Log-ratio analysis of microbiome data with many zeroes is library size dependent

Dennis E. te Beest[1a], Els H. Nijhuis[2], T.W.R. Möhlmann[3], Cajo J.F. ter Braak[1]

Running title: Log-ratio analysis of microbiome data.

[1] Biometris, Wageningen University & Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

[2] Biointeractions and Plant Health, Wageningen University & Research, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

[3] Laboratory of Entomology, Wageningen University & Research, P.O. box 16, 6700 AA, Wageningen, The Netherlands

[a] Corresponding author

1 Abstract

Microbiome composition data collected through amplicon sequencing are count data on taxa in which the total count per sample (the library size) is an artifact of the sequencing platform and as a result such data are compositional. To avoid library size dependency, one common way of analyzing multivariate compositional data is to perform a principal component analysis (PCA) on data transformed with the centered log-ratio, hereafter called a log-ratio PCA. Two aspects typical of amplicon sequencing data are the large differences in library size and the large number of zeroes. In this paper we show on real data and by simulation that, applied to data that combines these two aspects, log-ratio PCA is nevertheless heavily dependent on the library size. This leads to a reduction in power when testing against any explanatory variable in log-ratio redundancy analysis. If there is additionally a correlation between the library size and the explanatory variable, then the type 1 error becomes inflated. We explore putative solutions to this problem.

Keywords: Microbiome, Multivariate statistics, Zero-inflation, Log-ratio analysis.

2 Introduction

Microbiome composition data collected through amplicon sequencing are count data on taxa in which the total count per sample (the library size) is a technical, ill-understood artifact which carries no biological information and as a result such data are compositional. Some people have advocated the use of compositional data analyses in analyzing such data (Tsilimigras & Fodor, 2016; Gloor et al., 2017). For multivariate analysis this implies transforming the data with the centered log-ratio transformation (clr) followed by a standard least-squares method such as principal component analysis (PCA). Equivalently, the data (counts or proportions) are logarithmically transformed and double centered, followed by a PCA. This is often called log-ratio PCA or log-ratio analysis (Aitchison, 1983; Greenacre, 2018). Mathematically this is a solid approach when there are no zeroes, as it takes care of the arbitrary total per sample by only analyzing log-ratios (Greenacre, 2018). However, with zeroes, a pseudo-count must be added before taking the log-transformation.

Two aspects typical for amplicon sequencing data complicate the use of log-ratio PCA: the high amount of zeroes combined with a large variability in the library size. In this paper we show that using log-ratio PCA on such data has the unexpected and unwanted consequence that the library size again influences the analysis. In an unconstrained analysis (PCA) it is possible that the 1^{st} or 2^{nd} axes primarily display the library size. In an constrained analysis (e.g. log-ratio redundancy analysis (RDA) (van den Wollenberg, 1977; ter Braak, 1994) this effect complicates the assessment of significance of explanatory variables.

The primary aim of this paper is to make people aware of this problem of using log-ratio analysis and the clr transformation on amplicon sequencing data. We provide a mathematical explanation and illustrate the issue with simulated data and two amplicon sequencing data examples. We additionally explore some putative solutions.

3 Materials and methods

3.1 Log-ratio PCA

With the aim to compare samples, log-ratio PCA decomposes $\mathbf{Y}$, a matrix that contains compositional data with I samples (rows) and J taxa (columns), to a set of principal axes (Aitchison, 1983; Greenacre, 2018). We define $\mathbf{L} = \{l_{ij}\}$ as the log of $\mathbf{Y}$, $\mathbf{r}$ as the marginal mean of the rows of $\mathbf{L}$, and $\mathbf{c}$ as the

marginal mean of the columns of $\mathbf{L}$. The log-ratio transformation (clr) is defined as $log(y_i/g(y_i))$ (where y_i is the i-th row of $\mathbf{Y}$ and $g()$ is the geometric mean), which is equivalent to $l_{ij} - r_i$. Given that in a decomposition to principal axes we also need to center by taxa (columns), a log-ratio PCA involves double centering of $\mathbf{L}$.

$$s_{ij} = l_{ij} - r_i - c_j + l_{..} \tag{1}$$

where $l_{..}$ stands for the global mean of $\mathbf{L}$. The centered matrix $\mathbf{S} = \{s_{ij}\}$ can be decomposed with a singular value decomposition (SVD):

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

Matrix $\mathbf{U}$, of size $I \times K$, contains the "sample scores" where K stands for the number of latent dimensions. Matrix $\mathbf{V}$ is of size $J \times K$ and contains the "taxon scores". Matrix $\mathbf{\Sigma}$ is diagonal matrix with singular values (Greenacre, 2018, 2012). The main focus in our analyses is on comparing the sample scores.

3.2 Zeroes lead to library size dependence: mathematical explanation

A large number of zeroes combined with a large variability in library size, and thus in $\mathbf{r}$, creates a problem for log-ratio PCA. For count data it is common to add a pseudo count of 1. This preserves the zeroes and the sparsity of the data, and avoids needing to take the log of zero, but it destroys the proportionality to the library size which is key to log-ratio analysis, particularly for low count values, including zeroes. After row centering (i.e. deducting $\mathbf{r}$), taxa with many zeroes (and/or many ones and/or twos) will now primarily contain elements of $\mathbf{r}$, in particular, for zero counts $s_{ij} \approx -r_{ij}$. All taxa with many zero values (taxa with a low prevalence, rare taxa, for short) or with very low counts are therefore positively correlated among one another and all are negatively correlated to $\mathbf{r}$. If many such taxa exists, and there is a substantial variability in $\mathbf{r}$, a considerable part of the variance of $\mathbf{S}$ is related to $\mathbf{r}$. Both the variance in $\mathbf{S}$ that is explained by $\mathbf{r}$ and the correlation between $\mathbf{r}$ and $\mathbf{S}$ for rare taxa increases as both the variability in $\mathbf{r}$ and the number of zeroes increase.

Given enough variability in $\mathbf{r}$ and enough zeroes in the data, a log-ratio PCA identifies this artifact as a prominent effect. In this situation the effect of $\mathbf{r}$ is in competition with other effects, and may either influence any of the principal axes or even completely dominate the first axis. In a constrained analysis, e.g. log-ratio RDA, an explanatory variable that happens to be correlated with $\mathbf{r}$ is likely to be judged significant in permutation testing, even if it is unrelated to the taxa data (Type 1 error inflation). By contrast, there will be little power to detect explanatory variables that are uncorrelated with $\mathbf{r}$, but do

influence the microbiome.

We call the problem informally "library size dependence" and the cause "variability in library size", although the formal cause is variability in $\mathbf{r}$. It is important to note that data with an equal library size or equalized library size (rarefaction), may also show variability in $\mathbf{r}$ (Fig. S1). In most cases the library size and $\mathbf{r}$ are correlated, and if a correlation exists between $\mathbf{S}$ and the library size, there is likely also correlation between $\mathbf{S}$ and $\mathbf{r}$. Note that the problem we describe is not purely related to the amount of zeroes; it can also be ascribed to a lack of variability in taxon abundance, which violates the assumption in log-ratio analysis of proportionality to the library size. If the variance of a particular taxon is low, then, after double centering, its variance is largely explained by $\mathbf{r}$. In practice, a low variance is primarily observed for rare taxa.

3.3 Diagnostics

We propose two diagnostics to assess library size dependency in log-ratio PCA of sparse data. We cannot exclude that other data characteristics can cause the patterns described below, but in the context of a log-ratio PCA applied to amplicon sequencing data it is likely that a fit is influenced by the library size via row-centering if these patterns arise.

The first diagnostic is to calculate the correlation between each column of $\mathbf{S}$ and $\mathbf{r}$ (hereafter the correlations are collectively denoted by $\rho_{\mathbf{Sr}}$) and to plot this correlation against the log of the mean abundance per taxon (i.e. the log of marginal column mean of $\mathbf{Y}$). A negative value of $\rho_{\mathbf{Sr}}$ for a low abundance taxon suggests that this taxon primarily contains elements of $-\mathbf{r}$. If $\mathbf{S}$ contains the effect of $\mathbf{r}$, we expect that the low abundance taxa have a strong negative correlation with $\mathbf{r}$. Library size dependence is diagnosed if the graph of $\rho_{\mathbf{Sr}}$ against the log taxon mean shows an increasing trend starting from a low y-axis value (e.g. -0.5, see examples). This does not necessarily mean the 1$^{\text{st}}$ or 2$^{\text{nd}}$ PCA axis is influenced by $\mathbf{r}$, its effect may also be expressed on a subsequent axis. If this trend is absent then there is no dependence on $\mathbf{r}$ or the library size. Note that the correlation diagnostics can be used on any clr transformed matrix and is not specific for log-ratio PCA.

The second diagnostic we suggest is specific for log-ratio PCA; it is a plot of the (log) contribution of each taxon to a particular principal axis against the log of the mean abundance per taxon (i.e. the log of marginal column mean of $\mathbf{Y}$). The contribution of a taxon to an axis can be quantified with the square of its value in $\mathbf{V}$ (Greenacre, 2013a,b), which is output of the earlier described SVD (equation 2). A PCA axis is suspicious if all low abundance taxa have a relatively high and about equal contribution.

In such a case, these low abundance taxa are likely contributing due to their negative correlation with $\mathbf{r}$ and they are contributing to an axis that primarily contains the effect of $\mathbf{r}$. As taxon abundance and variance increase, the correlation with $\mathbf{r}$ reduces and the contribution drops. The most abundant taxa tend to have few zero values and are thus unaffected by $\mathbf{r}$. In extreme cases the resulting pattern is V-shaped. By contrast, if the mean contribution is either a gradually increasing (on the log scale) with taxon abundance or highly variable around a constant, the PCA axis is unsuspected.

Another possible diagnostic is to fit a log-ratio RDA with $\mathbf{r}$ as the constraining variable and estimate how much variance in $\mathbf{S}$ is related to $\mathbf{r}$. The problem with this diagnostic is that it is unclear what percentage of $\mathbf{r}$ related variance is low or high, i.e. we have nothing to compare with. It it also possible to quantify the amount of variance in $\mathbf{S}$ per taxon that can be explained by $\mathbf{r}$ (with $\rho^2_{\mathbf{Sr}}$); this is addressed with the first diagnostic.

3.4 Examples

One example in this paper is based on simulation and two examples are based on amplicon sequencing data. The aim of the simulation is to illustrate what may go wrong with log-ratio PCA. To make transparent how the row centering problem arises, we opt for a relatively simple simulation setting that allows us to asses the effect of a large number of zeroes with a large variation in the library size and, optionally, a correlation between $\mathbf{x}$ and $\mathbf{r}$. The two data examples demonstrate how the row centering problem manifests itself in amplicon sequencing data.

3.4.1 Simulation

In the simulation we draw a matrix of counts, $\mathbf{Y}$, with I samples and J taxa. By default we set $I = 50$ and $J = 500$. As microbiome data commonly show overdispersion compared to the Poisson distribution (McMurdie & Holmes, 2014), Matrix $\mathbf{Y}$ is sampled from a negative binomial distribution with mean μ_{ij} and variance $\mu_{ij} + \mu^2_{ij}$. We set the expectation μ_{ij} with a log-linear model: $log(\mu_{ij}) = a_i + t_j + b_j x_i$, where a_i reflects the library size and is drawn according to $a_i \sim N(0, \sigma_a)$, t_j reflects the overall abundance of taxon j and is drawn according to $t_j \sim N(0, \sigma_t)$, and x_i represents a binary (0/1) variable representing two treatment groups of equal size with b_j the treatment effect on taxon j. By default we set $\sigma_t = 2$, and we set σ_a to either 0, 0.5 or 1 so as to study the effects of library size. At random, 100 out of 500 taxa are made differentially abundant which are at random with equal probability either up or down regulated by setting b_j equal to b and $-b$, respectively; for the remaining taxa $b_j = 0$. Unless stated otherwise, taxa present in less than 5 samples are removed.

It is of interest to see how log-ratio PCA performs if the library size is correlated with the treatment, for example if the samples from one treatment group tend to have a higher library size than the samples from the other treatment group. We simulate such scenario by incorporating a correlation between $\mathbf{x}$ and $\mathbf{r}$. This is achieved by modelling $\mathbf{x}$ with a logistic function, according to $x_i \sim Bernoulli(g\frac{e^{\gamma a_i}}{1+e^{\gamma a_i}})$. With parameter γ we can set strength of the correlation ($\rho_{\mathbf{xr}}$). Parameter g is set for each simulated draw to ensure that the treatments groups are equal in size.

We first use the simulation model to demonstrate the diagnostics by simulating one data set per level of library size variability, i.e. $\sigma_a = 0$, 0.5, and 1, in the situation without correlation between $\mathbf{x}$ and $\mathbf{r}$, i.e. $\gamma = 0$. This results in example data sets with library sizes of, respectively, 2731 to 5842, 1215 to 13256, and 349 to 34907. After removing taxa that with less than 5 occurrences, these examples contain 445, 441, and 458 taxa and 42%, 42% and 44% zeroes, respectively. The fold change for the differentially abundant taxa in these simulations is set to 3 ($b = \log(3)$).

Next, we repeatedly simulate new data to estimate the type 1 error and power of a log-ratio RDA to detect the effect of the treatment $\mathbf{x}$ at the nominal significance level of 0.05. Here we explore two scenarios. First, we assess how variability in $\mathbf{r}$ affects the type 1 error and power by varying the fold change (in four steps from 1 to 2) for three levels of σ_a. In this scenario there is no correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 0$). In a second scenario, we explore what effect the correlation between $\mathbf{x}$ and $\mathbf{r}$ has on the type 1 error by varying γ between 0 and 3. As this scenario concerns type 1 error, there is no treatment effect (fold change = 1, $b = \log(1)$). With $\gamma = 2$, the average (Pearson's) correlation across 2000 simulations between $\mathbf{x}$ and $\mathbf{r}$ is 0.23, 0.41, 0.58 for, respectively, σ_a 0.25, 0.5, and 1. For a visualization of $\rho_{\mathbf{ax}}$ and $\rho_{\mathbf{rx}}$ for various values of γ we refer to supplemental figure S18. In the power and type 1 error simulations we also explore some putative solutions and asses how robust these are to the studied data charactersics. These solutions consist of alternative versions of log-ratio PCA and closely related methods. Details on these methods are available in the supplementary information.

3.4.2 Biting midges data

In the first real data example we examine a data set of 191 observations on laboratory reared biting midges. Each observation contains the pooled abdomens of 5 adult female biting midges that were fed for a period of time after hatching on sugar water supplemented with or without antibiotics to affect the gut microbiome. In total, 86 pools contained biting midges that received antibiotics and 105 pools received no antibiotics. Per pool fragment DNA was isolated, fragments of 16S were (amplified and) sequenced

(Illumina MiSeq), and grouped into Amplicon Sequence Variants (ASVs). For more information we refer to original publication (Möhlmann et al., 2020).

The original publication analyzed multiple biting midge species; here we only use the *Culicoides nubeculosus* samples. In the original study, only the samples were used with biting midges fed on sugar water for a period of 6 days, as this gave the best indication of the effect of antibiotics. For illustration purpose, we use all sequenced samples for this species that were collected during the course of the experiment (data from 2^{nd} to 11^{th} day feeding on sugar water with and without antibiotics). Analogous to the simulated example, we call the treatment variable $\mathbf{x}$.

We removed ASVs that were absent in 10 or more samples, leaving 155 taxa, containing 85% zeroes. The library size varies from 335 to 128.175 reads. Both the library size and $\mathbf{r}$ are correlated with the treatment variable (Figure 1), but with opposite signs. The (Pearson's) correlation between $\mathbf{x}$ and $\mathbf{r}$ is 0.54 in absolute value.

3.4.3 Rice data

In the second real data example we examine a data set about the root associated microbiome of 296 rice cultivars cultivated under field conditions. Each cultivar was grown with sufficient (control) and insufficient water (drought), giving 592 observations. Each observation contains the material of three pooled replicates. Per observation DNA was isolated, fragments of 18S were (amplified and) sequenced (Illumina MiSeq), and clustered into operational taxonomic unit (OTUs). Analogous to the simulated example, we call the treatment variable $\mathbf{x}$. For further details we refer to Andreo-Jiménez (x); Andreo-Jiménez et al. (2019).

Taxa that were absent in 10 or more samples were removed, leaving 650 taxa which together contained 92% zeroes. The library size varies from 651 to 92.224 reads. Both the library size and $\mathbf{r}$ are correlated with the treatment variable (Figure 1). The (Pearson's) correlation between $\mathbf{x}$ and $\mathbf{r}$ is 0.40 in absolute value.

3.5 Software

Log-ratio PCA was carried out using the function `dudi.pca` from `R` package `ade4` Dray & Dufour (2007) using a double centered log transformed counts matrix as input. For the log-ratio RDA (constrained analysis) we subsequently used the function `pcaiv`, and testing was done with `randtest` (both `ade4`), which performs a Monte Carlo permutation test (999 permutations). The testing was done on the

percentage of explained variance, i.e. constrained inertia in `ade4`.

4 Results

4.1 Diagnostics

4.1.1 Simulated examples

The simulated data examples illustrate how log-ratio PCA is influenced by variability in library size in the presence of zeroes (Figure 2). If the variation is low ($\sigma_a = 0$), the samples of the two treatment groups are clearly separated along the 1$^{\text{st}}$ PCA axis. There is no strong trend in $\rho_{\mathbf{Sr}}$ against the log taxon mean (Figure 2A-C) and the taxon contribution increases on average with taxon abundance. This demonstrates that for this scenario log-ratio PCA performs well, despite the presence of a large number of zeroes (42%).

If the variation in library size is increased ($\sigma_a = 0.5$, Figure 2D-F), the effect of **r** starts to compete with **x**. The first axis still largely contains the effect of **x**, but **r** is affecting the 2-d sample configuration. We see a clear trend in $\rho_{\mathbf{Sr}}$ with taxon abundance and the contributions to the 2$^{\text{nd}}$ axis display relatively high contributions for the low abundance taxa. If we further increase the variation in library size ($\sigma_a = 1$, Figure 2G-I), the effect of x is pushed to the 2$^{\text{nd}}$ axis. The 1$^{\text{st}}$ axis now reflects **r** and, thus the library size. The trend in correlations is more pronounced, with many abundunt taxa having positive correlation (up to 0.5), so that the contribution plot shows a V-shaped pattern.

4.1.2 Data examples

In both real data examples we see a good separation of the treatment groups in a two-dimensional log-ratio PCA, suggesting the treatment has a strong effect (Figure 3A & D). For the biting midges example this effect is on the 1$^{\text{st}}$ axis. For the rice example this effect seems to be tilted. For both data examples we see a clear trend in $\rho_{\mathbf{Sr}}$ against taxon abundance (Figure 3 B & E) and a relatively high and about equal contribution amongst the low abundance taxa (3 C & F). These patterns are similar to what we observed in the simulated example. These results suggests that the 1$^{\text{st}}$ axis, at least partly, contains the effect of **r**.

Given the correlation between **x** and **r** in these data sets (Figure 1), it is likely that in both data examples the 1$^{\text{st}}$ axis contains both the effect of **x** and **r**. In the rice example it is possible that the tilting of the effect is caused by the effect of **r** (similar to the simulated example with $\sigma_a = 0.5$, Figure 2D). For the

diagnostics, it is clear that the log-ratio PCA results are, at least partly, influenced by the library size.

4.2 Power and Type 1 error

Without treatment effect (fold change = 1, $b = 0$) and with a treatment that is independent of the library size, log-ratio RDA yields the correct type 1 error rate (0.05), irrespective of library size variability (σ_a)(Figure 4A). With low to moderate library size variability ($\sigma_a \leq 0.5$), log-ratio RDA has good power to detect the treatment effect. With a larger library size variability ($\sigma_a = 1$) the power strongly decreases, e.g. with a fold change of 1.5 it decreases from about 0.75 at $\sigma_a = 0.5$ to about 0.25 at $\sigma_a = 1$. If the treatment is correlated with the library size ($\gamma > 0$), log-ratio RDA reasonably controls the type 1 error rate, if there is low library size variability ($\sigma_a = 0.25$). If the library size variability is moderate to large ($\sigma_a = 0.5$ or 1), log-ratio RDA shows strong type 1 error rate inflation, with error rates running close to 1 whereas the nominal level is 0.05 (Figure 4 B).

Figure 5 compares type 1 error and power of some putative solutions (see supplementary information for details) with those of log-ratio RDA. In the absence of correlation between treatment and library size ($\rho_{\mathbf{xr}} = 0$, $\gamma = 0$), all methods (including log-ratio RDA) have a good control of the type 1 error, irrespectively of the amount of variation in library size (σ_a)(Figure 5A & B, Figure 6A & B). However, the power of most putative solutions do not decrease as much with increasing library size variation (σ_a) as log-ratio RDA does (Figure 5C & D). The methods log-ratio RDA with geometric Bayesian multiplicative zero imputation (GMB), log proportions RDA, and canonical correspondence analysis (CCA) on square rooted data are high-ranked in terms of power with both low and high library size variation. After an additional filtering step, the drop in power for an increased σ_a is minor or absent for all methods (6 C & D). The improvement here is most notable for log-ratio RDA.

With a correlation between treatment and library size ($\gamma = 2$), the putative solutions control the type 1 error for moderate library size variation (Figure 5E), but show moderate to large type 1 error inflation (> 0.10) for large library size variation ($\sigma_a = 1$ (Figure 5F), with the exceptions of CCA on counts and RCM that both perform badly in having a type 1 error rate that is too low (Figure 5F). Notably, log-ratio RDA with GBM imputation on proportions shows less type 1 error inflation than log-ratio RDA with GBM imputation on counts (Figure 5F). Part of the type 1 error inflation for all methods is caused by a difference in the number of zeroes between treatment groups of $\mathbf{x}$ that can occur as a result of $\rho_{\mathbf{xr}}$. In this scenario, the performance of all methods, but in particular of log-ratio RDA, can be improved by filtering out low abundance taxa (Figure 6 E & F).

5 Discussion

Log-ratio PCA is designed to give results that are library-size independent. However, as we demonstrated mathematically and with examples based on simulated and real data, log-ratio PCA becomes library-size dependent, if there are many infrequent taxa (many zeroes) and library sizes differ largely. In this situation, the row centering used in log-ratio PCA brings an effect of $\mathbf{r}$ (the row mean of the log-transformed counts) in the clr transformed matrix. Note that this effect is irrespectively of whether or not these infrequent taxa are genuine or due to sequencing noise or allocation error. This library-size dependence is unexpected in the sense that, after applying the clr, the transformed matrix is free of the effect of the row totals for strictly positive data ($y_{ij} > 0$ for all i and j). We additionally demonstrate that library size variability causes a loss in power in detecting an effect of $\mathbf{x}$ with log-ratio RDA. If there is additionally a correlation between treatment and the library size, the type 1 error for detecting the effect of $\mathbf{x}$ can be seriously inflated.

How serious is the issue in practice? It is important to note that we focus on fairly extreme scenarios in this paper. Both example data sets have a high proportion of zeroes, large variation in library size, and a correlation between treatment and library size. To some extent this can be seen as a worst case scenario, but at the same time this is a realistic situation that may occur frequently with amplicon sequencing data. These data characteristics may also occur outside the field of amplicon sequencing, although we are unaware of such data. Note that RNASeq data are closely related, but have less zeroes and less variability in the library size. Our simulated data are also extreme, aimed at describing the issues that may arise.

Our main message is that one has to be careful when analyzing data with the described characteristics with log-ratio PCA. We provide two diagnostics. If these diagnostics display the patterns described in this paper, additional actions are required. The most straightforward solution is stringent filtering out low abundance or infrequent taxa. Note that, if a particular data set is less extreme in the described data characteristics than the data in this paper, log-ratio PCA will likely work and, in these cases, it is a powerful tool in analyzing compositional data. We additionally explored various putative solutions (see also the supplementary information), some of which can also increase performance under the described circumstances.

There is a feature in the diagnostics that we do not fully understand mathematically, namely that many abundant taxa in situations with extreme library size variability show positive correlation ($\rho_{\mathbf{Sr}}$ up to 0.5) in the correlation diagnostic, resulting in extreme cases in a V-shaped pattern in the contribution plot.

These positive correlation occur in both the simulation and data examples (Figures 2 and 3) showing that the feature is real and not an artefact of our simulation. One possible explanation is that an effect of $-\mathbf{r}$ in low abundance taxa has to be compensated elsewhere, due the zero-sum constraint of the centered log-ratio, resulting in positive correlations amongst high abundance taxa.

Although the focus of this paper is on multivariate methods, there also consequences for other methods based on the clr. With high variation in library size and correlation between treatment and library size, low abundance clr transformed taxa will likely test significant in univariate analysis, even if there is no treatment effect, leading to type 1 error inflation. In case of graphical modelling with clr transformed taxa, we may detect spurious edges between low abundance taxa. The correlation diagnostic described in this paper can also be used prior to such analyses.

To some extent the large variability in library size and/or $\mathbf{r}$ and the large amount of zeroes are related to data quality. Currently the variation in library size is ill-understood, often not random, and it may even be correlated to a treatment variable, as in our examples. Future developments may lead to a better understanding of this variation and possibly, to more equal library sizes, which will reduce the problems we described.

6 Acknowledgment

We thank Beatriz Andreo Jimenez and Harro Bouwmeester for making the data on the rice example available.

7 Author Contributions

Initialize research: DETB, CJFTB, TWRM

Performed analyses: DETB

Wrote paper: DETB, CJFTB, EHN

Reviewed manuscript: All authors

8 Code and data Accessibility

Code and data are available on `github.com/DennisBeest` and in the supplement. The data of the midges example is also available from Möhlmann et al. (2020).

References

Aitchison, J (1983). Principal component analysis of compositional data, *Biometrika*, 70. https://doi.org/10.2307/2335943.

Andreo-Jiménez, B (x). Genetic mapping of the root fungal microbiome in rice: towards a better yield under drought, *manuscript in progress*, x, x.

Andreo-Jiménez, B, Vandenkoornhuyse, P, Lê Van, A, Heutinck, A, Duhamel, M, Kadam, N, Jagadish, K & Ruyter-Spira, H, C. Bouwmeester (2019). Plant host and drought shape the root associated fungal microbiota in rice, *PeerJ*, 7. https://doi.org/10.7717/peerj.7463.

Dray, S & Dufour, A (2007). The ade4 package: Implementing the duality diagram for ecologists, *Journal of Statistical Software*, 22(4), 1-20. https://doi.org/10.18637/jss.v022.i04.

Gloor, G, Macklaim, J, Pawlowsky-Glahn, V & Egozcue, J (2017). Microbiome datasets are compositional: And this is not optional, *Frontiers in Microbiology*, 8. https://doi.org/10.3389/fmicb.2017.02224.

Greenacre, M (2012). Biplots: The joy of singular value decomposition, *Wiley Interdisciplinary Reviews: Computational Statistics*, 4. https://doi.org/10.1002/wics.1200.

Greenacre, M (2013a), Contribution biplots, *Journal of Computational and Graphical Statistics*, 22. https://doi.org/10.1080/10618600.2012.702494.

Greenacre, M (2013b). The contributions of rare objects in correspondence analysis, *Ecology*, 94. https://doi.org/10.1890/11-1730.1.

Greenacre, M (2018). *Compositional Data Analysis in Practice*, Chapman and Hall/CRC.

McMurdie, P & Holmes, S (2014). Waste not, want not: Why rarefying microbiome data is inadmissible, *PLOS Computational Biology*, 10(4), 1-12. https://doi.org/10.1371/journal.pcbi.1003531.

Möhlmann, T, Vogels, C, Göertz, G, Pijlman, G, ter Braak, C, te Beest, D, Hendriks, M, Nijhuis, E, Warris, S, Drolet, B, van Overbeek, L & Koenraadt, C (2020). Impact of gut bacteria on the infection and transmission of pathogenic arboviruses by biting midges and mosquitoes, *Microbial Ecology*, 80. https://doi.org/10.1007/s00248-020-01517-6.

ter Braak, C (1994). Canonical community ordination. part 1: Basic theory and linear methods, *Écoscience*, 1. https://doi.org/https://10.1080/11956860.1994.11682237.

Tsilimigras, M & Fodor, A (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges, *Annals of Epidemiology*, 26. https://doi.org/10.1016/j.annepidem.2016.03.002.

van den Wollenberg, A (1977). Redundancy analysis: An alternative for canonical correlation analysis, *Psychometrika*, 42. https://doi.org/10.1007/BF02294050.

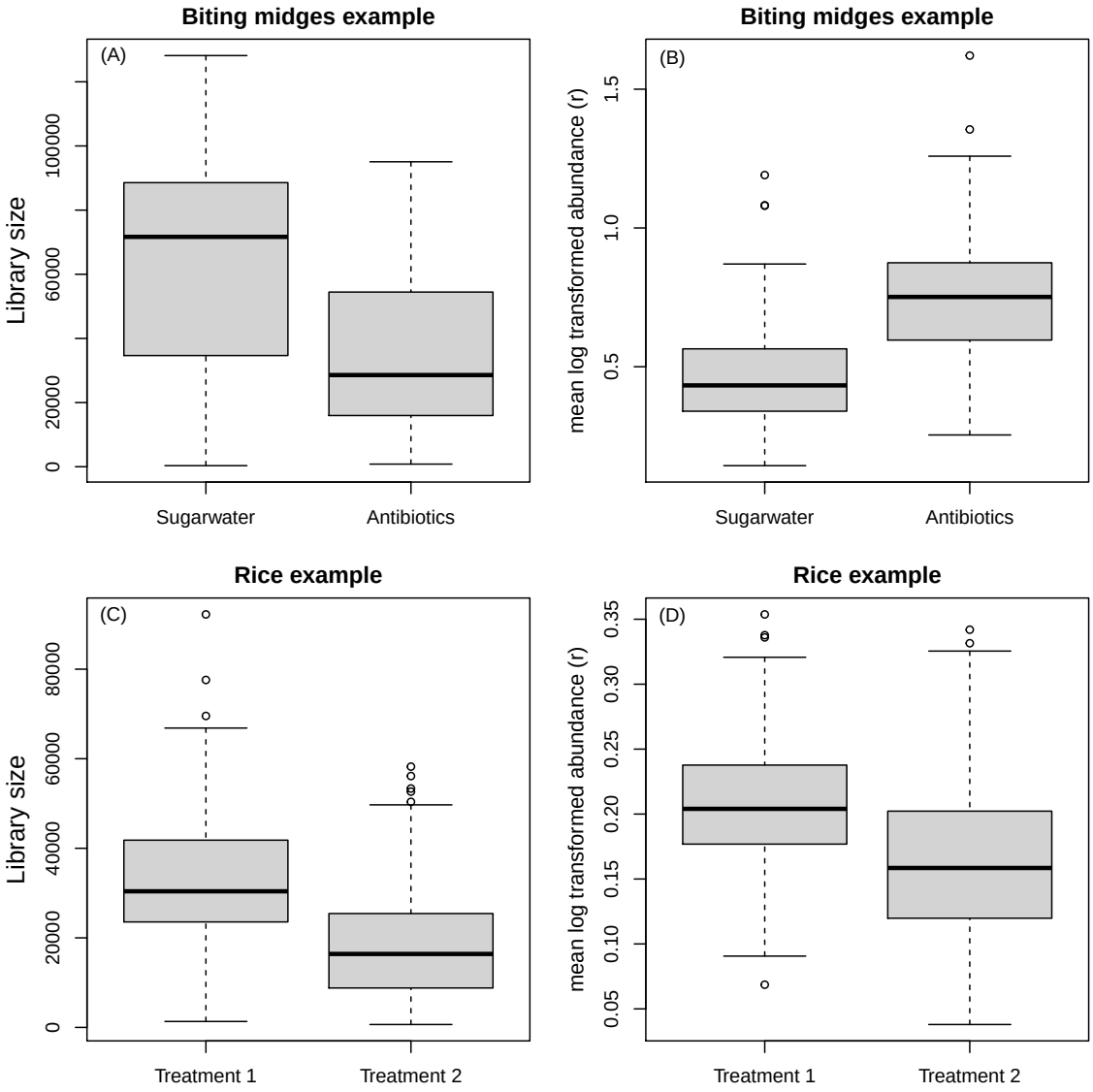

Figure 1: The library size (A & C) and the mean **r** (B & D) per treatment for both example data sets. In both examples, the library size and **r** are correlated with the treatment (**x**).

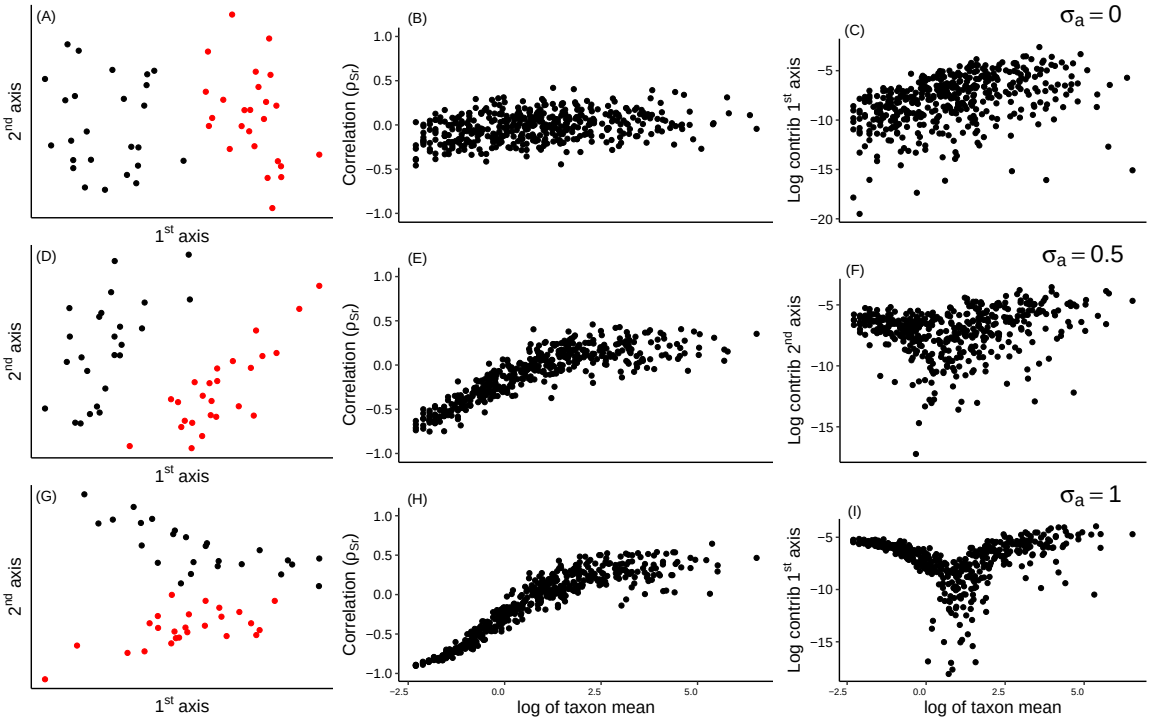

Figure 2: Simulated data. Log-ratio PCA and diagnostics (columns) for three levels of library size variability (rows: σ_a=0, σ_a=0.5, σ_a=1). The first column (A, D, G) displays the simulated observations on the 1$^{\text{st}}$ and 2$^{\text{nd}}$ principal axes, colors indicate treatment groups. The second column (B, E, H) displays the correlation between $\mathbf{S}$ (clr transformed abundances) and $\mathbf{r}$, and the third column (C, F, I) displays the contribution of a taxon versus its log mean abundance. For $\sigma_a = 1$, the 1$^{\text{st}}$ axis contains the effect of $\mathbf{r}$ and the effect of $\mathbf{x}$ is pushed to the 2$^{\text{nd}}$ axis.

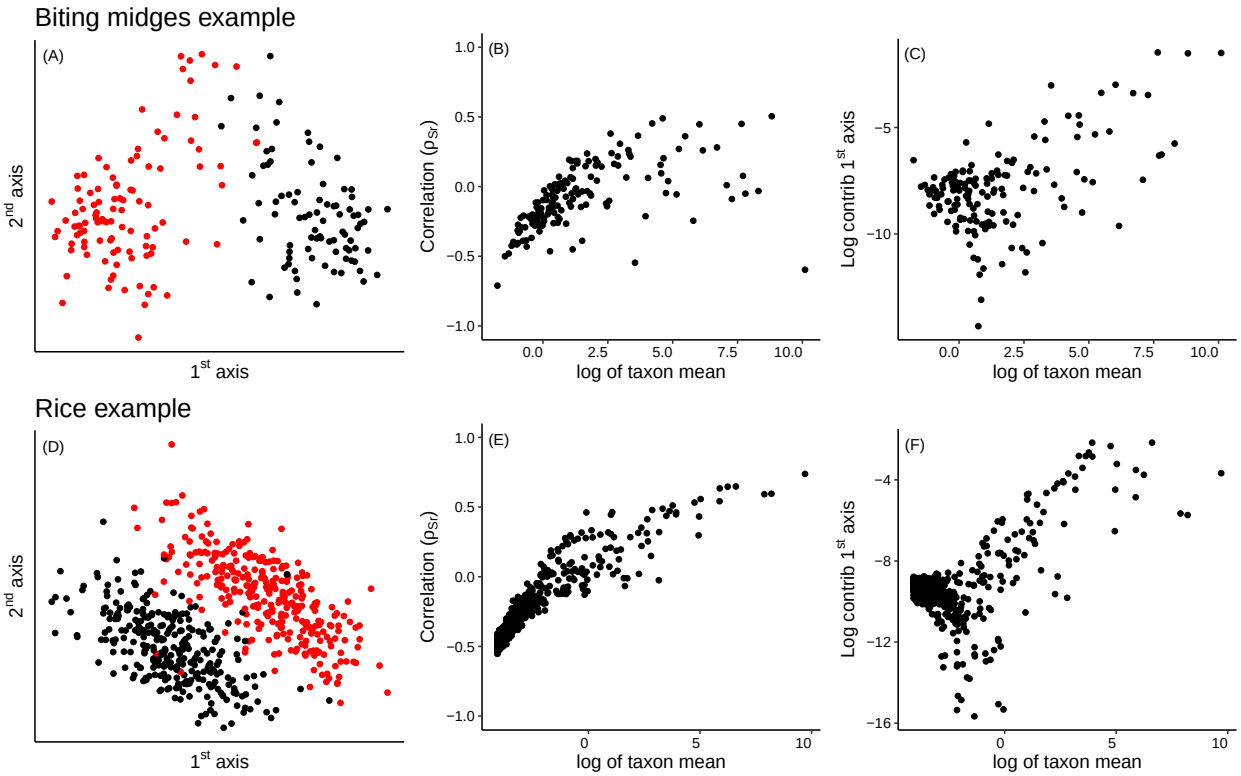

Figure 3: Log-ratio PCA and diagnostics (columns) for the real data examples. The first column (A & D) displays the observations on the 1st and 2nd principal axes, colors indicate treatment groups. The second column (B & E) displays the correlation between $\mathbf{S}$ (transformed abundances) and $\mathbf{r}$. The third column (C & F) displays the log contribution to the 1st axis per taxon versus its log mean abundance. The negative correlations and the relatively high and similar contributions amongst the low abundance taxa suggest there is an issue with row centering (and thus with log-ratio PCA) for both data examples.

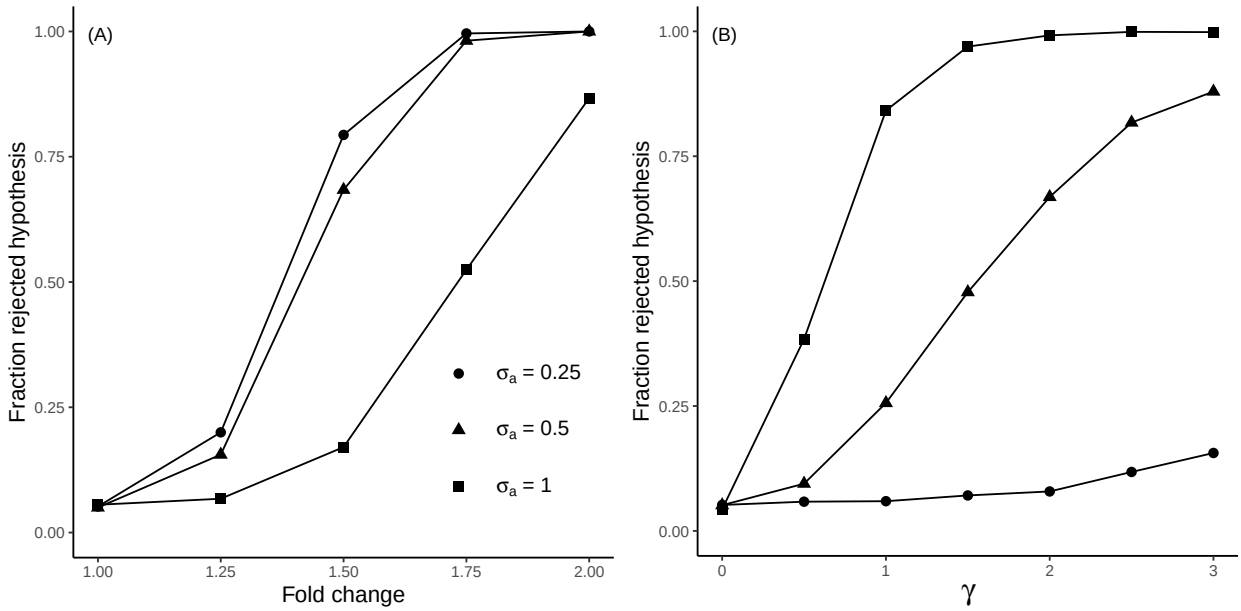

Figure 4: Rejection rate (number of p-values < 0.05 across 2000 simulations) in testing the treatment effect using log-ratio RDA. In (A) the fold change is increased for several levels of σ_a under independence of the treatment with the library size ($\gamma = 0$). The type 1 error is controlled, but the power is reduced as σ_a increases. In (B) there is no treatment effect (fold change is 1, $b = 0$), but there is an increasing correlation between treatment and library size (set with $\gamma \geq 0$) for three levels of σ_a. The type 1 error is controlled for $\gamma = 0$, but increases for higher values of γ and σ_a.

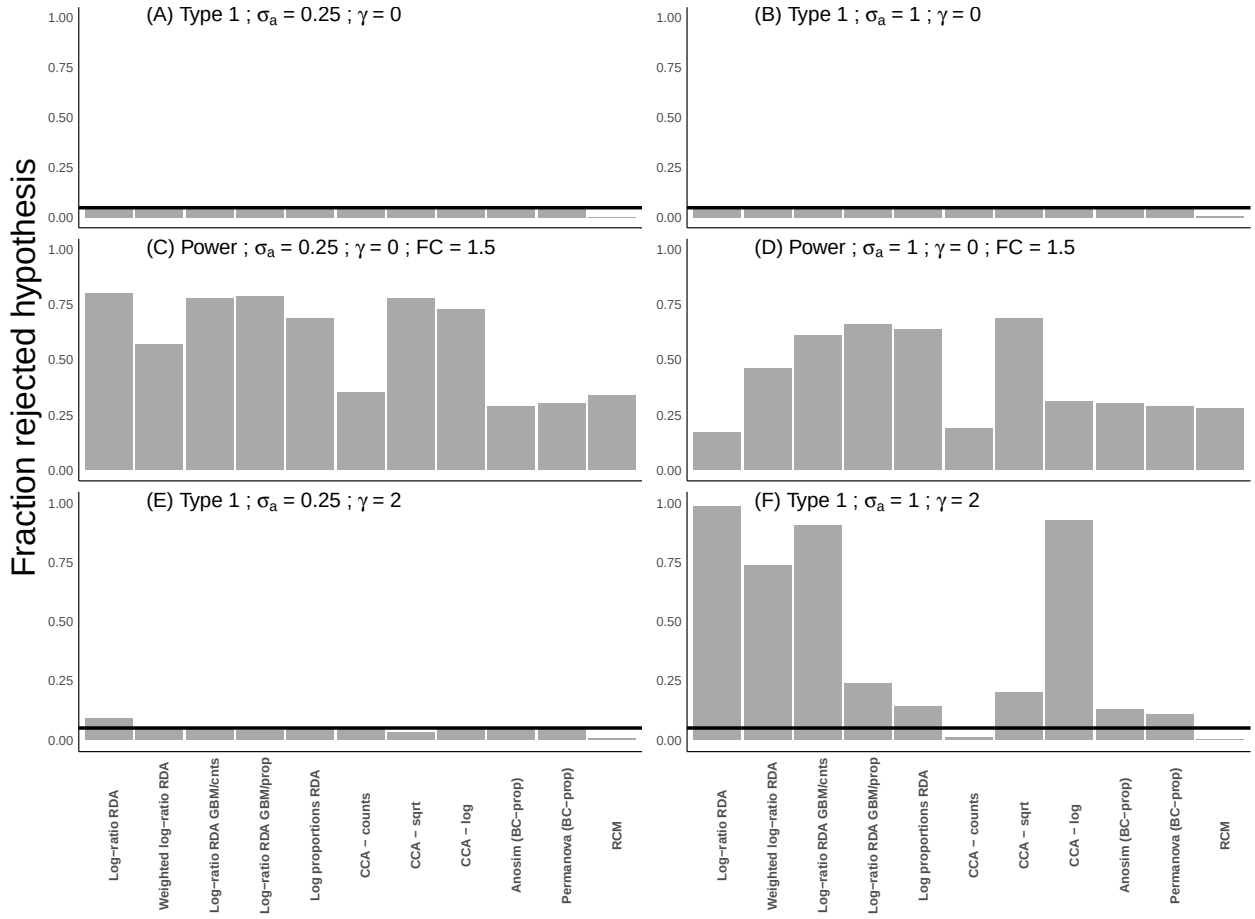

Figure 5: Type 1 and power for a set of methods closely related to log-ratio RDA for two levels of σ_a. (A) and (B) display the type 1 error without correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 0$). (C) and (D) display the power (Fold change = 1.5) without correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 0$). (E) and (F) display the type 1 error when there is a correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 2$). For all methods (expect RCM) the type 1 error and power were determined by counting the number of p-values below 0.05 across 2000 simulations. For RCM we did between 200-250 simulations, expect for (F) where most estimations failed.

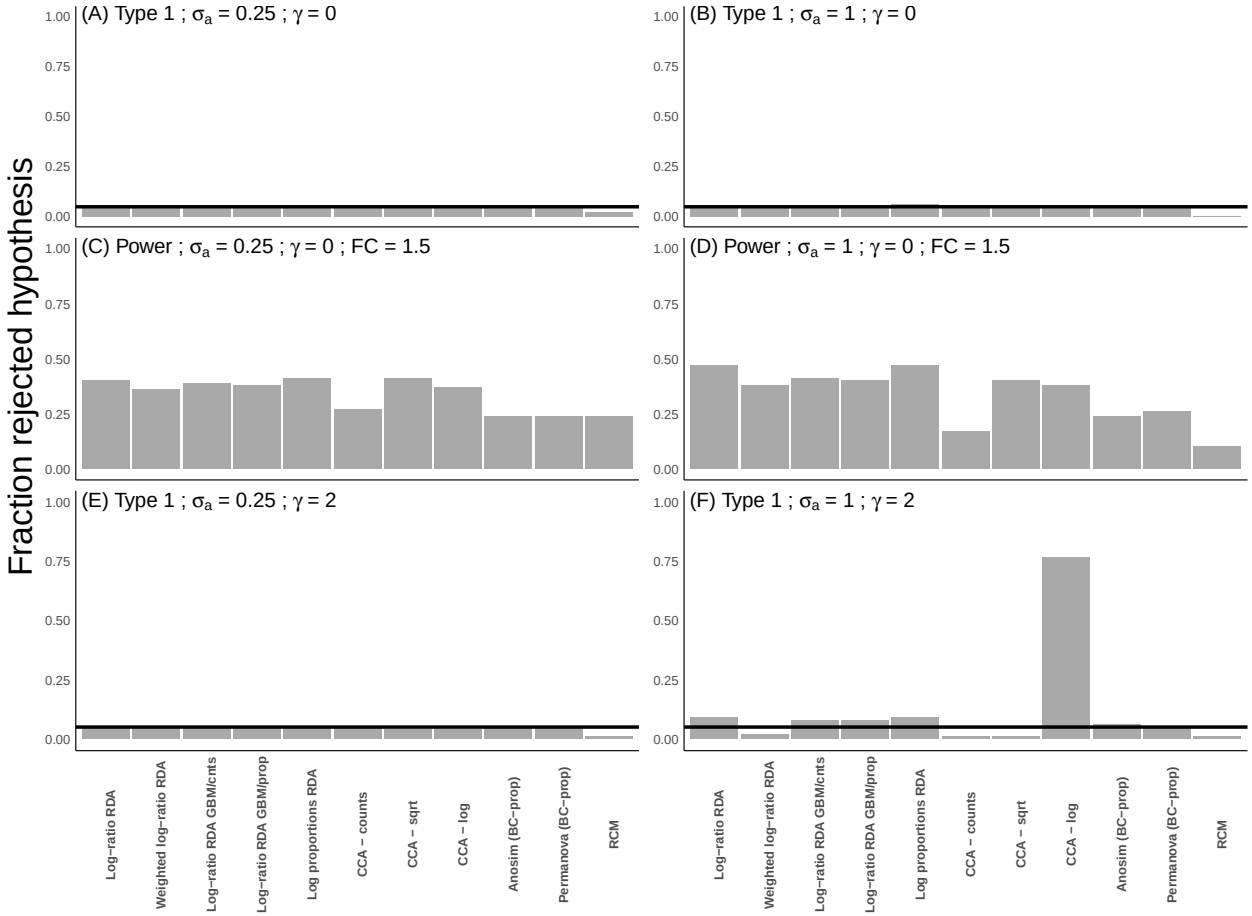

Figure 6: Type 1 and power for a set of methods closely related to log-ratio RDA for two levels of σ_a. Compared to 4, the simulated data were subject to an additional filtering step (see supplement for more information). (A) and (B) display the type 1 error without correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 0$). (C) and (D) display the power (Fold change = 1.5) without correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 0$). (E) and (F) display the type 1 error when there is a correlation between $\mathbf{x}$ and $\mathbf{r}$ ($\gamma = 2$). For all methods (expect RCM) the type 1 error and power were determined by counting the number of p-values below 0.05 across 2000 simulations. For RCM we did between 500 simulations.