genes

MDPI

Article

Robust Differential Abundance Analysis of Microbiome Sequencing Data

Guanxun Li [1], Lu Yang [2], Jun Chen [2,*] and Xianyang Zhang [1,*]

[1] Department of Statistics, Texas A&M University, College Station, TX 77843, USA; guanxun@stat.tamu.edu
[2] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA; yang.lu@mayo.edu
* Correspondence: chen.Jun2@mayo.edu (J.C.); zhangxiany@stat.tamu.edu (X.Z.)

Abstract: It is well known that the microbiome data are ridden with outliers and have heavy distribution tails, but the impact of outliers and heavy-tailedness has yet to be examined systematically. This paper investigates the impact of outliers and heavy-tailedness on differential abundance analysis (DAA) using the linear models for the differential abundance analysis (LinDA) method and proposes effective strategies to mitigate their influence. The presence of outliers and heavy-tailedness can significantly decrease the power of LinDA. We investigate various techniques to address outliers and heavy-tailedness, including generalizing LinDA into a more flexible framework that allows for the use of robust regression and winsorizing the data before applying LinDA. Our extensive numerical experiments and real-data analyses demonstrate that robust Huber regression has overall the best performance in addressing outliers and heavy-tailedness.

Keywords: compositional data; differential abundance analysis; Huber regression; robustness; winsorization

Citation: Li, G.; Yang, L.; Chen, J.; Zhang, X. Robust Differential Abundance Analysis of Microbiome Sequencing Data. *Genes* **2023**, *14*, 2000. https://doi.org/10.3390/genes14112000

Academic Editor: Silvia Turroni

Received: 2 October 2023
Revised: 20 October 2023
Accepted: 24 October 2023
Published: 26 October 2023

1. Introduction

The human microbiome is a complex and multifaceted ecosystem comprising diverse microorganisms, including bacteria, viruses, and fungi. These microorganisms inhabit different parts of the human body and play a crucial role in various biological functions essential for human health and disease prevention [1,2]. Understanding the diversity and abundance of microorganisms in the microbiome, such as the gut microbiome, is crucial for identifying potential pathogens that can cause harm or probiotics that promote good health [3,4].

Metagenomic sequencing is the primary method used to study the microbiome. It provides a comprehensive snapshot of the microbiome's composition by sequencing the genetic material of all microorganisms in a sample [5,6]. However, this technique only provides relative abundance data, with the abundance of each microorganism expressed as a proportion of the total number of microorganisms in the sample [7]. Absolute abundance measurement can be achieved through various experimental techniques like qPCR, spike-in, and flow cytometry. However, these techniques have yet to be widely adopted due to their severe limitations [8]. Therefore, the prevailing sequencing protocol is still only capable of measuring the relative abundances. Nevertheless, when combined with appropriate statistical methods, relative abundance data can still provide valuable insights into the composition and function of the microbiome and its impact on health and disease. Following the processing of sequence reads using a bioinformatic pipeline, such as DADA2 [9] for 16S-targeted sequencing and MetaPhlAn2 [6] for shotgun metagenomic data, an abundance table that records the frequencies of the detected microbial taxa is generated. This table is used for downstream statistical analyses along with metadata that captures sample-level characteristics.

Differential abundance analysis (DAA) is a central downstream task that seeks to identify microbial taxa whose abundance correlates with a variable of interest. However,

classical statistical tools, such as ANOVA and rank-based tests, are unsuitable for DAA as they do not account for the compositional nature of microbiome data and can lead to significant false discoveries. This is due to changes in one microorganism's abundance that can influence others' relative abundance. Several differential analysis methods have been developed that consider compositional effects [10] to address this issue. These methods use either robust normalization techniques, such as TMM, RLE, CSS, and GMPR [11–14] or (log) ratio-based approaches, including ALDEx2 [15], ANCOM-BC [16], MaAsLin2 [17], and LinDA [18].

Identifying differentially abundant taxa in real-world data can be challenging due to the presence of outliers (overly abundant taxa) that may negatively impact the effectiveness of existing DAA methods [19–21], where outliers mean an extremely high abundance of a particular taxon in a few samples. Despite their widespread use, there has been a need for more systematic investigations into the influence of outliers on DAA methods. Additionally, log-ratio-based approaches (e.g., LinDA, ANCOM-BC, MaAsLin2) rely on normality assumption for random errors, which may not be valid for real-world data. Our observations suggest that this assumption can be violated in real-world data, leading to a phenomenon we term heavy-tailedness, which can impair the performance of existing DAA methods. Here, heavy-tailedness means that the tail of the distribution of random errors is heavier than the normal distribution.

This study aims to investigate the impact of outliers and heavy-tailedness on differential abundance analysis (DAA) methods and identify effective strategies to mitigate these challenges. LinDA was used as the benchmark DAA method, and the primary aim was to identify the most effective strategy to address outliers and heavy-tailedness in DAA. The study revealed that the presence of outliers or heavy-tailedness significantly reduces the power of detecting differential taxa. To address these challenges, we proposed a general M-estimation framework for DAA, which encompasses differential analysis based on Huber regression as a special case. Huber regression is a widely used statistical method that guards against outliers and heavy-tailedness in regression problems [22]. Additionally, this study explored the effectiveness of the winsorization method, a statistical data preprocessing technique in combination with LinDA, for handling noisy data. Winsorization replaces extreme values with less extreme ones and has been shown to effectively handle outliers and heavy-tailedness in data analysis [23]. This study's simulations demonstrated that Huber regression exhibits superior robustness against outliers and heavy-tailedness compared to the LinDA and LinDA methods with winsorization. Therefore, this study recommends using the Huber method in instances wherein the dataset is subject to noise.

We summarize our main contributions as follows:

1. This study conducted comprehensive simulations to investigate the impact of outliers and different types of heavy-tailedness on the various DAA methods.
2. This study introduced a general M-estimation framework for DAA, which includes several methods that are robust to outliers and heavy-tailedness.
3. This study conducted extensive simulations to examine the performance of various DAA methods to address outliers and heavy-tailedness, including multiple levels of the winsorization technique and different M-estimation-based methods. The experiments demonstrated that the proposed Huber regression method based on the M-estimation framework is more stable for outliers and heavy-tailedness.

The remainder of this paper is organized as follows. In Section 2.1, we introduce the regression-based framework for DAA, which includes LinDA and the Huber regression-based method as special cases. Section 2.2 provides an overview of the winsorization method and its implementation details. Our simulation studies are presented in Section 3. Lastly, Section 4 presents some real data analyses. Additional numerical results can be found in the Appendixes A–C.

2. Materials and Methods

2.1. A Regression-Based Framework for Differential Analysis

In this section, we present a generalization of the LinDA method, utilizing a M-estimation framework for DAA based on the central log ratio (CLR) transformation. By replacing the l_2 loss with a robust loss function, we propose a more robust approach for DAA.

2.1.1. CLR-Based Log-Linear Models

Let X_{is} and Y_{is} denote the absolute abundance and observed read count of the i-th taxon in the s-th sample, respectively. For the s-th sample, the total read count of all taxon, $N_s = \sum_{i=1}^{m} Y_{is}$, is determined by the sequencing depth and DNA materials. Given N_s, it is natural to model the stratified count data over m taxon using a multinomial distribution as

$$P(Y_{1s} = y_{1s}, \ldots, Y_{ms} = y_{ms}) = \frac{N_s!}{\prod_{i=1}^{m} y_{is}!} \prod_{j=1}^{m} \left(\frac{X_{js}}{\sum_{i=1}^{m} X_{is}} \right)^{y_{js}}. \tag{1}$$

Under (1), we have

$$\log\left(\frac{Y_{is}}{\sum_{j=1}^{m} Y_{js}} \right) = \log\left(\frac{X_{is}}{\sum_{j=1}^{m} X_{js}} \right) + e_{is}, \tag{2}$$

where the ratio on the left side represents the sample proportion, while the ratio on the right side represents the population proportion, and e_{is} denotes the estimation error. The sample proportion is the maximum likelihood estimator (MLE) of the population proportion based on the multinomial model (1). By the consistency of the MLE, e_{is} is expected to diminish as N_s increases. We consider the log linear model on the absolute abundance

$$\log(X_{is}) = u_s \alpha_{0i} + \mathbf{c}_s^\top \boldsymbol{\beta}_{0i} + \epsilon_{is}, \tag{3}$$

where $\mathbf{c}_s = (1, c_{s1}, \ldots, c_{sd})^\top$ includes intercept and the d-dimensional covariates to be adjusted, u_s is the variable of interest, and ϵ_{is} is the error term. Here, we assume a log-linear model on the absolute abundance, which is a reasonable and widely adopted approach in abundance data analysis [24,25]. The objective here is to identify taxa that show differential abundance relative to u_s. To this end, we simultaneously test the following m hypotheses:

$$H_{0,i} : \alpha_{0i} = 0 \text{ versus } H_{a,i} : \alpha_{0i} \neq 0, \quad 1 \leq i \leq m.$$

Set $\varepsilon_{is} = \epsilon_{is} + e_{is}$. Under (2) and (3), the CLR data satisfy the following linear model:

$$\begin{aligned} W_{is} :=& \log\left\{ \frac{Y_{is}}{(\prod_{j=1}^{m} Y_{js})^{1/m}} \right\} = \log\left(\frac{Y_{is}}{\sum_{k=1}^{m} Y_{ks}} \right) - \frac{1}{m} \sum_{j=1}^{m} \log\left(\frac{Y_{js}}{\sum_{k=1}^{m} Y_{ks}} \right) \\ =& \log(X_{is}) - \frac{1}{m} \sum_{j=1}^{m} \log(X_{js}) + e_{is} - \frac{1}{m} \sum_{j=1}^{m} e_{js} \\ =& u_s(\alpha_{0i} - \bar{\alpha}) + \mathbf{c}_s^\top (\boldsymbol{\beta}_{0i} - \bar{\boldsymbol{\beta}}) + \varepsilon_{is} - \bar{\varepsilon}_s, \end{aligned}$$

where $\bar{\alpha} = m^{-1} \sum_{i=1}^{m} \alpha_{0i}$, $\bar{\boldsymbol{\beta}} = m^{-1} \sum_{i=1}^{m} \boldsymbol{\beta}_{0i}$, and $\bar{\varepsilon}_s = m^{-1} \sum_{i=1}^{m} \varepsilon_{is}$.

2.1.2. M-Estimation Framework for Differential Analysis

Define $\tilde{\alpha}_i = \alpha_{0i} - \bar{\alpha}$ and $\tilde{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_{0i} - \bar{\boldsymbol{\beta}}$. We propose to estimate $\tilde{\alpha}_i$ and $\tilde{\boldsymbol{\beta}}_i$ by solving the following M-estimation problem:

$$(\tilde{\alpha}_i, \tilde{\boldsymbol{\beta}}_i) = \arg\min_{\alpha_i, \boldsymbol{\beta}_i} \frac{1}{n} \sum_{s=1}^{n} \mathcal{L}\left(W_{is} - u_s \alpha_i - \mathbf{c}_s^\top \boldsymbol{\beta}_i \right), \tag{4}$$

where $\mathcal{L}$ is a loss function chosen by the practitioners and the resulting estimators are referred to as M-estimators. When $\mathcal{L}$ is the l_2 loss, i.e., $\mathcal{L}(r) = r^2$, $(\tilde{\alpha}_i, \tilde{\boldsymbol{\beta}}_i)$ become the ordinary least squares (OLS) estimators, which has been considered in [18].

Example 1. *To handle heavy-tailedness and outliers in datasets, we can employ the robust loss function in (4), which is commonly used in robust regression. Robust regression is designed to estimate the parameters of a regression model in the presence of influential observations that can distort the results. Unlike ordinary least squares regression, which assumes normally distributed errors with constant variance, robust regression offers more flexible assumptions about the error distribution and is less sensitive to outliers. It becomes particularly advantageous when data contain outliers that cannot be readily removed or explained, or when ensuring that the regression coefficients are not overly influenced by a few observations. Below are some popular robust regression loss functions.*

1. *Huber's loss is defined as*

$$l_{\text{Huber}}(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le c \\ c|r| - \frac{1}{2}c^2 & \text{if } |r| > c \end{cases},$$

   where c is the hyperparameter and the default value is 1.345*, which is widely used in robust regression studies. Notice that the Huber estimator down-weights the influence of observations with large residuals, resulting in less impact on the estimated regression coefficients. Additionally, Huber [22] argued that if the true distribution was normal, this loss function is asymptotically 95% as efficient as least squares.*
2. *Tukey's bisquare loss is defined as*

$$l_{bi}(r) = \begin{cases} \frac{c_0^2}{6}\left(1 - \left(1 - \left(\frac{r}{c_0}\right)^2\right)^3\right) & \text{if } |r| \le c_0 \\ \frac{c_0^2}{6} & \text{if } |r| > c_0 \end{cases},$$

   where $c_0 = 4.685$ is the standard constant for this loss function. It is worth noting that this function has been shown to possess an asymptotic efficiency of 95% with respect to linear regression for the normal distribution.
3. *Quantile regression loss is defined as*

$$l_\tau(r) = \begin{cases} \tau r & \text{if } r \ge 0 \\ (\tau - 1)r & \text{if } r < 0 \end{cases},$$

   where τ represents the quantile level of interest. Notably, when $\tau = \frac{1}{2}$, the loss function is equivalent to the l_1 loss, expressed as $l_1(r) = |r|$. The l_1 loss corresponds to the loss function utilized in the least absolute deviations (LAD) regression method.

Let $\mathbf{z}_s = (u_s, \mathbf{c}_s^\top)^\top$ and $\theta_i = (\alpha_i, \boldsymbol{\beta}_i^\top)^\top$. We define $\tilde{\theta}_i$ (and $\bar{\theta}_i$) in the same way as θ_i by replacing α_i with $\tilde{\alpha}_i$ (and $\bar{\alpha}_i$), and $\boldsymbol{\beta}_i$ with $\tilde{\boldsymbol{\beta}}_i$ (and $\bar{\boldsymbol{\beta}}_i$), respectively. Denote $r_{is} = W_{is} - \mathbf{z}_s^\top \theta_i$, and define $\bar{r}_{is}$ and $\tilde{r}_{is}$ analogously by replacing θ_i with $\bar{\theta}_i$ and $\tilde{\theta}_i$, respectively. We can then rewrite each summand of the objective function in (4) as $\mathcal{L}(r_{is}) := \mathcal{L}(W_{is} - \mathbf{z}_s^\top \theta_i)$. Observe that

$$\frac{\partial}{\partial \theta_i}\mathcal{L}(r_{is}) = \mathbf{z}_s \psi(r_{is}),$$

where ψ is referred to as the influence curve [26]. Under certain regularity conditions (see, for example, Theorem 5.21 in Van der Vaart [27]), the asymptotic normality of the M-estimator is given as follows [26]:

$$\sqrt{n}(\tilde{\theta}_i - \bar{\theta}_i) \to^d \mathcal{N}(0, \Sigma_i), \tag{5}$$

where

$$\Sigma_i = \left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1}\frac{\mathbb{E}[\psi^2(\bar{r}_{is})]}{\mathbb{E}[\psi'(\bar{r}_{is})]^2}. \tag{6}$$

In practice, we will not know the true distribution of the error term and the true regression parameters $\bar{\theta}_i$. Therefore, we propose using the plug-in method to estimate the asymptotic variance of $\tilde{\theta}_i$ (rather than $\sqrt{n}\tilde{\theta}_i$), which is given by

$$\hat{\Sigma}_i = \left(\sum_{s=1}^{n}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1}\frac{\frac{1}{n}\sum_{s=1}^{n}\psi^2(\tilde{r}_{is})}{\left(\frac{1}{n}\sum_{s=1}^{n}\psi'(\tilde{r}_{is})\right)^2}. \tag{7}$$

When $\mathcal{L}$ is the l_2 loss, the estimators used in LinDA and (7) are asymptotically equivalent.

From the above discussions, $\tilde{\alpha}_i$ obtained by minimizing (4) is an asymptotically unbiased estimator for $\alpha_i - \bar{\alpha}$. To estimate α_i, it remains to find an estimator for $\bar{\alpha}$. To this end, we adopt the mode correction method proposed by Zhou et al. [18]. Specifically, assuming that only a small portion of taxa exhibit differential abundance, meaning that most values of α_i are equal to 0. Under this assumption, the mode of $\tilde{\alpha}_i$ is expected to be close to $-\bar{\alpha}$. Hence, we can estimate $\bar{\alpha}$ through estimating the mode of $\{\tilde{\alpha}_i : 1 \le i \le m\}$. Specifically, we use the kernel smoothing approach to estimate the mode and let

$$\tilde{\alpha} := -\frac{\widehat{\text{mode}}(\{\sqrt{n}\tilde{\alpha}_i\}_{i=1}^m)}{\sqrt{n}}$$

be the estimate for $\bar{\alpha}$. Here,

$$\widehat{\text{mode}}(\{\sqrt{n}\tilde{\alpha}_i\}_{i=1}^m) = \arg\max_{a\in\mathbb{R}}\frac{1}{mh}\sum_{i=1}^{m}\mathcal{K}\left(\frac{a-\sqrt{n}\tilde{\alpha}_i)}{h}\right),$$

where $\mathcal{K}$ is a kernel function satisfying $\int_{-\infty}^{\infty}\mathcal{K}(x)\,dx = 1$ and h is the bandwidth. Then, the resulting bias-corrected estimator of α_i is given by $\hat{\alpha}_i = \tilde{\alpha}_i + \tilde{\alpha}$. In our implementation, we use *mlv* function in R package *modeest* to estimate the mode.

Once the bias-corrected estimator is obtained, we implement the t-test by defining the test statistic as $T_i = \hat{\alpha}_i/\hat{\sigma}_i$, where $\hat{\sigma}_i^2$ is the variance estimator of $\tilde{\alpha}_i$ in the regression problem, corresponding to the $(1,1)$th entry of $\hat{\Sigma}_i$ defined in (7). Although we have the asymptotic normality of the M-estimator, our simulations revealed that the t-distribution provides a better approximation for the sampling distribution of T_i and offers better FDR control for small samples. Consequently, we propose using the t-test in our method. The p-value is then calculated as $2P(T \ge |T_i|)$, where T follows a t-distribution with $n-d-2$ degrees of freedom. To control the FDR, we recommend using the Benjamini–Hochberg (BH) procedure to adjust the p-values obtained for each taxon. The taxa with an adjusted p-value less than a certain threshold are referred to as differential taxa.

We summarize our procedure as follows:

1. For each taxon, solve the optimization problem defined in (4) to obtain the estimator $\tilde{\alpha}_i$ and its variance estimator $\hat{\sigma}_i^2$.
2. Calculate the mode based on $\{\tilde{\alpha}_i : 1 \le i \le m\}$, and perform mode correction to obtain the bias-corrected estimator $\hat{\alpha}_i$.
3. Compute the test statistics T_i and the p-value for each taxon.
4. Apply the BH procedure to adjust the p-values.

Remark 1. *Given the practical challenge of determining the optimal hyperparameter to use in the loss function (e.g., c in the Huber loss function), we propose utilizing the Cauchy combination rule [28] for aggregating the p-values obtained from different hyperparameters. This approach*

addresses the difficulty of selecting the most suitable hyperparameter by combining the results from multiple options.

Remark 2. *Based on the regression-based framework, our method can be readily extended to apply to the mixed-effect model. Please refer to Appendix B for more details.*

2.2. Winsorization

Winsorization is a statistical data preprocessing technique utilized for handling outliers and heavy-tailed data. This method involves sorting the dataset in either ascending or descending order based on the analytical requirements. Next, the extreme values, namely outliers or values in the tails of the distribution, are substituted with the smallest or largest non-outlier value, correspondingly. Moreover, winsorization can be complemented with DAA techniques to enhance the precision and robustness of statistical analysis, particularly when confronted with datasets containing outliers or heavy-tailed distributions.

In microbiome data analysis, winsorization typically entails replacing the top $1-\tau$ largest values with the value at the τ quantile [20]. Specifically, for a given taxon i, Y_{is} was arranged in increasing order, and the τ quantile of $\{Y_{is}\}_{s=1}^{n}$ was computed and denoted as $q_i(\tau)$. We replace the observed count Y_{is} by its winsorized value defined as

$$\widetilde{Y}_{is} = \begin{cases} Y_{is} & \text{if } Y_{is} \leq q_i(\tau), \\ q_i(\tau) & \text{if } Y_{is} > q_i(\tau). \end{cases}$$

While winsorization is known to reduce the influence of outliers, it remains unclear whether this method leads to the loss of essential data information. Furthermore, despite its widespread use, the impact of winsorization on DAA methods has not been thoroughly investigated. Consequently, this study seeks to comprehensively examine the impact of multiple levels of winsorization on DAA methods in various scenarios, both with and without outliers/heavy-tailedness, through several numerical examples.

3. Results

To comprehensively evaluate the performance of different M-estimation-based methods and winsorization at different levels, we conducted extensive simulations under various scenarios. Before describing the simulation settings, we provide a list of the methods we compared.

We evaluated six methods in our comparative analysis, namely LinDA [18] without winsorization (referred to as LinDA), LinDA with winsorization at the 97% quantile (referred to as LinDA97), LinDA with winsorization at the 90% quantile (referred to as LinDA90), M-estimation method with Huber's loss (referred to as Huber), M-estimation method with Tukey's bisquare loss (referred to as Bi_square), and quantile regression method (referred to as QR). To implement the Huber and bisquare methods, we used the *rlm* function in the *MASS* package (version: 7.3-60) in R to perform the regression estimation. For the selection of hyperparameters, we considered 10 values of c equally spaced within the interval $[1.345,\ 5]$ on a log scale for the Huber method. For the Bi_square method, we took 10 values of c_0 equally spaced within the interval $[4.685,\ 20]$ on a log scale. We used the *rq* function in the *quantreg* package (version: 5.95) in R to implement quantile regression. Similar to the Huber method, we took into account the quantile level τ that is evenly distributed across the interval $[0.25,\ 0.75]$ with an adjacent difference of 0.05 (resulting in a total of 11 values).

Remark 3. *We also compared with other differential abundance analysis methods, including ALDEx2 [15], ANCOM-BC [16], and MaAsLin2 [17], and the results are deferred to Appendix C. Based on our simulation, we found that LinDA outperforms other methods that do not consider outliers and heavy-tailedness. Hence, we used LinDA as the benchmark method in the following and focus on comparing various log-linear model-based approaches used for addressing outliers.*

To handle zero values, we employed the hybrid method proposed by Zhou et al. [18], which combines two different approaches. The first approach involves adding a pseudo-count of 0.5 to all counts, which is a commonly used technique in microbiome data analysis on the log scale. The second approach is the imputation-based method, which involves replacing the zeros with fractions equal to $N_s / \max\{N_k : Y_{ik} = 0\}$ for the i-th taxon in the s-th sample, where larger fractions are used for samples with larger library sizes. Zhou et al. [18] used a statistical test to determine which method to apply. Specifically, they tested the association between the covariate of interest and the library size based on the log-linear model. If the p-value was less than 0.1, they used the imputation approach; otherwise, they used the pseudo-count approach. More details can be found in Zhou et al. [18].

3.1. Simulations Based on Log-Linear Models

In this section, we simulated datasets from log-linear models. We adopted the data-generating process proposed by Zhou et al. [18], while using different methods to generate error terms. We assumed that the baseline absolute abundance X_{is}^* is generated from

$$\log(X_{is}^*) \sim^{\text{i.i.d.}} \mathcal{N}(\beta_i^*, \sigma_{i*}^2),$$

and the true proportion is obtained as

$$\pi_{is}^* = \frac{X_{is}^*}{\sum_{j=1}^m X_{js}^*}.$$

Letting $\bar{\pi}_i^* = \sum_{s=1}^n \pi_{is}^*/n$. We denote the signal strength as μ. In order to construct a power curve, we included six signal strengths in the figures, which are evenly spaced within the interval $[1.05, 2]$. Given that low-abundance taxa exhibit lower statistical power, we assigned greater weight to their effects to prevent dominance by the abundant ones. Specifically, for the i-th taxon, we set

$$\mu_i = \log(2\mu)\mathbb{1}(\bar{\pi}_i^* > 5 \times 10^{-3}) + \log\left(2\mu\left(5 \times 10^{-3}/\bar{\pi}_i^*\right)^{1/3}\right)\mathbb{1}(\bar{\pi}_i^* \leq 5 \times 10^{-3})$$

for $n = 50$ and

$$\mu_i = \log(\mu)\mathbb{1}(\bar{\pi}_i^* > 5 \times 10^{-3}) + \log\left(\mu(5 \times 10^{-3}/\bar{\pi}_i^*)^{1/3}\right)\mathbb{1}(\bar{\pi}_i^* \leq 5 \times 10^{-3})$$

for $n = 200$.

We randomly selected the differential taxon from the entire set and denoted γ_i as an indicator of whether the taxon is differentially abundant ($\gamma_i = 1$) or not ($\gamma_i = 0$). The underlying truth of γ_i was generated from a Bernoulli distribution with parameter p_γ, where we set $p_\gamma = 0.05$ or $p_\gamma = 0.2$ to correspond to sparse and dense signal settings, respectively. We then denoted the true signal strength for the differentially abundant taxon by $\alpha_i = \mu_i\gamma_i$.

To generate the absolute abundance X_{is}, we considered two cases: with or without confounders, and three types of error terms. We defined ϵ_{is} as the error term corresponding to the s-th sample of the i-th taxon, which we will define later. The absolute abundance X_{is} was then obtained by

$$\log(X_{is}) = \begin{cases} \beta_{i0} + u_s\alpha_i + \epsilon_{is} & \text{without confounders,} \\ \beta_{i0} + u_s\alpha_i + \mathbf{c}_s^\top\boldsymbol{\beta}_i + \epsilon_{is} & \text{with confounders,} \end{cases}$$

where β_{i0} is the intercept term and

$$u_s \sim \text{Bernoulli}(0.5)$$

if there is no confounder and

$$u_s \sim \text{Bernoulli}\big(1/\big(1+\exp(-0.5c_{s1}-0.5c_{s2})\big)\big)$$

if there are confounders. Here, the confounders $\mathbf{c}_s = (c_{s1}, c_{s2}) \in \mathbb{R}^{n\times 2}$, with $c_{s1} \sim$ Bernoulli(0.5) and $c_{s2} \sim \mathcal{N}(0,1)$. The taxon-specific coefficients β_{i1} and β_{i2} are independently generated from the normal distributions with means of 1 and 2 and variances of 1, respectively. Then, the observed operational taxonomic unit (OTU) data were generated from

$$(Y_{1s}, \cdots, Y_{ms}) \sim \text{Multinomial}(N_s, \pi_{1s}, \cdots, \pi_{ms}),$$

where $\pi_{is} = X_{is}/\sum_{j=1}^{m} X_{js}$. Moreover, the parameters β_i^*, σ_{i*}^2, and N_s used in our study are identical to those used in Zhou et al. [18]. Specifically, Zhou et al. [18] estimated the aforementioned parameters based on a real dataset (COMBO) that studied the gut microbiota in a general population.

We begin by examining the performance of the methods in the absence of heavy-tailedness and outliers. We aim to assess how much power is lost when using robust regression/winsorization. For this purpose, we sampled errors from a normal distribution with a mean of 0 and a variance of σ_{i*}^2. The results are presented in Appendix A.1. We consider sample sizes of $n \in \{50,\ 200\}$, referred to as the small sample case and the large sample case, respectively. We set the number of taxa as $m = 500$. In the small sample size scenario, LinDA97 shows the best performance in the sparse-signal setting, whereas LinDA outperforms other methods in the dense-signal setting. For larger sample sizes, LinDA, LinDA97, Huber, and Bi_square methods all perform similarly and outperform LinDA90. The QR method has the lowest power in all scenarios and this method has FDR inflation when $n = 200$ in the sparse-signal setting.

3.1.1. Heavy-Tailedness Setting

To demonstrate the advantages of our method in handling heavy-tailedness, instead of assuming a standard normal distribution for ϵ_{is}, we generate error terms using three different ways:

- Case 1: Student's t-distribution with degrees of freedom 3.
- Case 2: Log-normal distribution with log mean parameter of 0 and log standard deviation parameter of 0.8. We recentered the samples so that it has a zero mean.
- Case 3: Weibull distribution with shape parameter of 0.5 and scale parameter of 0.3. We recentered the samples so that it has a zero mean.

These three are all heavy-tailed distributions, also discussed in Fan et al. [29]. It is important to underscore the significance of choosing appropriate parameters for generating error terms. For example, when the shape parameter of the Weibull distribution is small (e.g., 0.25), the resulting data can become excessively noisy, causing all methods to have no power to detect differential taxa. On the other hand, when the shape parameter of the Weibull distribution exceeds 1 (meaning the Weibull distribution is no longer heavy-tailed), the noise level diminishes significantly. As a result, all methods tend to achieve a power nearing 1.

We consider two sample size scenarios: $n \in \{50, 200\}$, denoted as the small and large sample cases, respectively. Additionally, we fix the number of taxa at $m = 500$. We conducted 100 simulation runs for each setting, calculated the mean power to detect differential taxa, and computed the empirical FDR. The results are presented in plots. This section showcases results for settings without confounders in the sparse scenario ($p_\gamma = 0.05$). Similar phenomena were observed in both the dense scenario and the case with confounders. These are detailed in Appendices A.2 and A.3, respectively.

Figure 1 presents the results obtained when error terms are generated from a t-distribution (Case 1). All methods show an increased power to detect differential taxa as the sample size grows. The Huber, Bi_square, and LinDA90 methods perform comparably for small sample sizes, each showing higher power than the LinDA97 and LinDA methods.

However, as the sample size grows, the performances of the LinDA90, LinDA97, Huber, and Bi_square methods converge, all surpassing the power of the LinDA method. Notably, the QR method consistently exhibits the lowest power and suffers from significant FDR inflation.

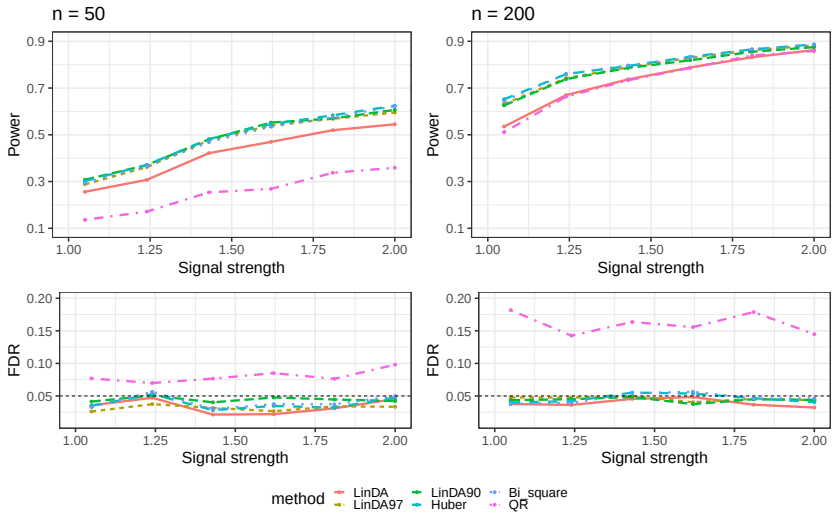

Figure 1. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a t-distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

The results for error terms generated from a Log-normal distribution (Case 2) are presented in Figure 2. Like the previous case, all methods exhibit an increased power to detect differential taxa with a larger sample size. In both small and large sample size scenarios, the Huber and Bi_square methods lead in power. LinDA90 follows, outperforming LinDA97. Notably, when the sample size is small, LinDA90 has an FDR inflation. Upon closer inspection, the Huber method outperforms the Bi_square method regarding power and FDR control. Although the QR method has higher power than LinDA when $n = 200$, it comes at the cost of considerable FDR inflation.

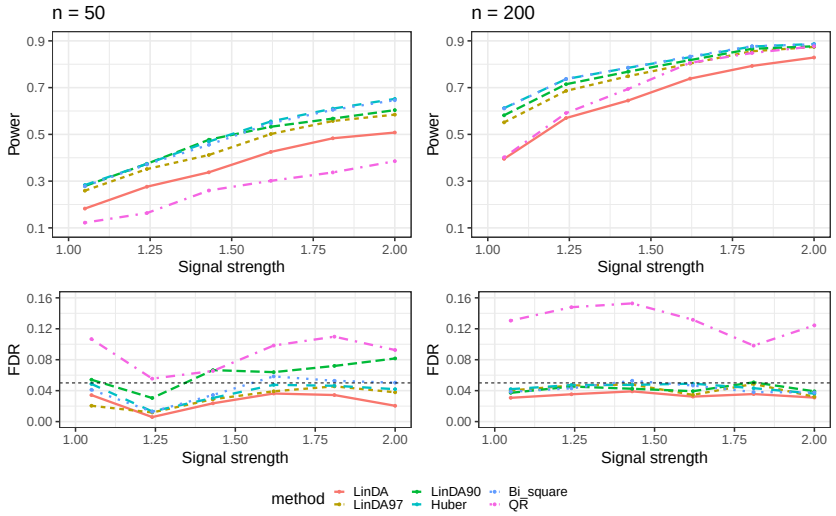

Figure 2. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a Log-normal distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

The outcomes for the setting in which error terms are sampled from a Weibull distribution (Case 3) are shown in Figure 3. As the sample size increases, all methods exhibit enhanced power in detecting differential taxa. When the sample size is small, all methods display limited power for detecting differential taxa, with a power of approximately 0.1. In contrast, with larger sample sizes, the Huber method outperforms all other methods, followed by the Bi_square and LinDA97 methods. Notably, Bi_square shows FDR inflation when the sample size is small, whereas the LinDA90 method exhibits significant FDR inflation with large sample sizes. Since the FDR of the QR method exceeds 0.3 when $n = 200$, it was excluded from the plot.

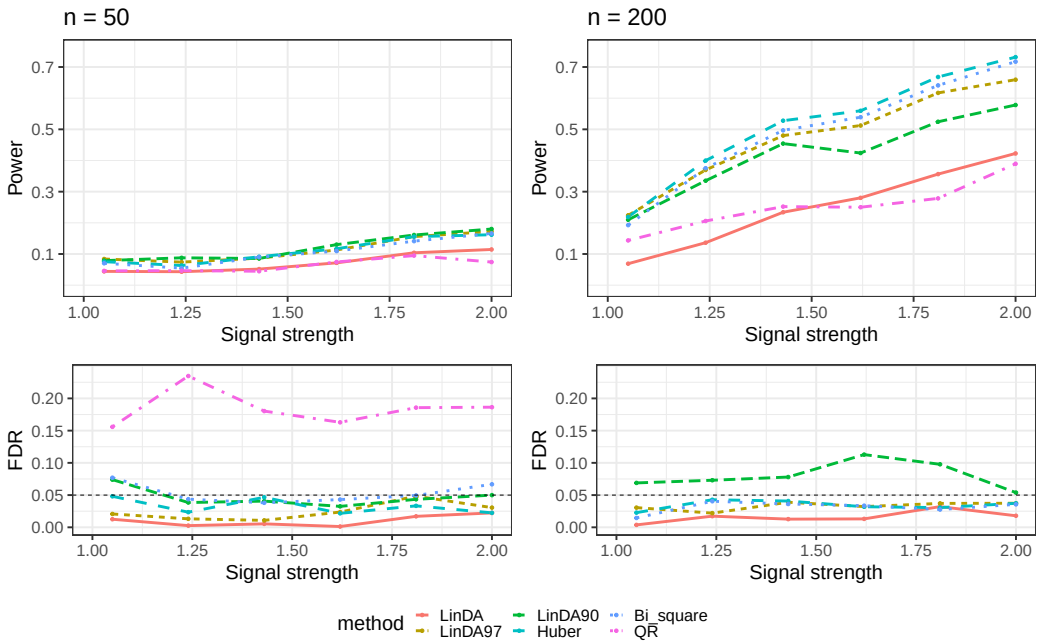

Figure 3. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a Weibull distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

3.1.2. With Outliers Setting

We employ the following procedure to generate the data to demonstrate the impact of the number of outliers on the power of the DAA method. The number of taxa is fixed at $m = 500$, and the number of samples is set to $n = 100$. Following the data-generation process discussed in Section 3.1, we begin by sampling error terms from a normal distribution with a mean of 0 and a variance of σ_{i*}^2 for the i-th taxon. The outliers are generated by randomly selecting a subset of nonzero counts and multiplying them by a fold change of 20. More specifically, we randomly choose 250, 500, 1000, and 2000 nonzero counts, corresponding to an average of 0.5, 1, 2, and 4 outliers per taxon, respectively.

Let ρ denote the average number of outliers per taxon. The results for the case when outliers exist without confounders are presented in Figure 4. As the number of outliers increases, the power of all methods will decrease. When the number of outliers is small ($\rho = 0.05$), LinDA97 exhibits slightly better performance across all methods. However, as the number of outliers increases, the performance of Huber and LinDA90 improves. Specifically, when $\rho = 4$, LinDA90 demonstrates the highest power for small signal strengths, while Huber exhibits the highest power for large signal strengths.

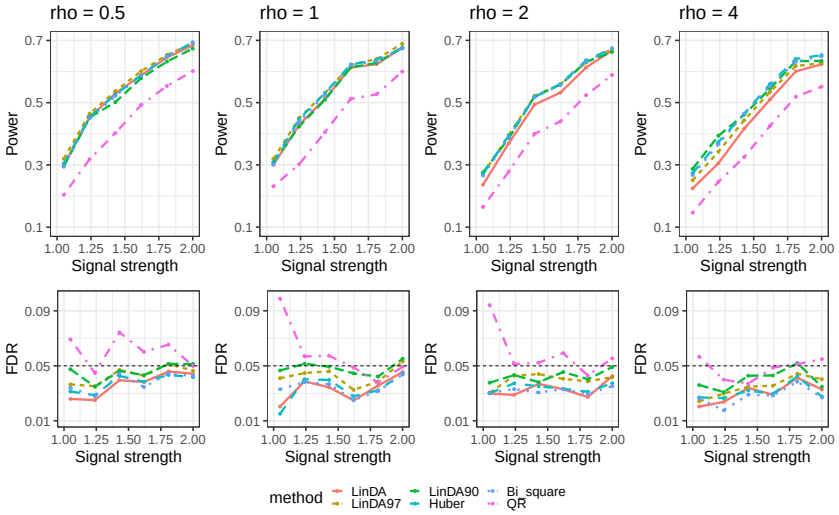

Figure 4. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$) with outliers. ρ is the average number of outliers per taxon.

4. Real Data Analysis

We utilized three real datasets, including independent samples from studies on *C. difficile* infection (CDI) [30], inflammatory bowel disease (IBD) [31], and rheumatoid arthritis (RA) [12]. The CDI and RA datasets were downloaded from the links provided in the original paper, while the IBD dataset was obtained from the Qiita database [32] using study IDs 1460 and 524. To ensure data quality, we excluded samples with less than 1000 read counts and taxa that appeared in less than 10% of the samples for each dataset. The variable that we are interested in is binary phenotypes across all datasets. In the case of the IBD dataset, the confounding factor is the usage of antibiotics.

First, we compared the detection power of all methods discussed in Section 3 across the three datasets. Figure 5 presents the number of discoveries at various FDR levels (0.01–0.25). For the CDI dataset, the LinDA97, LinDA, and Huber methods exhibited the highest number of discoveries at an FDR level of 0.1. As the FDR level increased, the Bi_square method identified the most discoveries. For the IBD dataset, all methods except QR had a similar number of discoveries at an FDR level of 0.1. However, the Bi_square and Huber methods outperformed others regarding discoveries at different FDR levels in the overall analysis. Conversely, the LinDA and Huber methods consistently identified the most discoveries across all FDR levels in the RA dataset. Therefore, the Huber method displayed superior discovery capability overall. It is worth mentioning that the QR method consistently yielded the fewest discoveries in all three datasets, which is consistent with our simulation results.

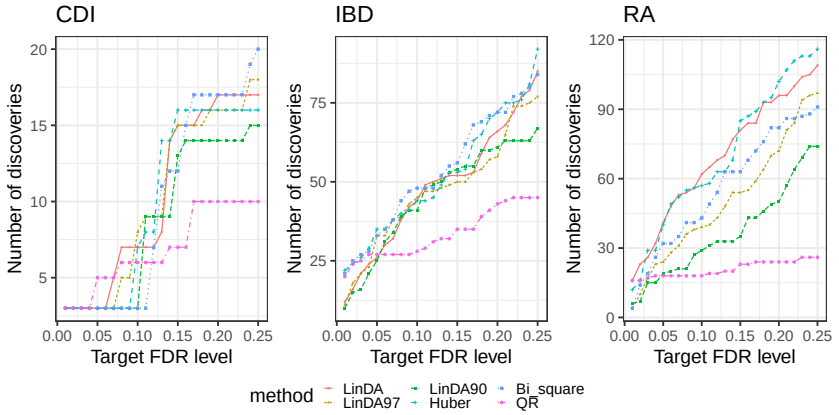

Figure 5. Number of discoveries with respect to different FDR levels (0.01–0.25) for three real datasets.

Subsequently, to illustrate the overlapping discoveries among different methods, we employed the UpSet plot [33] to depict the overlap at the target FDR level of 0.1. Figure 6 presents the overlap of differentially abundant taxa across the three real datasets at this FDR level. It is evident that, in most cases, the taxa identified by the Huber method are consistently identified by other methods as well. This consistency implies that the taxa identified by the Huber method are more likely to be "truly" differentially abundant. Conversely, the remaining methods exhibit independent findings, lacking support from other methods, suggesting a higher likelihood of false discoveries. Consequently, the Huber method demonstrates greater robustness in practical applications.

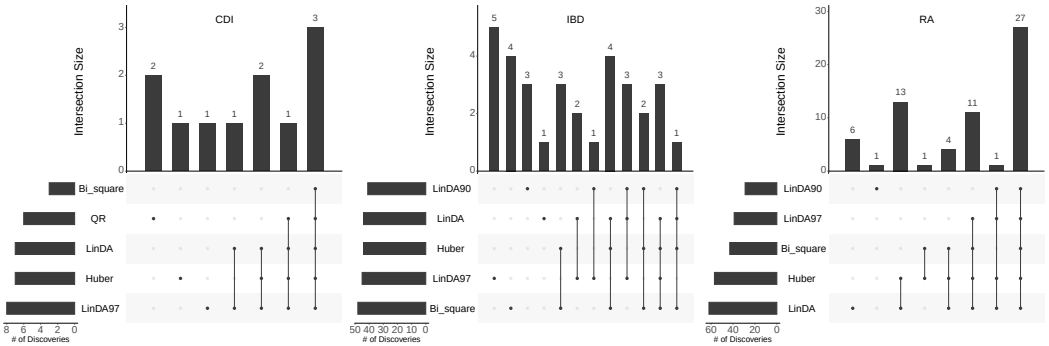

Figure 6. The overlap of differentially abundant taxa detected by various DAA methods across three real datasets at an FDR level of 0.1.

As an illustration, we present a boxplot of the taxa expression after applying the CLR transformation, focusing solely on taxa identified only by the LinDA method. Figure 7 displays an example of a taxa expression after CLR, identified only by the LinDA method in the RA dataset. The left panel shows the original data, while the right panel represents the data with potential outliers removed. In this case, we remove samples whose taxa expression after CLR exceeds 1.5. After removing these samples that significantly deviate from the rest, we observe similar means between the two groups. However, LinDA identifies this taxon as differentially abundant due to outliers, which may be a false discovery.

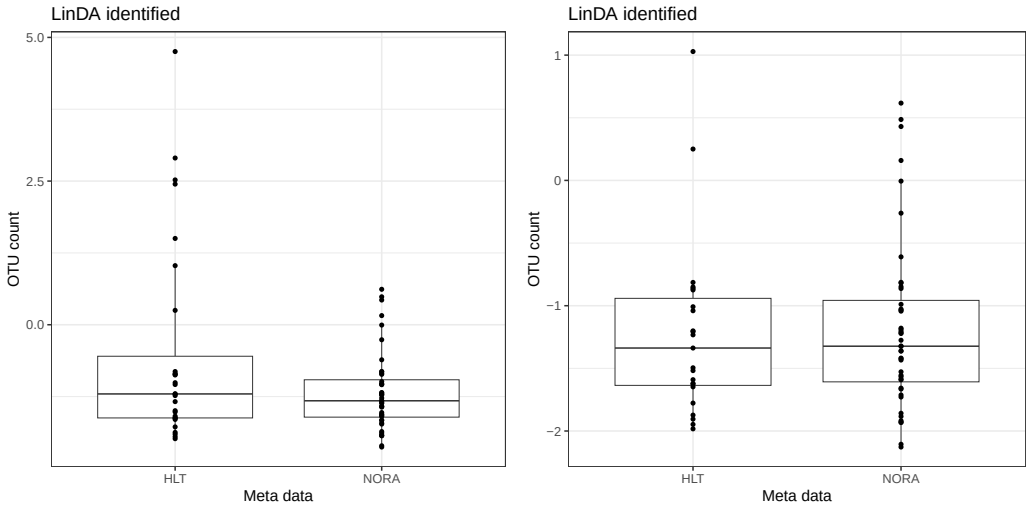

Figure 7. An example of a taxa expression after CLR. The taxa are only identified by the LinDA method in the RA dataset. The left panel shows the original data. The right panel represents the data with potential outliers removed.

5. Conclusions and Discussion

This study investigates the influence of heavy-tailedness and outliers on differential abundance analysis using different methods and proposes effective strategies to address them. The presence of heavy-tailedness and outliers can substantially reduce the power of existing DAA methods. To resolve this issue, we propose two techniques to enhance the robustness of DAA methods. First, we introduce a general regression framework for DAA by extending the LinDA method. This framework includes differential analysis based on Huber regression as a special case, which is more stable in handling outliers. Second, we propose the winsorization method, which involves winsorizing the data before applying DAA methods.

Our findings from both simulation and real data indicate that, among the M-estimation-based approaches, the Huber method, in particular, demonstrates more robust performance in handling heavy-tailedness and the presence of outliers in the dataset. Specifically, compared to LinDA with winsorization, the Huber method exhibits higher power while maintaining FDR control. Furthermore, the Huber method demonstrates better FDR control than the Bi_square method. Consequently, we recommend utilizing the Huber method as an effective approach to address outliers and heavy-tailedness. Additionally, the optimal quantile level for winsorization remains unclear. Therefore, another advantage of the Huber method is that it does not require the selection of such hyperparameters.

Author Contributions: Conceptualization, J.C. and X.Z.; methodology, G.L. and X.Z.; software, G.L.; formal analysis, G.L., J.C. and X.Z.; real data, G.L., L.Y. and J.C.; writing—original draft preparation, G.L.; writing—review and editing, G.L. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Institute of Health R01GM144351 (Chen and Zhang), National Science Foundation DMS-1830392, DMS2113359, DMS1811747 (Zhang and Li), and National Science Foundation DMS2113360 and Mayo Clinic Center for Individualized Medicine (Chen).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The function for implementing Robust DAA is available at https://github.com/guanxunli/robustDAA (accessed on 24 June 2023). The code for running the simulation is available at https://github.com/guanxunli/robust_daa (accessed on 23 October 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Numerical Results for Log-Linear Model

Appendix A.1. Errors Generated Form Normal Distribution

To ensure the completeness of our study, we conducted tests on all methods by sampling errors from a normal distribution in the absence of outliers. The results for scenarios without confounders are presented in Figure A1.

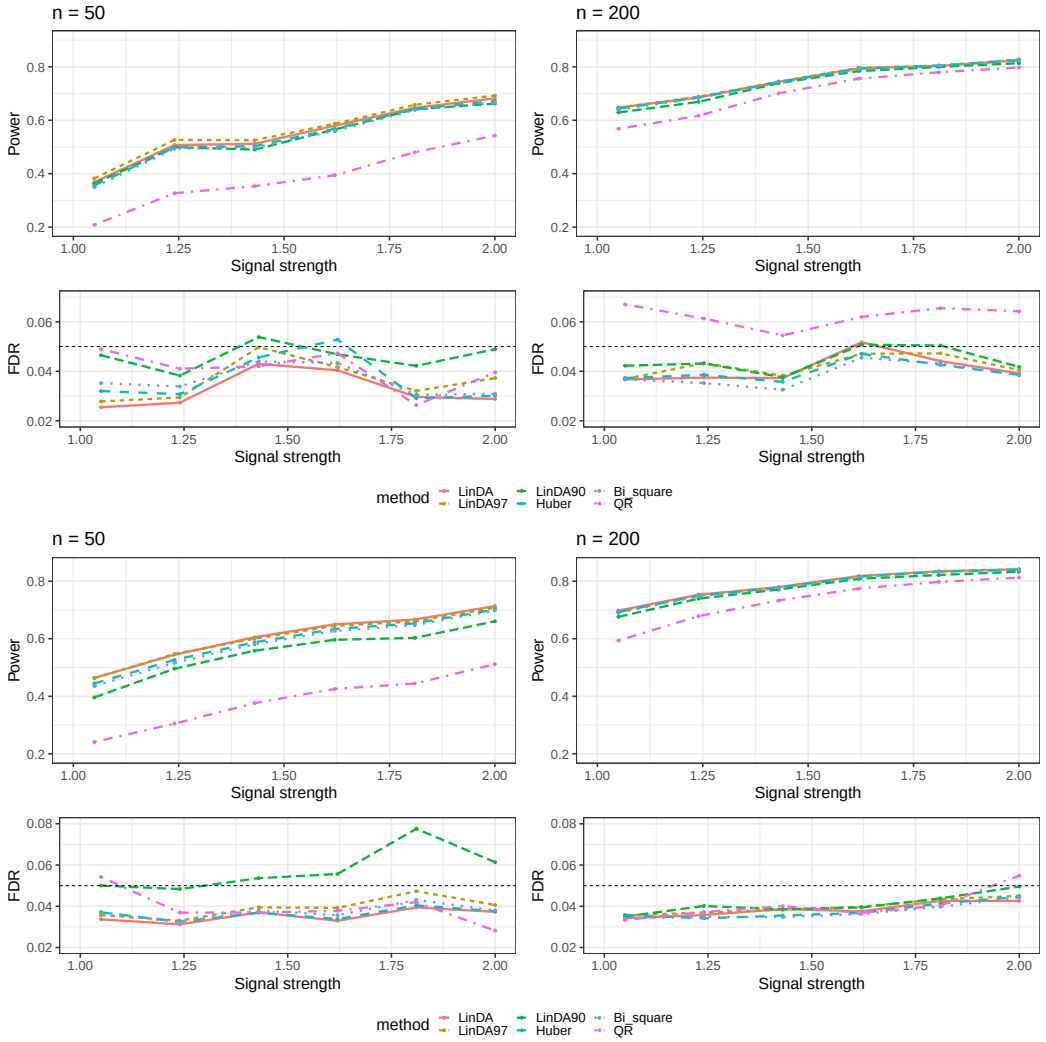

Figure A1. Results for the log-linear model with errors generated from a normal distribution without confounders. The left column corresponds to a sample size of $n = 50$ and the right column corresponds to a sample size of $n = 200$. The first group of figures shows results in the sparse-signal setting and the second group of figures shows results in the dense-signal setting.

Appendix A.2. Numerical Results for Dense-Signal Setting

Figure A2 shows the results for error terms generated from t-distribution; Figure A3 shows the results for error terms generated from Weibull distribution; Figure A4 shows the results for the case with outliers.

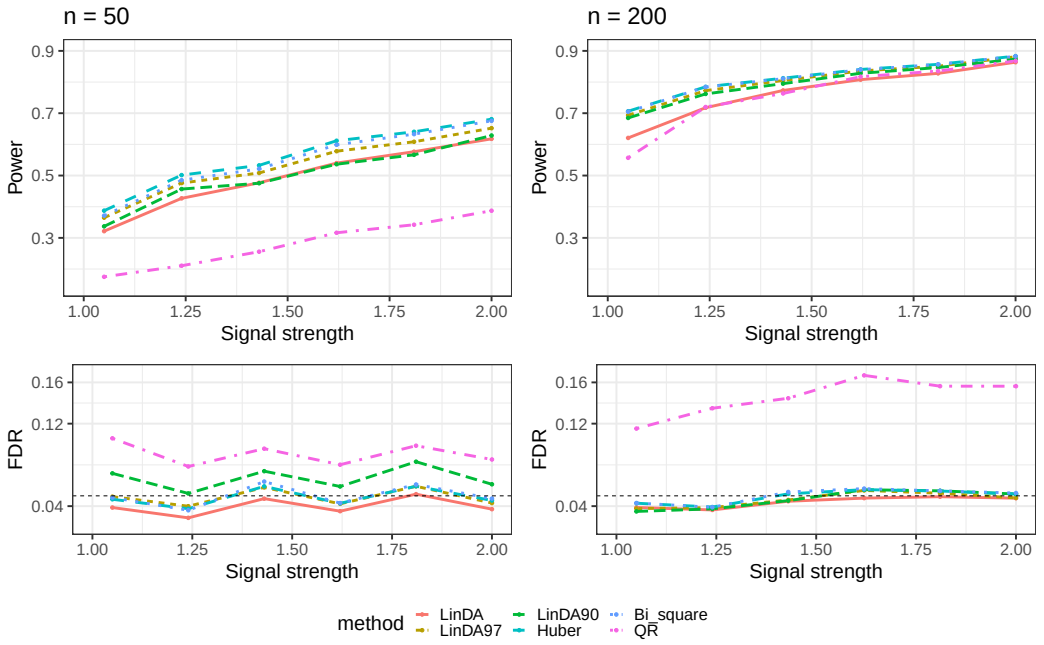

Figure A2. The log-linear model yielded results in the absence of confounding variables in the dense-signal setting ($p_\gamma = 0.2$), where errors were generated from a t-distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

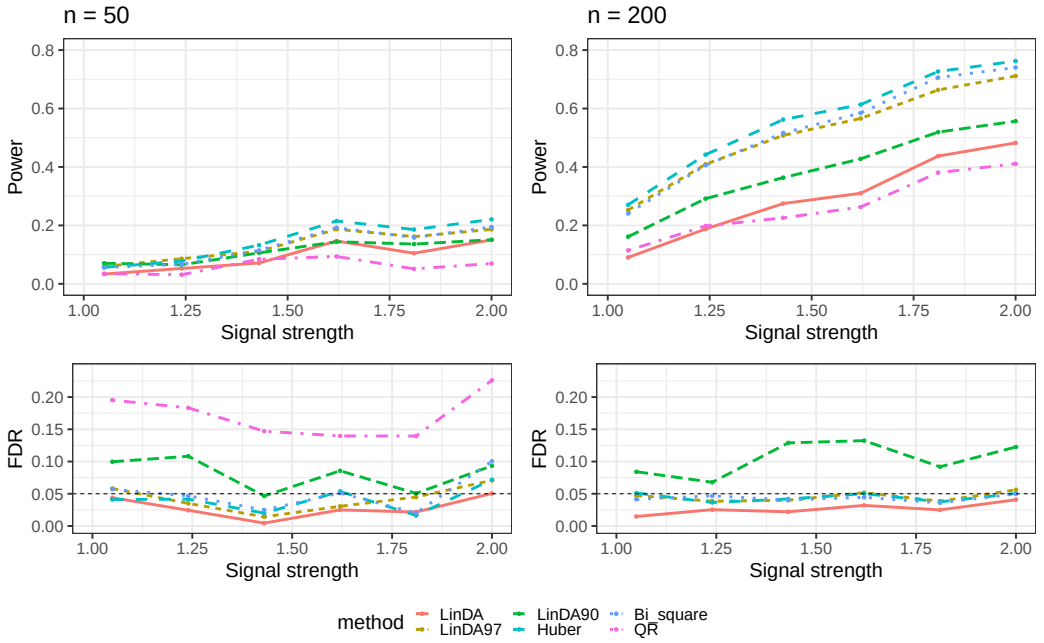

Figure A3. The log-linear model yielded results in the absence of confounding variables in the dense-signal setting ($p_\gamma = 0.2$), where errors were generated from a Weibull distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

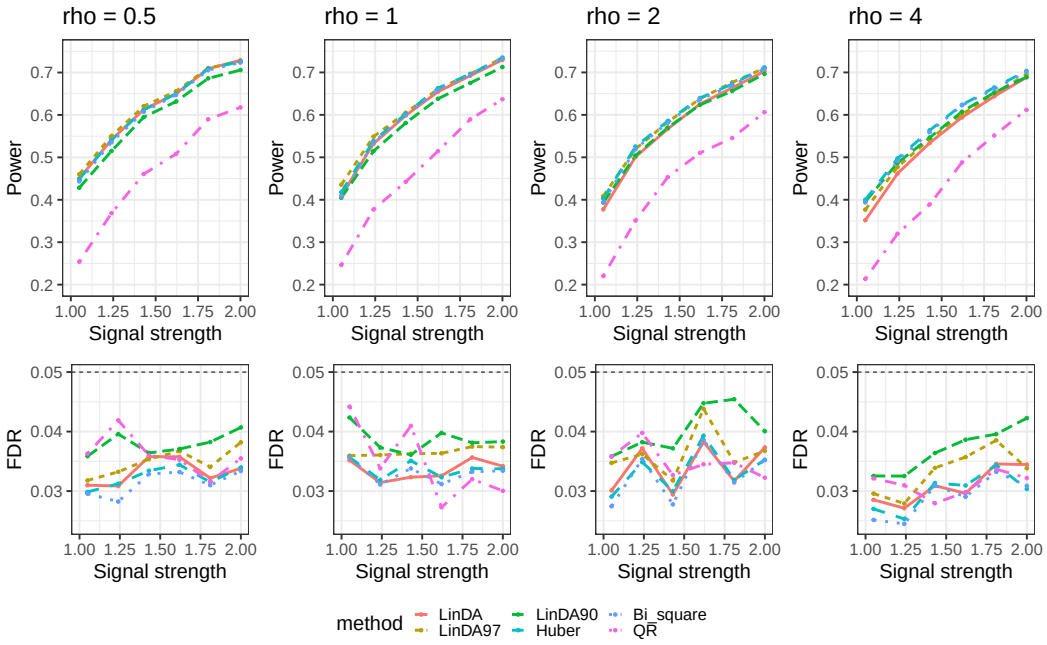

Figure A4. The log-linear model yielded results in the absence of confounding variables in the dense-signal setting ($p_\gamma = 0.2$) with outliers. ρ is the average number of outliers per taxon.

Appendix A.3. Numerical Results with Confounders

Figure A5 shows the results for error terms generated from *t*-distribution; Figure A6 shows the results for error terms generated from Weibull distribution; Figure A7 shows the results for the case with outliers.

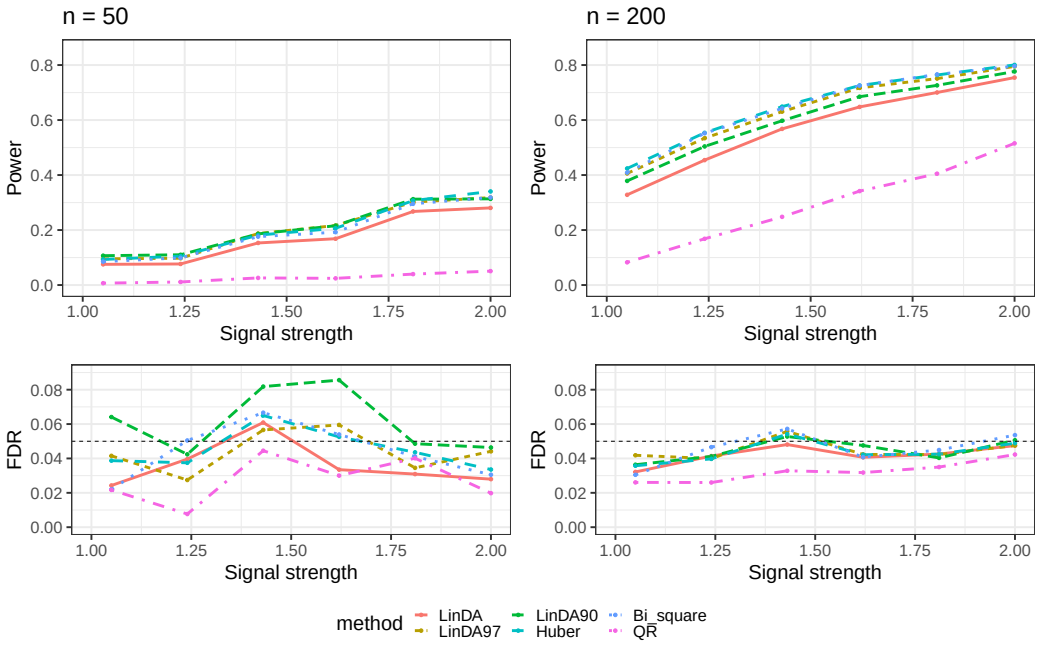

Figure A5. The log-linear model yielded results with confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a *t*-distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

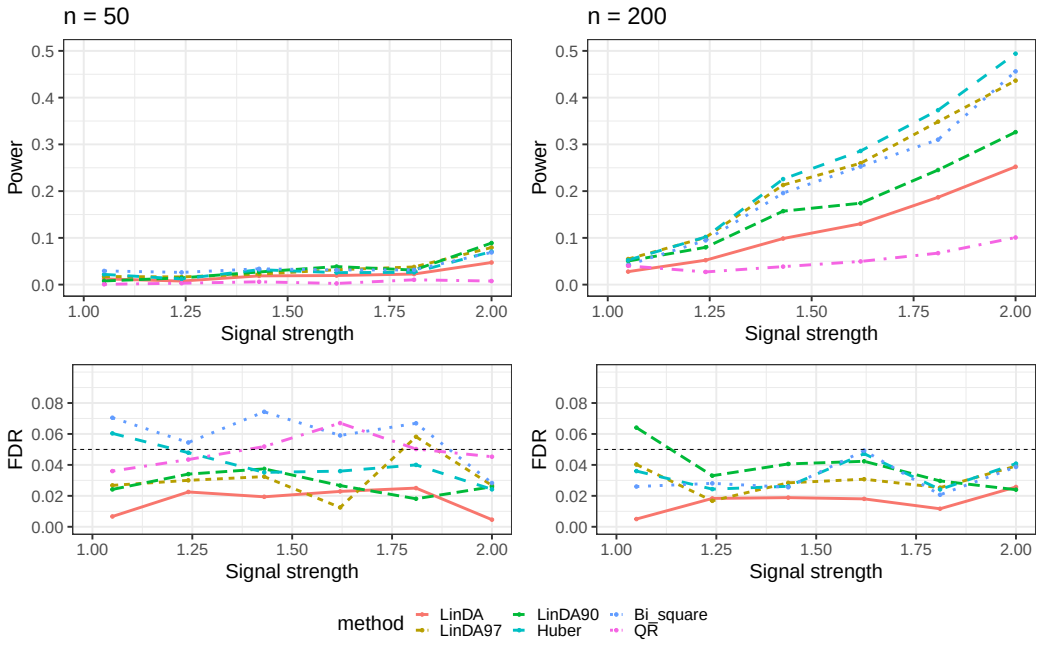

Figure A6. The log-linear model yielded results with confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a Weibull distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

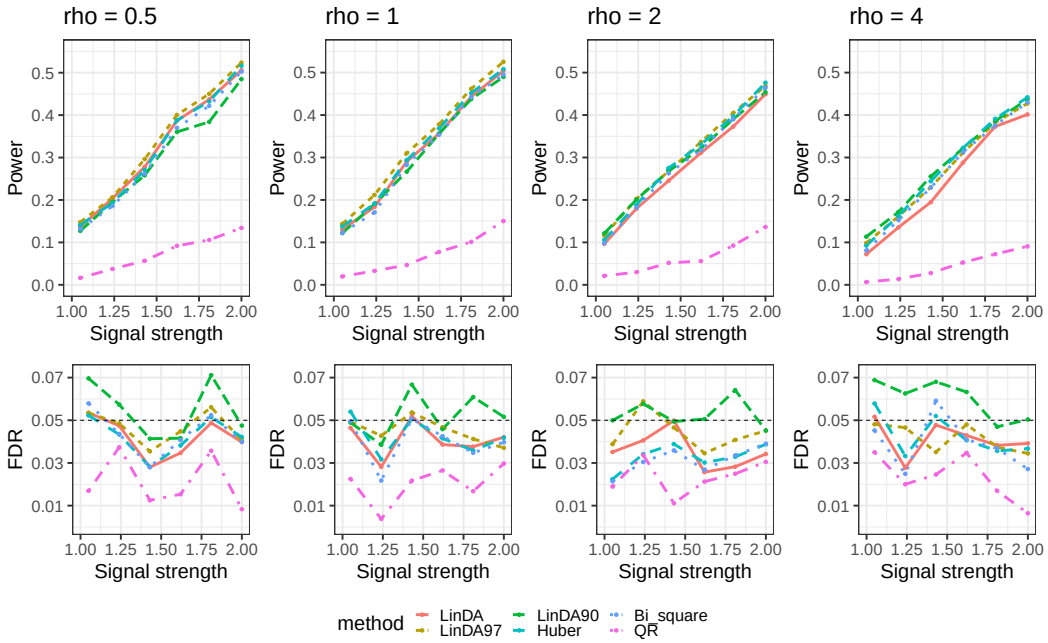

Figure A7. The log-linear model yielded results with confounding variables in the sparse-signal setting ($p_\gamma = 0.05$) with outliers. ρ is the average number of outliers per taxon.

Appendix B. Mixed-Effects Models

Based on the regression-based framework, our method can be readily extended to apply to the mixed-effect model. Consider the linear mixed-effect model:

$$\log(X_{is}) = u_s a_i + \mathbf{c}_s^\top \boldsymbol{\beta}_i + \mathbf{v}_s^\top \boldsymbol{\gamma}_i + \epsilon_{is},$$

where $\boldsymbol{\gamma}_i$ represents the random effect and $\mathbf{v}_s$ represents the corresponding design matrix. Mixed-effects analysis can be used to analyze correlated microbiome data from studies in-

volving replicates or spatial sampling, as well as family-based and longitudinal microbiome studies. Similarly, we need to address the regression problem:

$$W_{is} = u_s(a_i - \bar{a}) + \mathbf{c}_s^\top(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}) + \mathbf{v}_s^\top(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}}) + \epsilon_{is} - \bar{\epsilon}_s,$$

where any method for dealing with linear mixed effects can be applied to obtain the estimate here.

For robustly fitting the linear mixed-effects models, we utilize the *rlmer* function provided by the R package *robustlmm* (version: 3.2-0) [34], which is a Huber loss-based method. Furthermore, we use the *get_Lb_ddf* function available in the R package *pbkrtest* (version: 0.5.2) to compute the degrees of freedom using the Kenward–Roger approximation [35]. Due to the considerable computational time required for fitting a robust linear mixed-effects model, we opted to select specific hyperparameters. Following the recommendation by Koller [34], we chose $c = 1.345$ and $c = 2.28$. Subsequently, we employed the Cauchy method to combine the p-values.

To evaluate the performance of our proposed method, we employed the data generation process proposed by Zhou et al. [18] but used different methods to generate the random effect terms and the error terms. Consider the scenario involved in replicate sampling, where we generated 50 subjects with 4 replicates, for $n = 200$. The absolute abundance X_{is} was generated by

$$\log(X_{is}) = \beta_{i0} + u_s\alpha_i + \omega_{is} + \epsilon_{is},$$

where ω_{is} is the random effect and ϵ_{is} is the error term.

Appendix B.1. Heavy-Tailedness Setting

In the heavy-tailedness setting, the random effect terms and error terms of the mixed-effects model were evaluated in two different settings, corresponding to Case 2 and 3, as described in Section 3.1.

1. Case1: We sampled ω_{is} and ϵ_{is} from Log-normal distribution with log mean parameter of 0 for both terms and log standard deviation parameter of 0.5 and 0.8, respectively. We recentered the samples so that it has a zero mean.
2. Case2: We sampled ω_{is} and ϵ_{is} from a Weibull distribution with scale parameter of 0.3 for both terms and shape parameter of 0.8 and 0.5, respectively. We recentered the samples so that it has a zero mean.

For all three cases, the observed OTUs were generated following the same method as the log-linear model described in Section 3.1. Both sparse-signal ($p_\gamma = 0.05$) and dense-signal ($p_\gamma = 0.2$) settings were considered.

Figure A8 illustrates the outcomes of the mixed-effects model, with random effects and errors generated from a Log-normal distribution. In both the sparse and dense-signal settings, the Huber method displays the highest power, while maintaining absence of FDR inflation. In the sparse-signal setting, the power of LinDA90 outperforms that of LinDA97, while both methods exhibit comparable performance in the dense-signal setting.

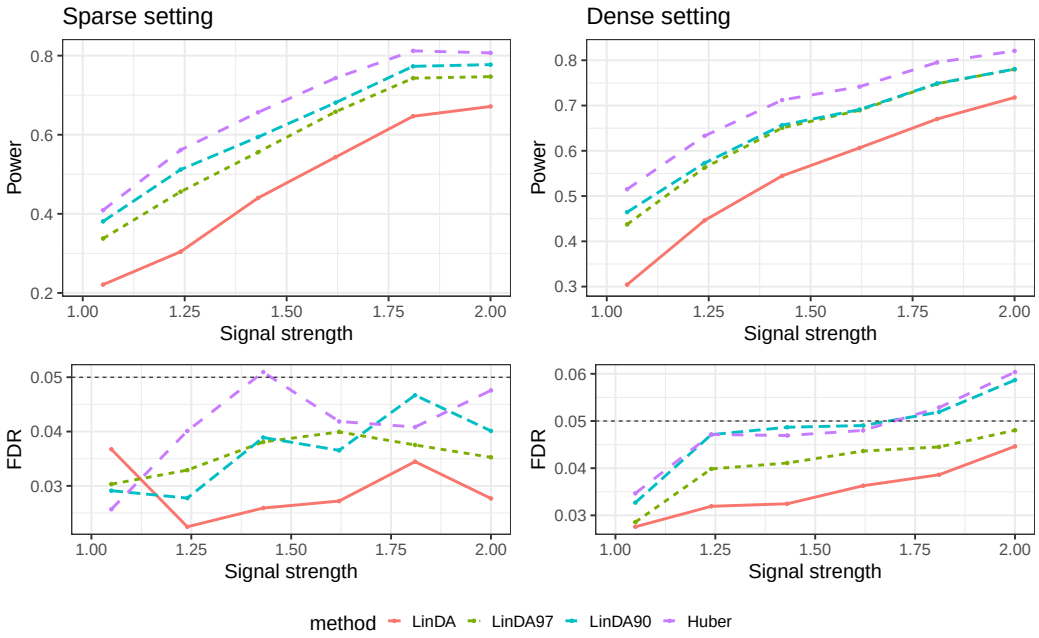

Figure A8. Results for the mixed-effects model where random effects and errors were generated from a Log-normal distribution. The left panel represents the sparse-signal setting and the right panel corresponds to dense-signal setting.

Figure A9 gives the results of the mixed-effects model, where the random effects and errors were generated from a Weibull distribution. In scenarios with low signal strength, LinDA97 demonstrates the best performance, while Huber surpasses all methods as the signal strength increases. Notably, LinDA90 is the only method exhibiting FDR inflation in this particular setting.

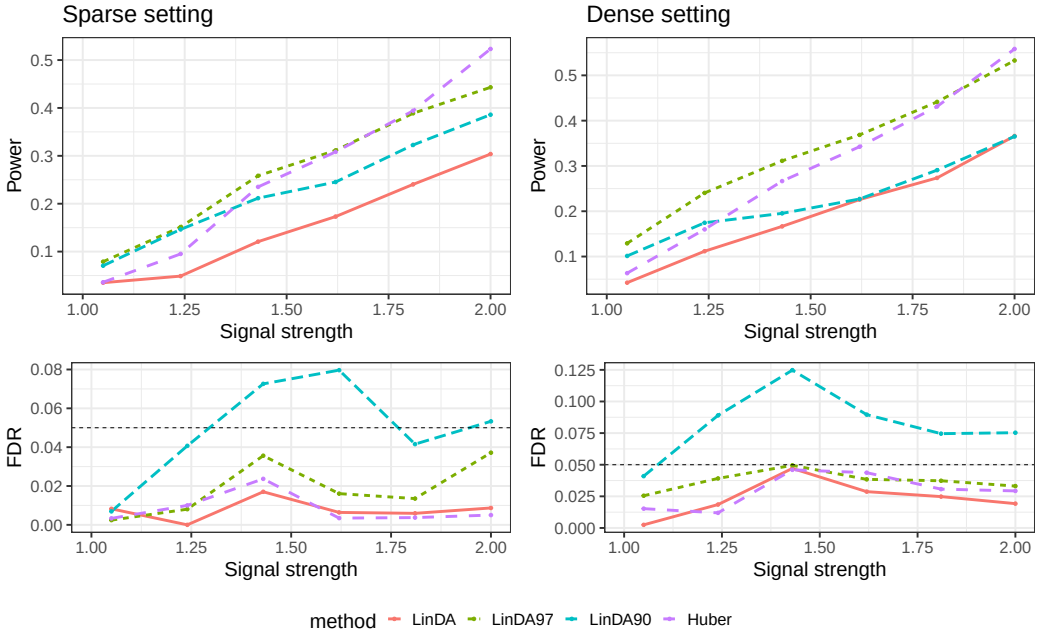

Figure A9. Results for the mixed-effects model where random effects and errors were generated from a Weibull distribution. The left panel represents the sparse-signal setting and the right panel corresponds to dense-signal setting.

Appendix B.2. With Outliers Setting

In the setting with outliers, ω_{is} was generated from a normal distribution with a mean of 0 and a variance of τ_i, where $\tau_i = a_i\sigma_{i*}^2$ and a_i follows a uniform distribution in the range of $[0,\ 1]$. The error terms ϵ_{is} were sampled from a normal distribution with a mean of 0 and a variance of σ_{i*}^2. The outliers are generated by the same way described in Section 3.1, corresponding to an average of 1, 2, 4, and 8 outliers per taxon, respectively.

Let ρ denote the average number of outliers per taxon. The results for the sparse case and dense case are presented in Figures A10 and A11, respectively. As the number of outliers increase, the power of all methods will decrease. When the number of outliers is small ($\rho = 1$), LinDA97 exhibits slightly better performance across all methods. However, as the number of outliers increases, the performance of Huber method improves. Specifically, when $\rho = 4$, Huber exhibits the highest power over all methods.

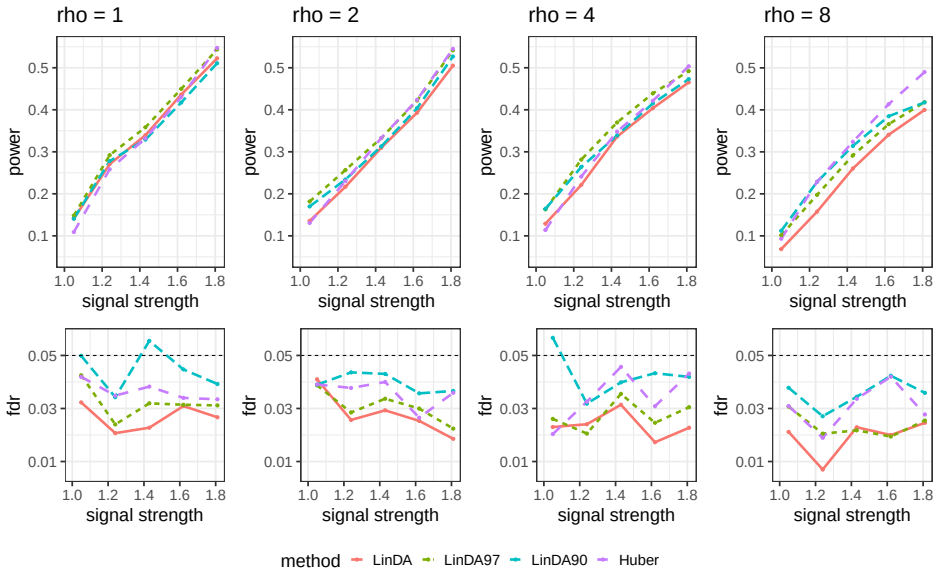

Figure A10. Results for the mixed-effects model with outliers in the sparse setting. ρ is the average number of outliers per taxon.

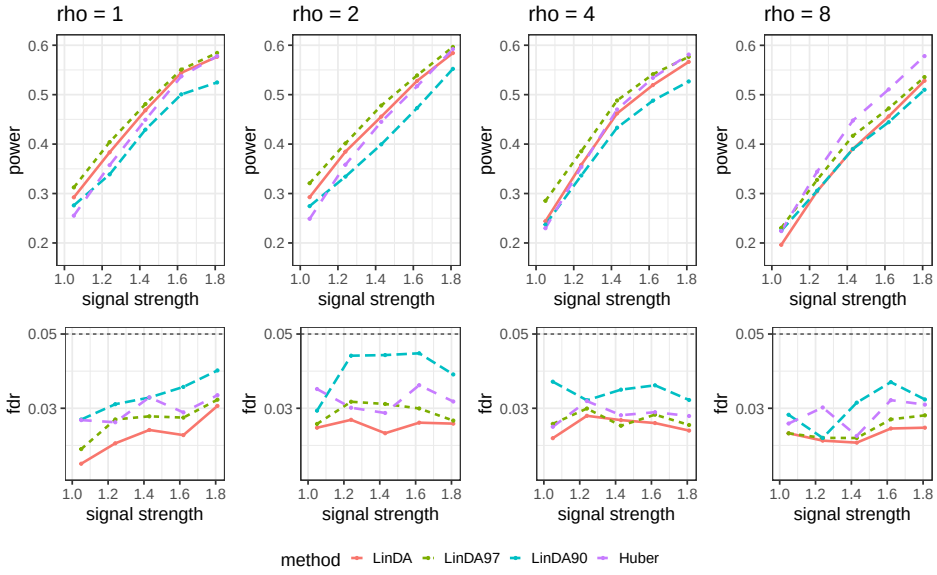

Figure A11. Results for the mixed-effects model with outliers in the dense setting. ρ is the average number of outliers per taxon.

Appendix C. Additional Numerical Results for Comparison with Other Methods

In this section, we compared our method with other differential abundance analysis methods, including ALDEx2 [15], ANCOM-BC [16], and MaAsLin2 [17]. We focus on the simulation settings based on log-linear models we present in Section 3.1.

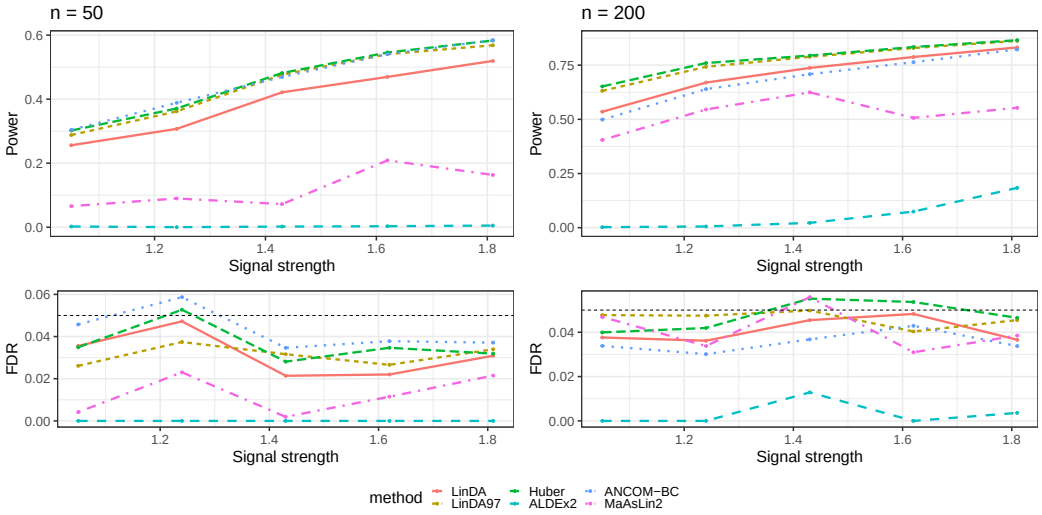

Figure A12. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a t-distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

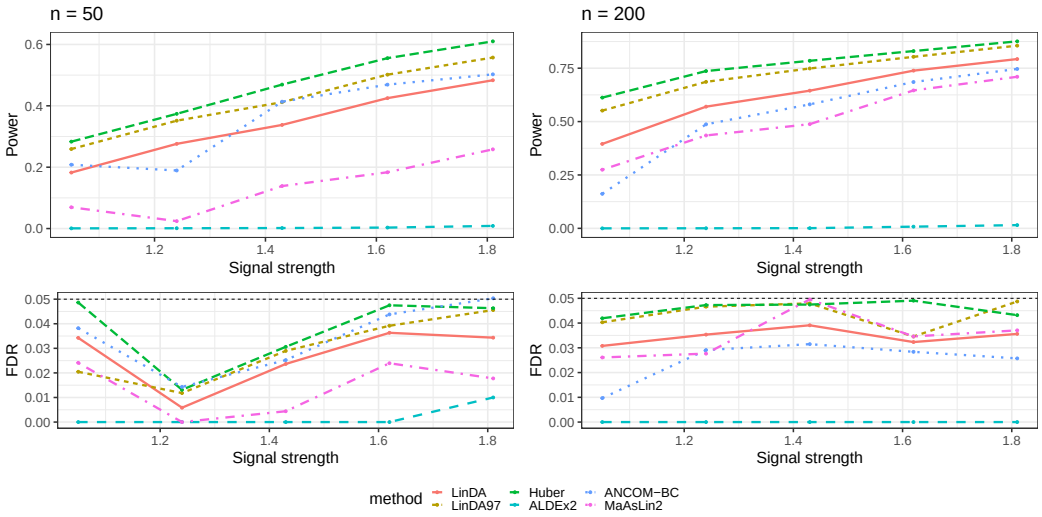

Figure A13. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a Log-normal distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

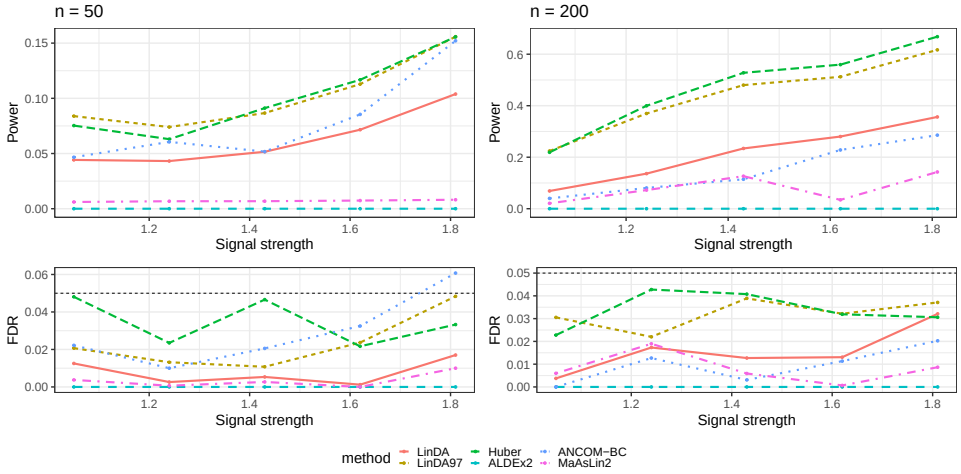

Figure A14. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$), where errors were generated from a Weibull distribution. The left panel represents a sample size of $n = 50$ and the right panel corresponds to a sample size of $n = 200$.

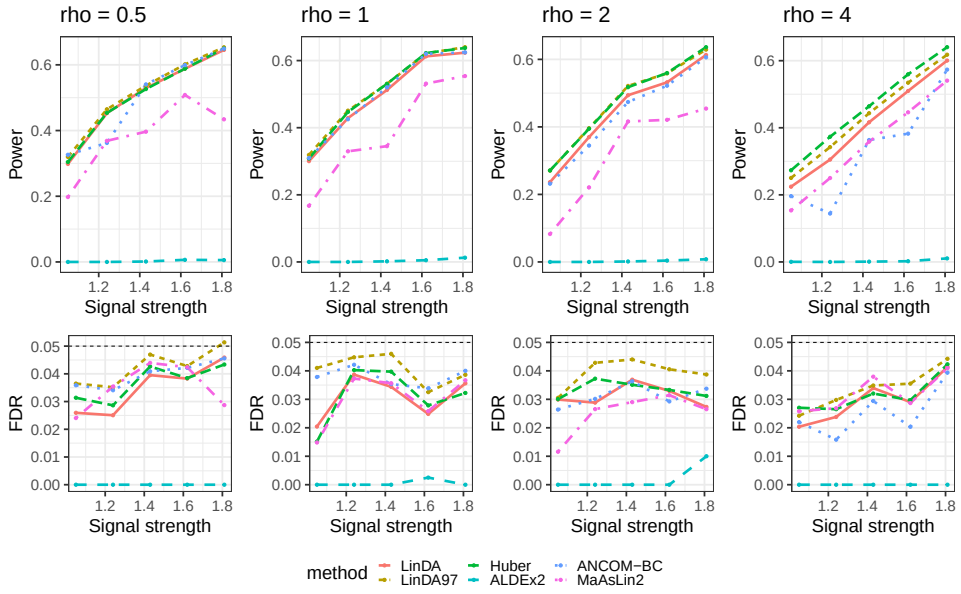

Figure A15. The log-linear model yielded results in the absence of confounding variables in the sparse-signal setting ($p_\gamma = 0.05$) with outliers. ρ is the average number of outliers per taxon.

References

1. Cho, I.; Blaser, M.J. The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* **2012**, *13*, 260–270. [CrossRef] [PubMed]
2. Valdes, A.M.; Walter, J.; Segal, E.; Spector, T.D. Role of the gut microbiota in nutrition and health. *BMJ* **2018**, *361*, k2179. [CrossRef] [PubMed]
3. Knights, D.; Lassen, K.G.; Xavier, R.J. Advances in inflammatory bowel disease pathogenesis: Linking host genetics and the microbiome. *Gut* **2013**, *62*, 1505–1510. [CrossRef] [PubMed]
4. Fan, Y.; Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **2021**, *19*, 55–71. [CrossRef] [PubMed]
5. Kuczynski, J.; Lauber, C.L.; Walters, W.A.; Parfrey, L.W.; Clemente, J.C.; Gevers, D.; Knight, R. Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **2012**, *13*, 47–58. [CrossRef] [PubMed]
6. Truong, D.T.; Franzosa, E.A.; Tickle, T.L.; Scholz, M.; Weingart, G.; Pasolli, E.; Tett, A.; Huttenhower, C.; Segata, N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **2015**, *12*, 902–903. [CrossRef] [PubMed]
7. Tsilimigras, M.C.; Fodor, A.A. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Ann. Epidemiol.* **2016**, *26*, 330–335. [CrossRef] [PubMed]

8. Morton, J.T.; Marotz, C.; Washburne, A.; Silverman, J.; Zaramela, L.S.; Edlund, A.; Zengler, K.; Knight, R. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **2019**, *10*, 2719. [CrossRef]
9. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
10. Yang, L.; Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: Current status and potential solutions. *Microbiome* **2022**, *10*, 130. [CrossRef]
11. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25. [CrossRef] [PubMed]
12. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Nat. Preced.* **2010**, 1. [CrossRef]
13. Paulson, J.N.; Stine, O.C.; Bravo, H.C.; Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **2013**, *10*, 1200–1202. [CrossRef] [PubMed]
14. Chen, L.; Reeve, J.; Zhang, L.; Huang, S.; Wang, X.; Chen, J. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **2018**, *6*, e4600. [CrossRef] [PubMed]
15. Fernandes, A.D.; Reid, J.N.; Macklaim, J.M.; McMurrough, T.A.; Edgell, D.R.; Gloor, G.B. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2014**, *2*, 15. [CrossRef] [PubMed]
16. Lin, H.; Peddada, S.D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **2020**, *11*, 3514. [CrossRef] [PubMed]
17. Mallick, H.; Rahnavard, A.; McIver, L.J.; Ma, S.; Zhang, Y.; Nguyen, L.H.; Tickle, T.L.; Weingart, G.; Ren, B.; Schwager, E.H.; et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **2021**, *17*, e1009442. [CrossRef] [PubMed]
18. Zhou, H.; He, K.; Chen, J.; Zhang, X. LinDA: Linear models for differential abundance analysis of microbiome compositional data. *Genome Biol.* **2022**, *23*, 95. [CrossRef]
19. Montassier, E.; Al-Ghalith, G.A.; Hillmann, B.; Viskocil, K.; Kabage, A.J.; McKinlay, C.E.; Sadowsky, M.J.; Khoruts, A.; Knights, D. CLOUD: A non-parametric detection test for microbiome outliers. *Microbiome* **2018**, *6*, 137. [CrossRef]
20. Chen, J.; King, E.; Deek, R.; Wei, Z.; Yu, Y.; Grill, D.; Ballman, K. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* **2018**, *34*, 643–651. [CrossRef]
21. Nearing, J.T.; Douglas, G.M.; Hayes, M.G.; MacDonald, J.; Desai, D.K.; Allward, N.; Jones, C.M.; Wright, R.J.; Dhanani, A.S.; Comeau, A.M.; et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* **2022**, *13*, 342. [CrossRef] [PubMed]
22. Huber, P.J. Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799–821. [CrossRef]
23. Dixon, W.J.; Yuen, K.K. Trimming and winsorization: A review. *Stat. Hefte* **1974**, *15*, 157–170. [CrossRef]
24. Kimura, D.K. Analyzing relative abundance indices with log-linear models. *N. Am. J. Fish. Manag.* **1988**, *8*, 175–180. [CrossRef]
25. Rivest, L.P.; Lévesque, T. Improved log-linear model estimators of abundance in capture-recapture experiments. *Can. J. Stat.* **2001**, *29*, 555–572. [CrossRef]
26. Fox, J.; Weisberg, S. Robust regression. *R S-Plus Companion Appl. Regres.* **2002**, *91*, 6.
27. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000; Volume 3.
28. Liu, Y.; Chen, S.; Li, Z.; Morrison, A.C.; Boerwinkle, E.; Lin, X. ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **2019**, *104*, 410–421. [CrossRef] [PubMed]
29. Fan, J.; Li, Q.; Wang, Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 247. [CrossRef]
30. Schubert, A.M.; Rogers, M.A.; Ring, C.; Mogle, J.; Petrosino, J.P.; Young, V.B.; Aronoff, D.M.; Schloss, P.D. Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *mBio* **2014**, *5*, e01021-14. [CrossRef]
31. Morgan, X.C.; Tickle, T.L.; Sokol, H.; Gevers, D.; Devaney, K.L.; Ward, D.V.; Reyes, J.A.; Shah, S.A.; LeLeiko, N.; Snapper, S.B.; et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **2012**, *13*, R79. [CrossRef]
32. Gonzalez, A.; Navas-Molina, J.A.; Kosciolek, T.; McDonald, D.; Vázquez-Baeza, Y.; Ackermann, G.; DeReus, J.; Janssen, S.; Swafford, A.D.; Orchanian, S.B.; et al. Qiita: Rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **2018**, *15*, 796–798. [CrossRef]
33. Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; Pfister, H. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1983–1992. [CrossRef]
34. Koller, M. robustlmm: An R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* **2016**, *75*, 1–24. [CrossRef]
35. Halekoh, U.; Højsgaard, S. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package pbkrtest. *J. Stat. Softw.* **2014**, *59*, 1–32. [CrossRef]