) Power

Figure 15: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j(X + |X^3|) + \beta_j e^Z + \epsilon$ and $X \sim N(\rho(Z + Z^2), 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

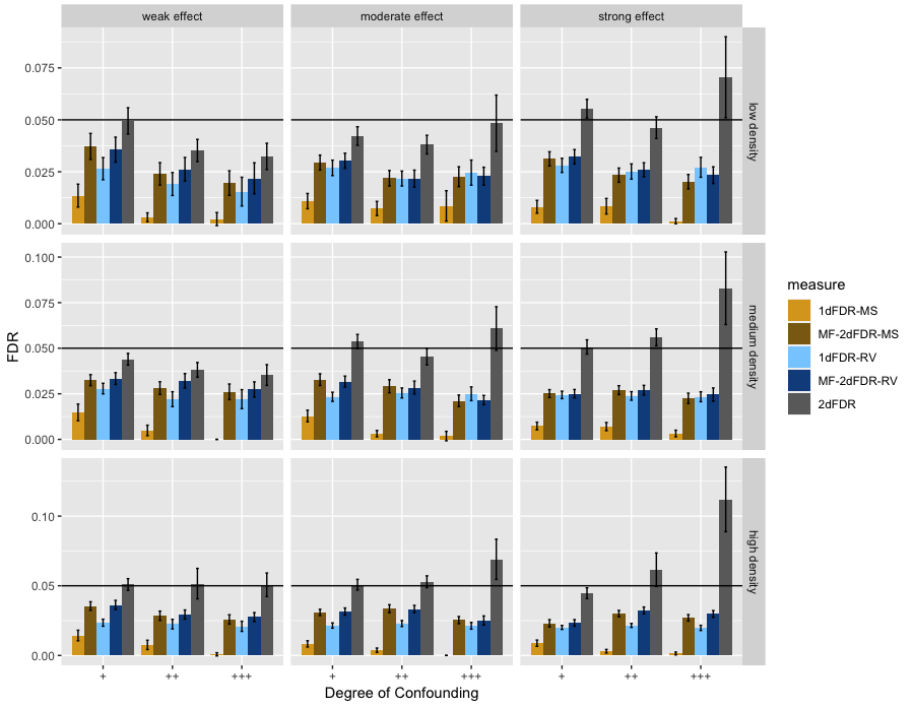


(a) FDR

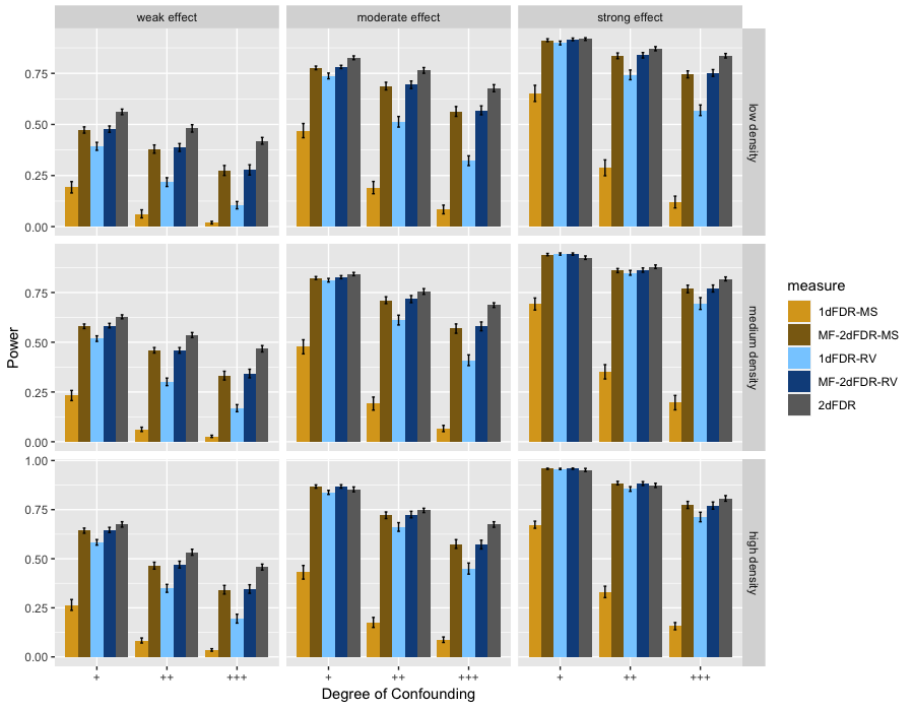


(b) Power

Figure 16: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, 2dFDR, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

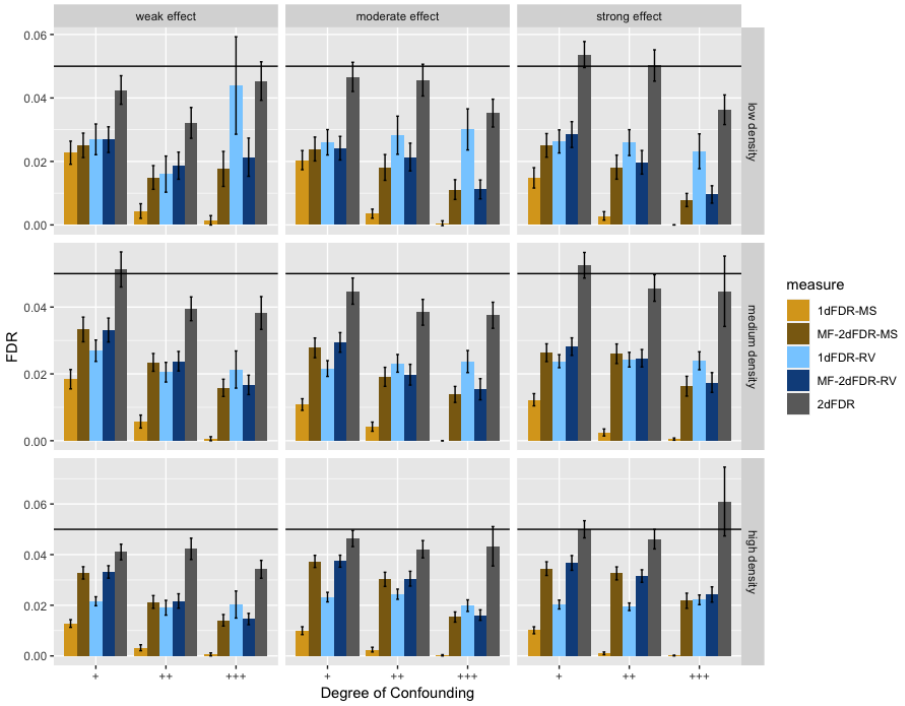


(a) FDR

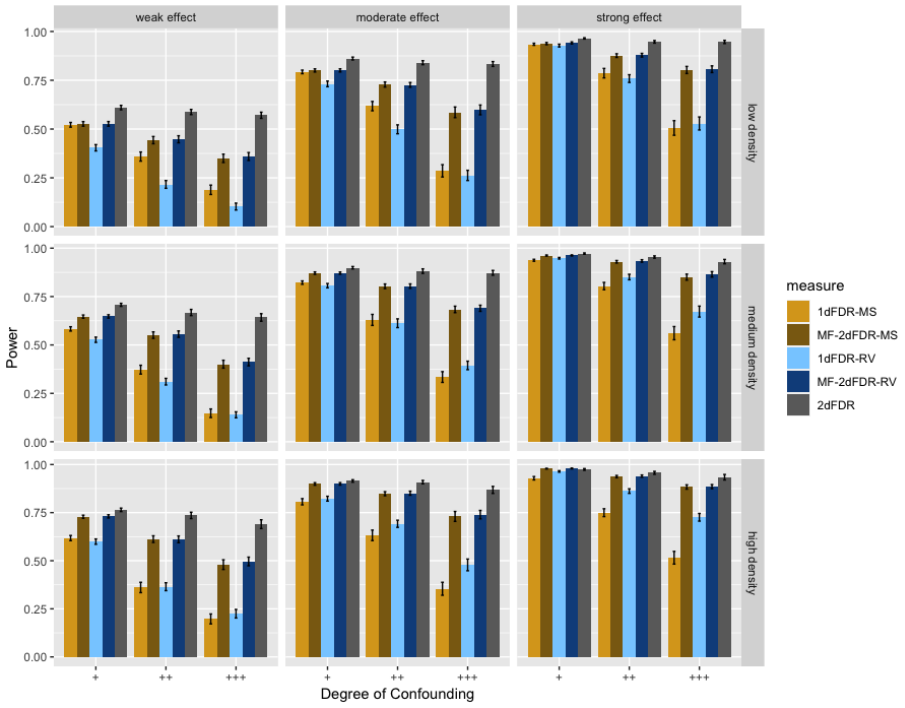


(b) Power

Figure 17: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, 2dFDR, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j e^Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1+e^{-\rho Z})^{-1})$, where $Z \sim N(0,1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

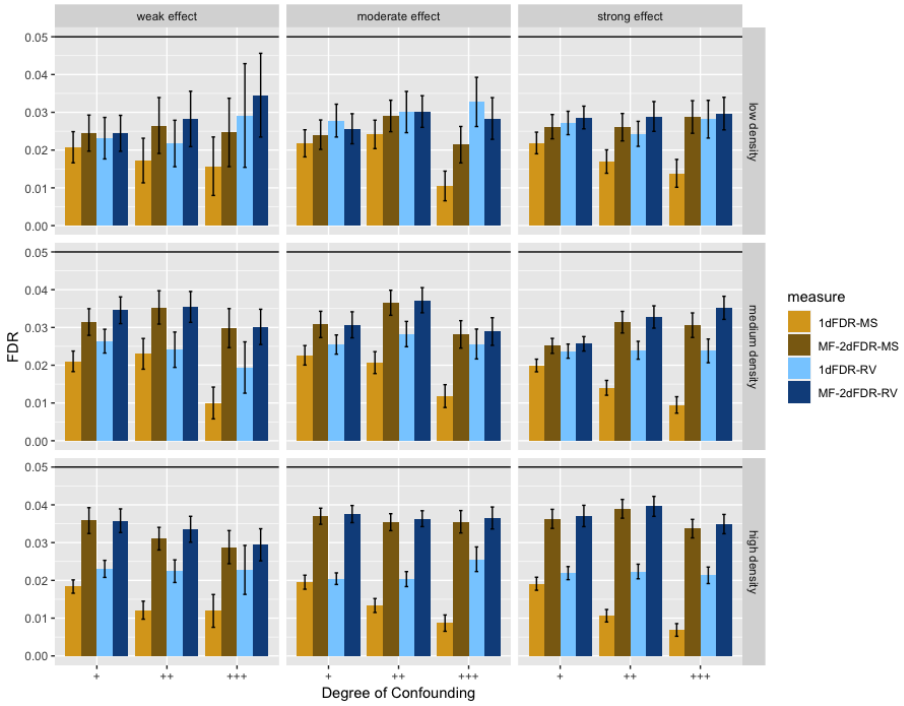


(a) FDR

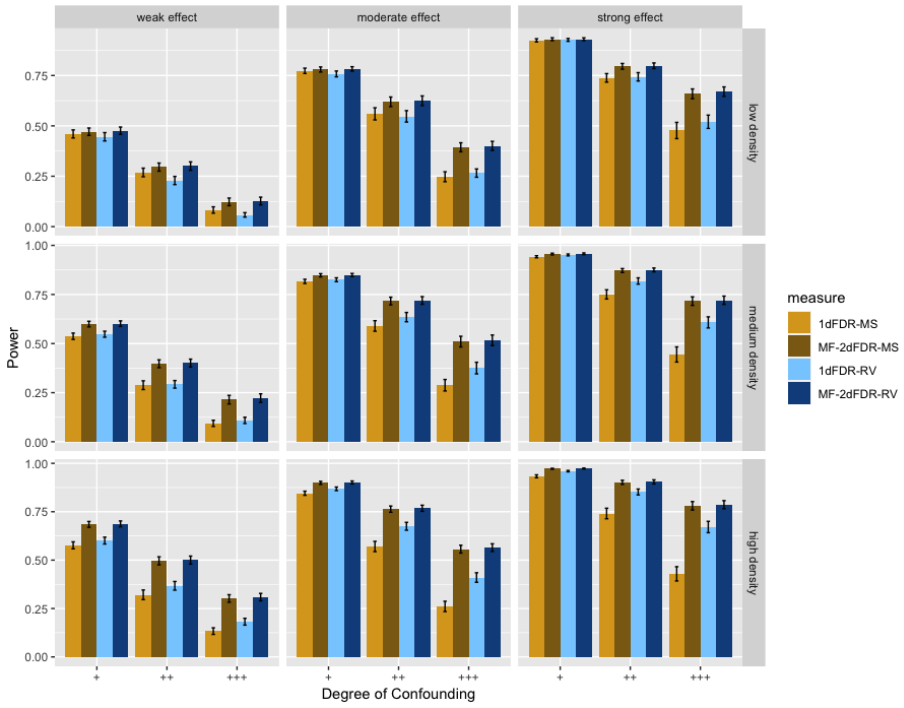


(b) Power

Figure 18: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, 2dFDR, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j Z^2 + \epsilon_j$ and $X \sim \text{Bernoulli}((1+e^{-\rho Z})^{-1})$, where $Z \sim N(0,1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

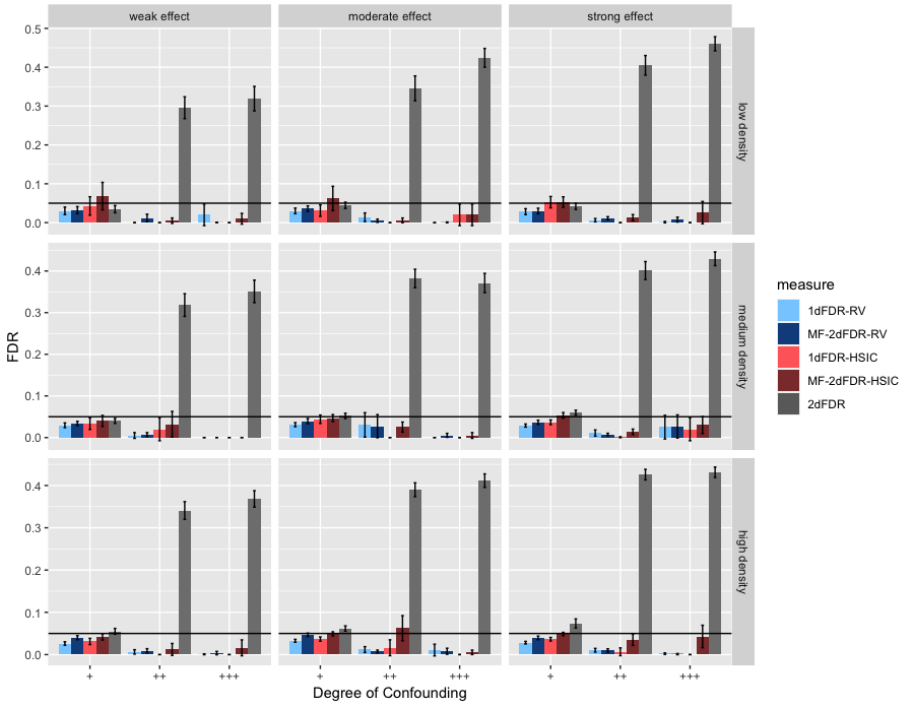


(a) FDR

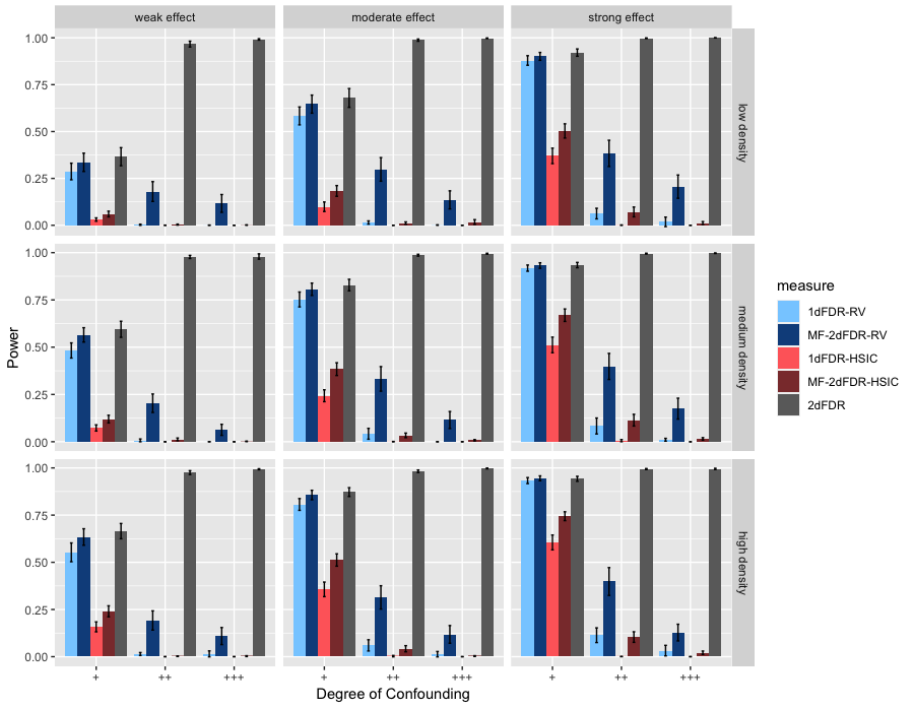


(b) Power

Figure 19: Empirical FDR and power for 1dFDR-MS, 1dFDR-RV, MF-2dFDR-MS, MF-2dFDR-RV under the model $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ and $X \sim \text{Bernoulli}((1 + e^{-\rho Z})^{-1})$, where $Z \sim \text{Bernoulli}(0.7)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

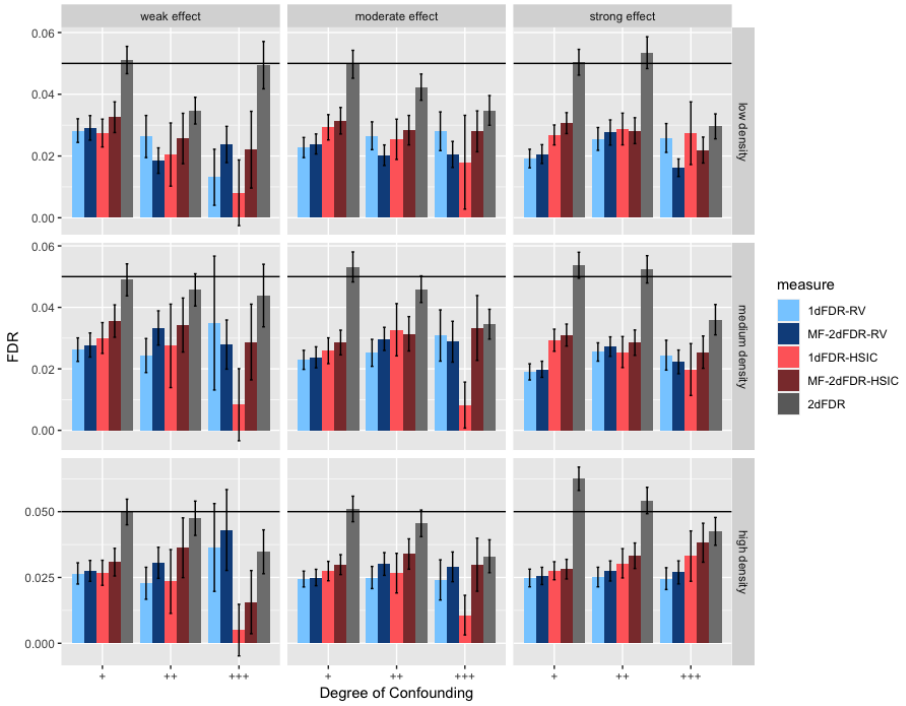


(a) FDR

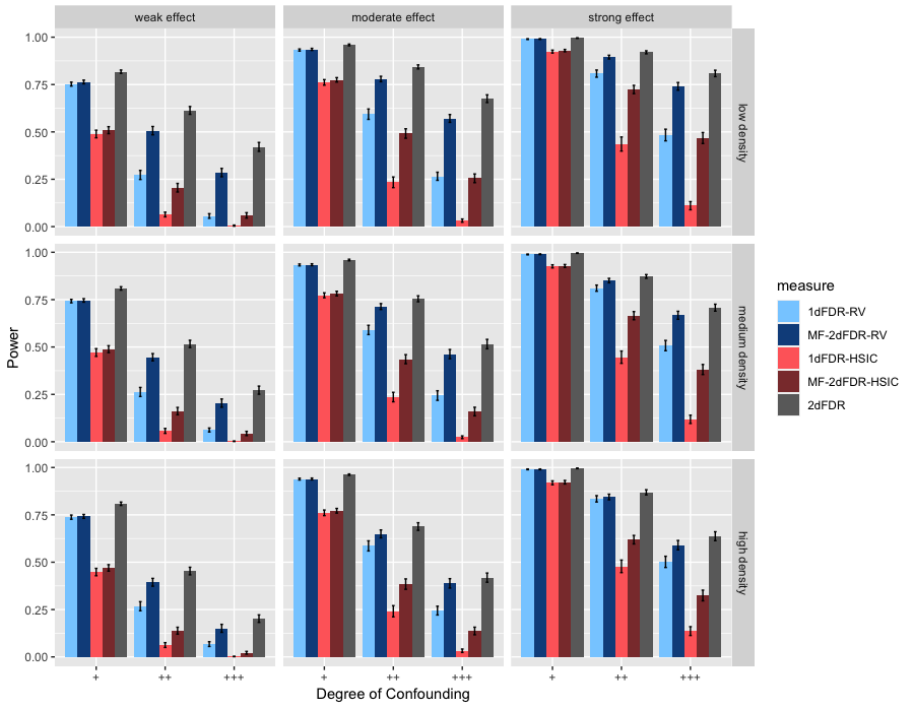


(b) Power

Figure 20: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j e^Z + \epsilon_j$, where ϵ_j follows an AR(1) model with the AR(1) coefficient being 0.7, $X \sim N(\rho(Z + Z^2), 1)$ and $Z \sim N(0, 1)$ Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

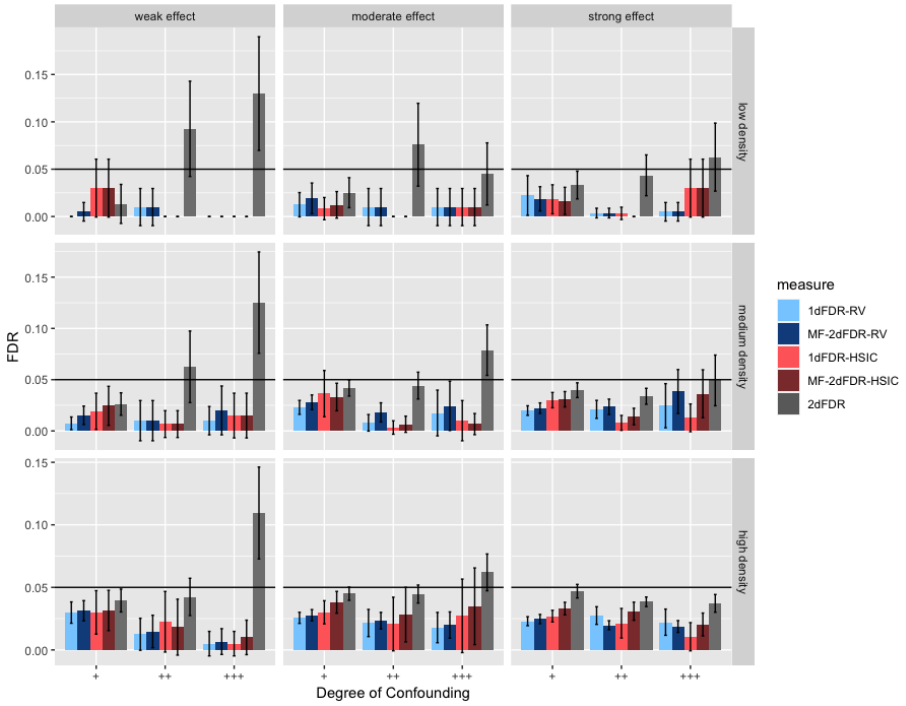


(a) FDR

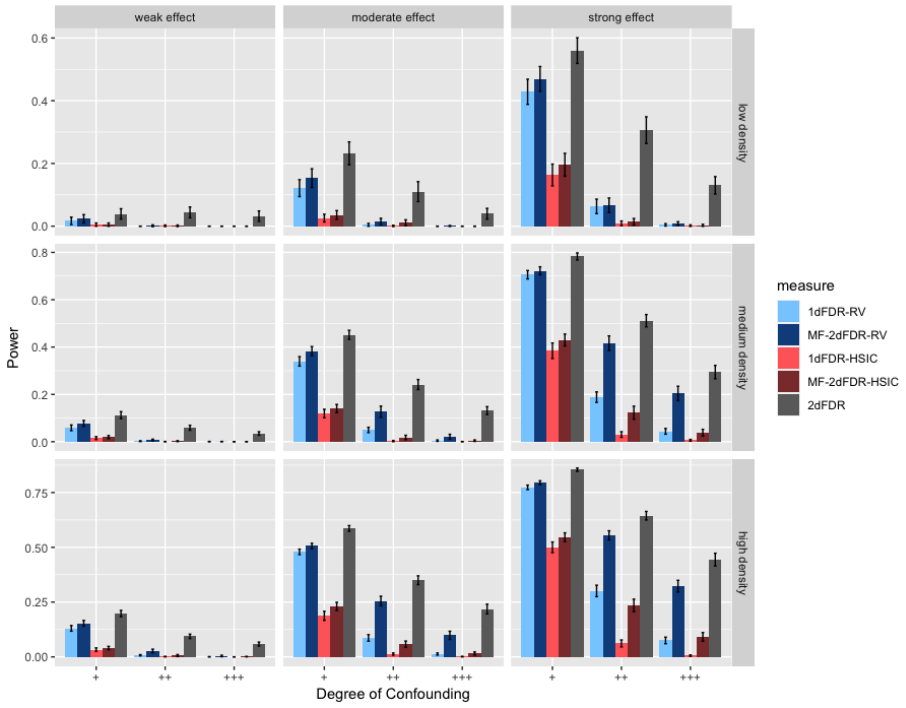


(b) Power

Figure 21: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ where $X \sim N(\rho Z, 1)$ and $Z \sim N(0, 1)$. The signal density of α_j has been fixed at 10 % while the signal density of β_j has been varied through 1%, 5% and 10%. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

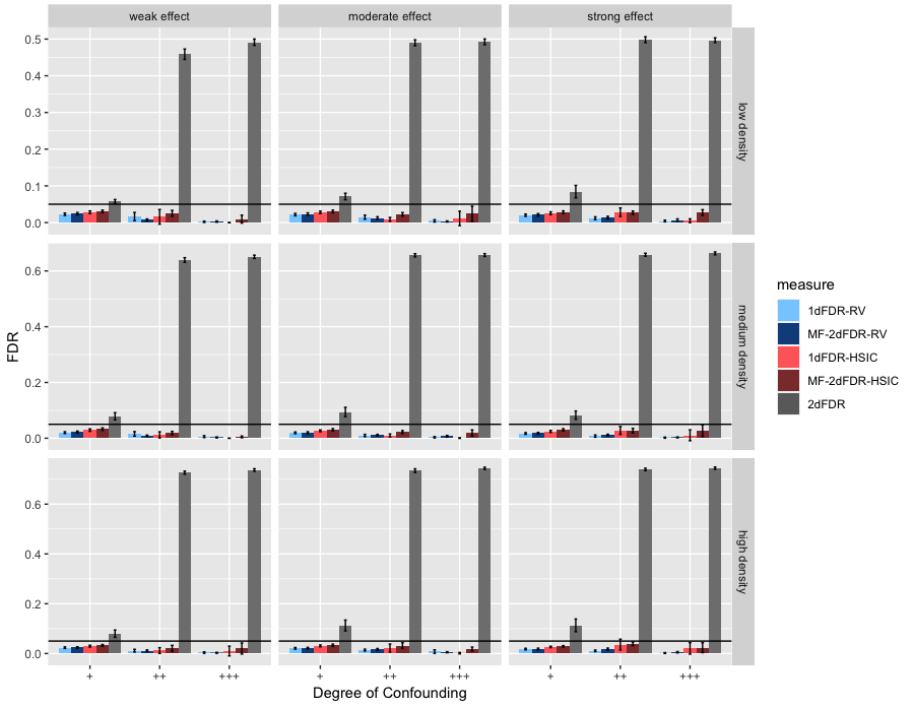


(a) FDR

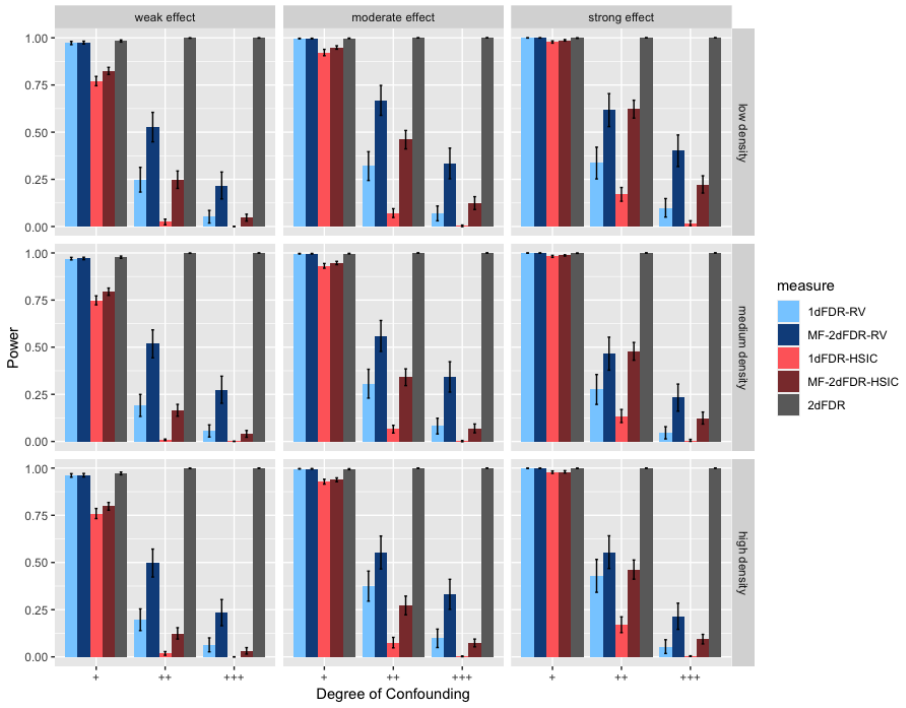


(b) Power

Figure 22: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$ where $X \sim N(\rho Z, 1)$ and $Z \sim N(0, 1)$. The signal density of β_j has been fixed at 10 % while the signal density of α_j has been varied through 1%, 5% and 10%. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

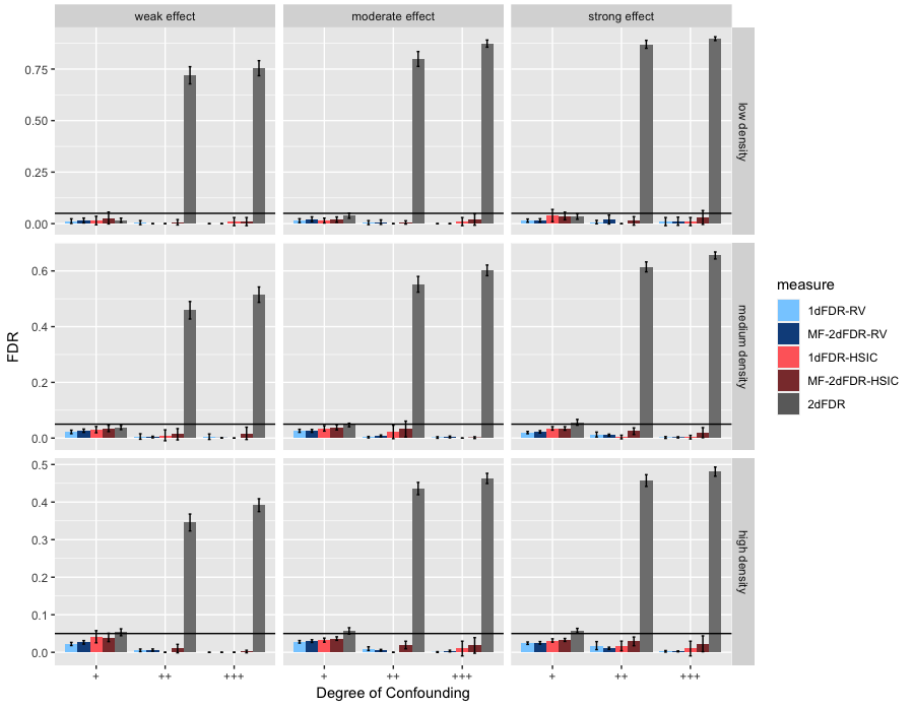


(a) FDR

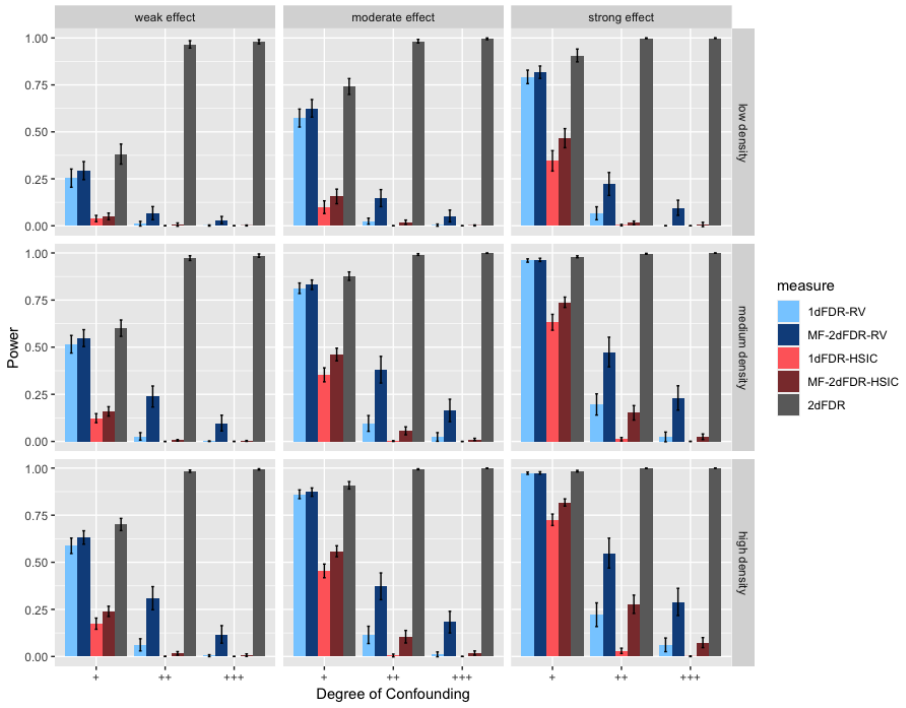


(b) Power

Figure 23: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j Z^2 + \epsilon_j$ where $X \sim N(\rho Z^2, 1)$ and $Z \sim N(0, 1)$. The signal density of α_j has been fixed at 10 % while the signal density of β_j has been varied through 1%, 5% and 10%. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

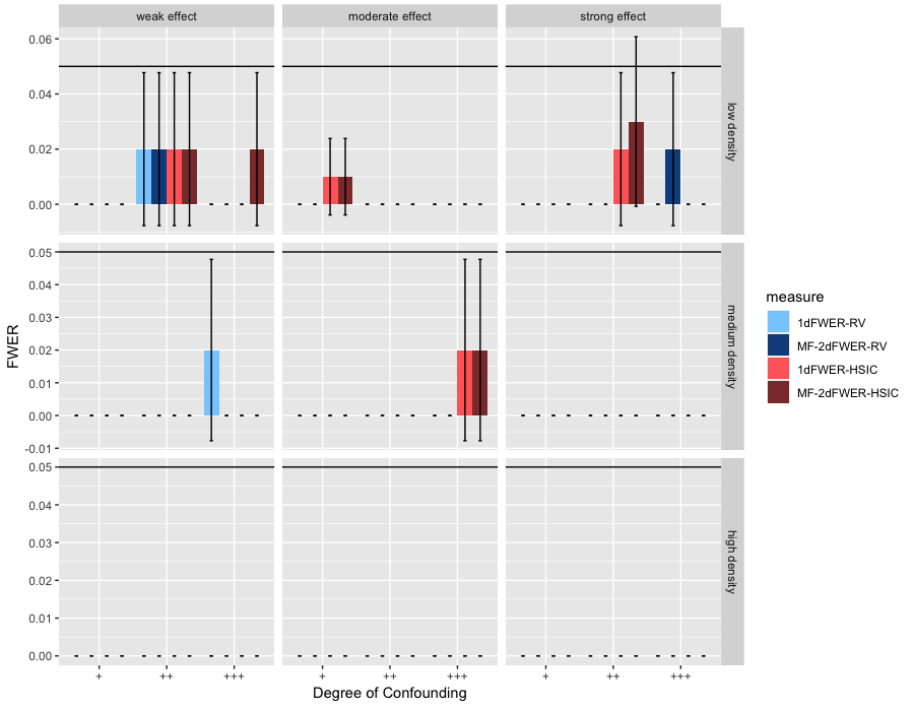


(a) FDR

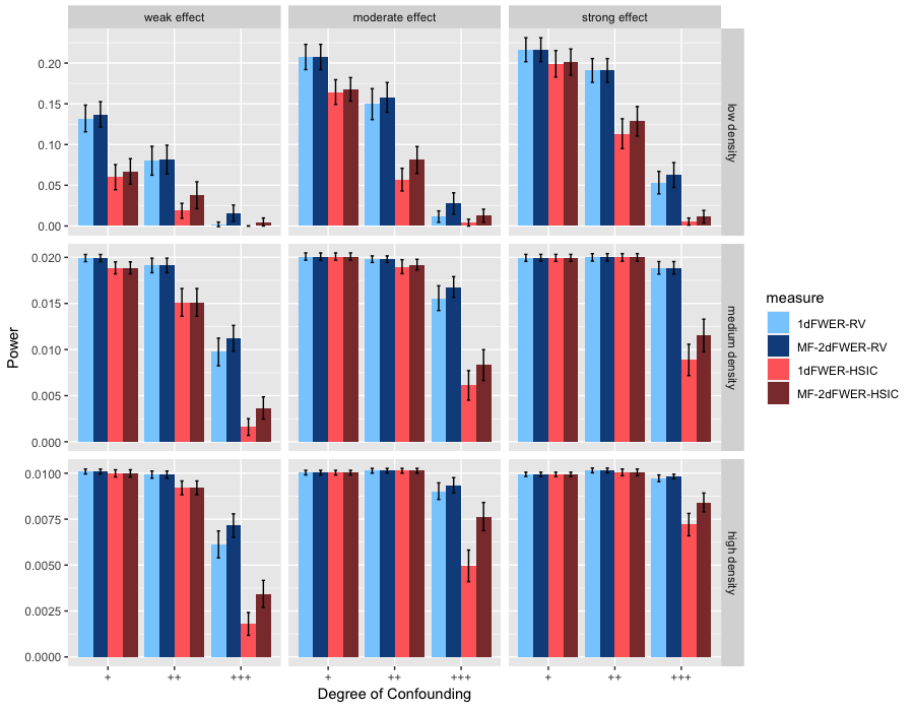


(b) Power

Figure 24: Empirical FDR and power for 1dFDR-HSIC, 1dFDR-RV, 2dFDR, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j Z^2 + \epsilon_j$ where $X \sim N(\rho Z^2, 1)$ and $Z \sim N(0, 1)$. The signal density of β_j has been fixed at 10 % while the signal density of α_j has been varied through 1%, 5% and 10%. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.

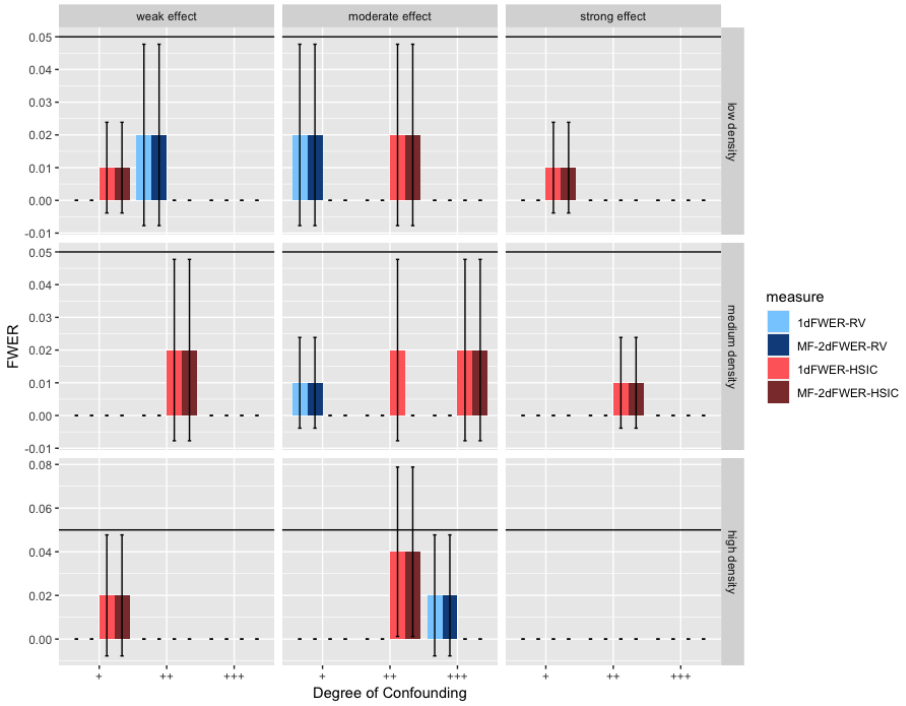


(a) FWER

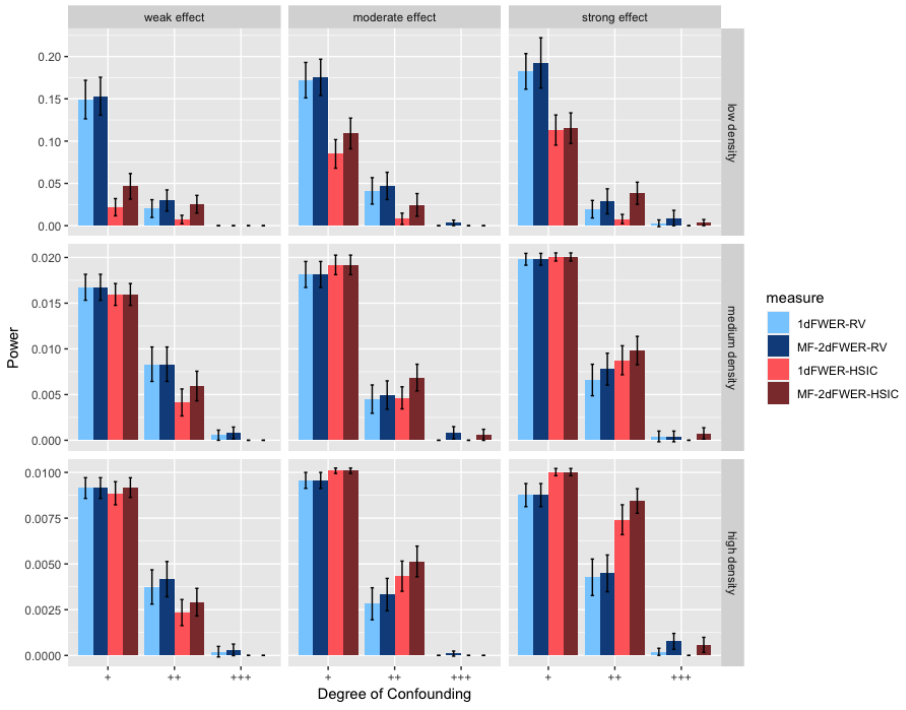


(b) Power

Figure 25: Empirical FWER and power for 1dFDR-HSIC, 1dFDR-RV, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j X + \beta_j Z + \epsilon_j$, where $X \sim N(\rho Z, 1)$ and $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FWER level of 0.05.



(a) FWER



(b) Power

Figure 26: Empirical FWER and power for 1dFDR-HSIC, 1dFDR-RV, MF-2dFDR-HSIC, MF-2dFDR-RV under the model $Y_j = \alpha_j e^X + \beta_j Z^2 + \epsilon_j$, where $X \sim N(\rho Z^2, 1)$ and $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FWER level of 0.05.