JDSSV

Journal of Data Science, Statistics, and Visualisation

May 2025, Volume V, Issue VI. *doi: 10.52933/jdssv.v5i6.133*

Comparing Model-Based Unconstrained Ordination Methods in the Analysis of High-Dimensional Compositional Count Data

Wenqi Tang
University of Jyväskylä

Pekka Korhonen
University of Jyväskylä

Jenni Niku
University of Jyväskylä

Klaus Nordhausen
University of Helsinki

Sara Taskinen
University of Jyväskylä

Abstract

Model-based ordination of ecological community data has gained significant popularity among practitioners recently, largely due to increased availability and utilization of computational resources. Specifically, generalized linear latent variable models (GLLVMs)—a factor-analytic and rank-reduced form of mixed effect models—have proven to be both accurate and computationally efficient. GLLVMs have been implemented for a wide range of response types common to ecological community data; presence-absence, biomass, overdispersed and/or zero-inflated counts serving as examples. In this paper, we demonstrate how GLLVMs can be applied in the analysis of high-dimensional compositional count data. These methods are useful, for example, in the analysis of microbiome data, which are typically collected using modern lab-based sampling tools and are inherently compositional due to the finite capacity of sequencing instruments. We use simulation studies to compare the ordination methods based on GLLVMs with algorithmic compositional data analysis methods that rely on log-transformations. Also recently developed fast model-based ordination methods that utilize Gaussian copula models are included in our comparisons. The methods are illustrated with a microbiome data example.

Keywords: Community-level modeling, copula, latent variable model, overdispersion, zero-inflation.

1. Introduction

Advancements in modern sampling and classification techniques, such as high-throughput sequencing (HTS, Fernandes et al. 2014; Conesa et al. 2016; Pollock et al. 2018), which is an experimental technique capable of rapid and large-scale generation of DNA or RNA sequence data (Gloor et al. 2017), have significantly advanced microbiome research. The raw HTS data, where each sample generates thousands or even millions of gene sequences (reads), typically represent multiple species within a microbial community. Through preprocessing steps, such as removing duplicates and performing clustering analysis on these sequences, they can be grouped into distinct operational taxonomic units (OTUs, Gloor et al. 2017) or classified using alternative methods to achieve different partitions of the raw HTS data such as amplicon sequence variants (ASVs). These approaches enable the partitioning of HTS data into biologically meaningful units. In practical analyses, microbiome data are often represented as OTUs or ASVs, and consist of relative counts; exhibiting unique characteristics that pose significant challenges for statistical analysis. First, the data are inherently compositional due to the finite capacity of sequencing instruments (Gloor et al. 2017). Additionally, the data are often high-dimensional and sparse; the number of variables can exceed thousands, and many observations may contain zeros due to biological or technical factors such as under-sampling.

The analysis of microbiome data has become a vibrant area of research. Methods inspired by compositional data analysis (Aitchison 1986) are widely used, as evidenced by recent works (Gloor et al. 2016; Greenacre et al. 2021). More recently, probabilistic models tailored to compositional data have gained popularity (Lutz et al. 2022; Zeng et al. 2023). For comprehensive reviews, see Swift et al. (2023) and Peterson et al. (2024).

As an example of high-dimensional, sparse, and overdispersed data, consider the microbial community data set described in Kumar et al. (2017). The data consist of normalized read counts of 985 bacterial species sampled from 56 soil sites across three distinct climatic regions: Ny-Ålesund (high Arctic), Kilpisjärvi (low Arctic), and Mayrhofen (European Alps) (Kumar et al. 2017). These regions have distinct climate and environmental conditions making them ideal to study the impact of habitat on microbial communities. For recording species' abundances from each sample, bacteria were identified based on the similarity of their 16S rRNA genetic sequences and classified into operational taxonomic units (OTUs). The data are illustrated in Figure 1, highlighting the sparsity (63.3% zeros) and the overdispersion (variance exceeding the mean for most bacteria). This data set is available via `data("microbialdata")` in the R package **gllvm** (Niku et al. 2024).

In this paper, we focus on visualizing high-dimensional microbiome data using ordination methods. Ordination refers to dimensionality reduction techniques that reduce

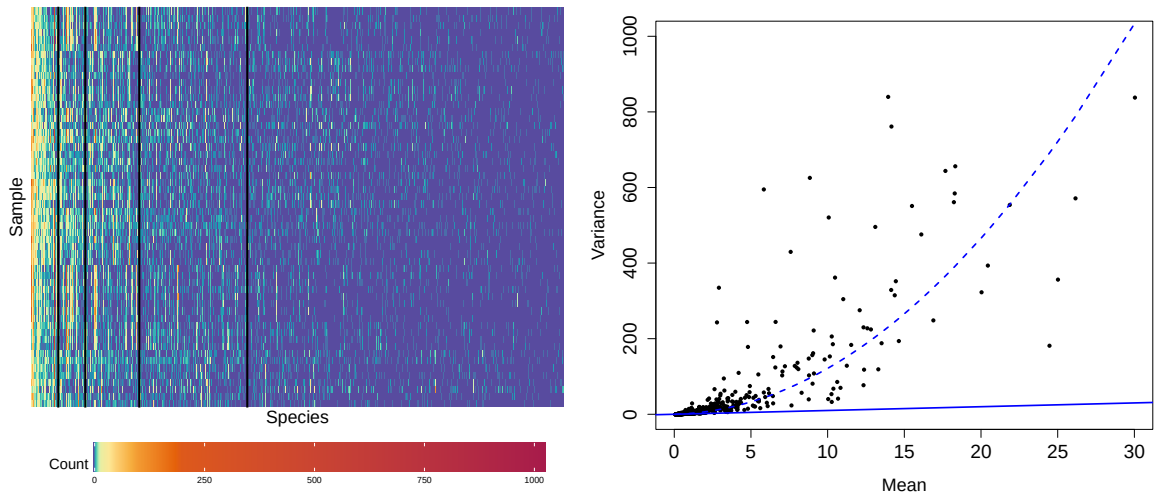

Figure 1: Left: Heatmap showing counts of $m = 985$ bacteria across $n = 56$ samples. Blue indicates absence of bacteria, while dark red indicates high abundances. The black vertical lines represent subsets of sizes $m = 50$, 100, 200, and 400, to be used as a basis for simulation studies. Right: Species-wise means plotted against species-wise variances. The solid line represents the Poisson mean-variance relationship, and the dotted line represents the negative binomial relationship with a dispersion parameter 1.1.

data from many variables to (typically) two dimensions for visualization. These dimensions, known as ordination scores, facilitate the detection of patterns in community composition (Legendre and Legendre 2012). When covariates are not included, the process is referred to as unconstrained ordination (Clarke 1993), also known as indirect gradient analysis. Conversely, when covariates are incorporated, constrained ordination (Ter Braak and Prentice 1988), also referred to as direct gradient analysis (Ter Braak and Prentice 1988), or concurrent ordination methods (van der Veen et al. 2023), are employed.

Ordination has usually relied on dissimilarity-based methods such as non-metric multidimensional scaling (nMDS, Kruskal 1964a,b). These approaches aim to simplify a complex dissimilarity matrix—which captures all pairwise distances between observation units—by reducing its dimensionality, while preserving as much of the original distance information as possible. As dissimilarity-based methods rely heavily on prechosen dissimilarity measure, they do not necessarily account for important properties of microbiome data, such as mean-variance relationships or compositional constraints.

We instead focus on model-based approaches, which explicitly account for key data characteristics such as sparsity, overdispersion, and compositional structure. The methods also provide tools for inference, model selection, and diagnostics. Two model-based approaches are compared in this paper:

- The first approach builds upon a joint modeling framework using generalized linear latent variable models (GLLVMs, Skrondal and Rabe-Hesketh 2004), which can be seen as a factor-analytic extension of generalized linear mixed models. GLLVMs account for correlations among high-dimensional responses by introducing a small number of latent variables. The latent variables capture the un-

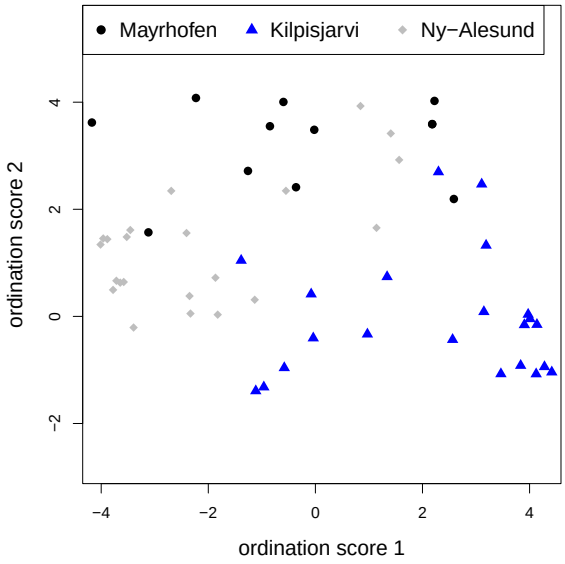

Figure 2: An example of ordination plot based on microbial community data.

derlying structure of the data, enabling dimensionality reduction while explicitly modeling response dependencies.

- The second approach combines marginal generalized linear models (GLMs) with a multivariate model to address correlations across responses. Specifically, the Gaussian copula latent variable model (GCLVM, Popovic et al. 2022) maps responses to copula values with a Gaussian distribution. Classical factor analysis is then applied to copulas to model the dependency structure. This two-step approach allows the GCLVM to handle both marginal distributions and multivariate dependencies effectively.

Both approaches tackle the key challenges inherent in microbiome data, such as overdispersion and sparsity, by selecting appropriate distributions for the responses. Moreover, the compositional nature of the data is addressed through the specification of the linear predictor in GLLVMs or GLMs. By comparing these two methods, we aim to provide insights into their strengths and suitability for visualizing and analyzing complex high-dimensional data.

The paper is organized as follows. In Section 2, we recall the algorithmic-based and model-based methods included in the comparisons. Section 3 presents simulation studies and goodness-of-fit comparisons of the methods, and Section 4 illustrates the model-based approach in the unconstrained and concurrent ordination. The paper is concluded with some discussion in Section 5.

2. Ordination Methods

As described above, ordination, when applied to ecological community data, is a dimensionality reduction technique that reduces multivariate data to (usually) two dimensions to reveal patterns in community composition. As an example of an

ordination plot based on microbial community data, as described above, see Figure 2. The plot displays 56 sampling sites according to their ordination scores on two latent axes. Sampling sites that are close to each other in the ordination can be interpreted as having more similar microbial count compositions or relative abundances (Hui et al. 2015; Warton et al. 2015) than those that are farther apart. We discuss this example again in Section 4.

2.1. Algorithmic methods

We start by a short introduction to the classical methods that have been traditionally used for unconstrained ordination due to their easy-to-use interfaces and computationally efficient algorithms, but before that, let us first introduce some notation needed later. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\top$ denote a m-vector of responses, where y_{ij} represents a count of variable $j = 1, \ldots, m$ (here representing bacteria) recorded at an observational unit $i = 1, \ldots, n$ (here representing sampling site). In addition, we can have information in the form of k variables for each observational unit, denoted here as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$.

By far, the most traditional way of performing unconstrained ordination is using so-called dissimilarity-based methods, such as non-metric multidimensional scaling (nMDS, Kruskal 1964a,b). The first step for dissimilarity-based methods is to calculate a dissimilarity matrix between observational units. After that, the dimension reduction is applied to dissimilarity matrix using an algorithm that attempts to preserve information about relative distances. By using some dissimilarity measure, nMDS proceeds with repositioning ordination scores until the relative distances in the ordination have the strongest possible monotonic fit to the pairwise dissimilarities between observational units. When applying nMDS to compositional count data, a classical approach is to apply as a dissimilarity measure the Bray-Curtis distance (Bray and Curtis 1957)

$$d_{ii'} = \frac{\sum_{j=1}^{m} |y_{ij} - y_{i'j}|}{\sum_{j=1}^{m} (y_{ij} + y_{i'j})},$$

that implicitly applies row-standardization (Ricotta and Podani 2017). Another method for handling compositional count data with nMDS is to use the Aitchison distance (Aitchison 1982) to measure dissimilarity. In a practical simulation, to compute the Aitchison distance between compositional data vectors, one can apply a centered log-ratio (clr) transformation as defined in (1), to map the compositional count data from the Aitchison space to real space, and then compute the Euclidean distance as dissimilarity between the two vectors.

Due to the compositional nature of microbiome data, Gloor et al. (2016, 2017) advise applying centered log-ratio (clr) transformation (Aitchison 1982) to the data followed by principal component analysis (PCA) to obtain a low-dimensional representation of multivariate data. The clr-transformation of a m-vector of responses $\boldsymbol{y}_i$, $i = 1, \ldots, n$, can be obtained as

$$\boldsymbol{y}_{i,clr} = \mathrm{clr}(\boldsymbol{y}_i) = \left(\log \frac{y_{i1}}{\sqrt[m]{\prod_{j=1}^{m} y_{ij}}}, \ldots, \log \frac{y_{im}}{\sqrt[m]{\prod_{j=1}^{m} y_{ij}}} \right). \tag{1}$$

The clr-transformed values are logarithms of the ratios of the original components to the geometric mean of the composition. This interpretation facilitates understanding the relative abundances of components within the composition and is based on the assumption that the information lies in the relative proportions rather than the absolute values. In addition to its scale invariance, the clr-transformation preserves the Euclidean distance between compositions in the transformed space. This preservation allows for the application of standard multivariate techniques, such as PCA, to the clr-transformed data. For further details on the clr-transformation, we refer to Chapter 3 of Filzmoser et al. (2018).

A notable drawback of the clr-transformation is that it is not defined if one or more of the observed values equal zero. If the frequency of zeros is relatively low and zeros result from counts below some detection limit or other sampling issues, which can occur, for example, in high-throughput sequencing, a common approach is to replace them with a small value (e.g., one). Alternatively, one can treat zeros as missing values and address them using imputation approaches (Templ et al. 2016; Filzmoser et al. 2018; Lubbe et al. 2021). A more challenging case is when a zero indicates that the count is truly zero. Such zeros are known as structural or essential zeros and should not be replaced with other values. For a more detailed discussion about the differences between sampling zeros and structural zeros, we refer to Chapter 13 of Filzmoser et al. (2018) and references therein. Note that distinguishing between structural and sampling zeros is a difficult problem, and expert knowledge should be used when analyzing sparse data with compositional data analysis methods. When applying clr-transformation, we add, for simplicity, to each count value one to avoid having zero counts. Then PCA is applied to $\boldsymbol{y}_{i,clr}$, $i = 1, \ldots, n$, and the resulting first two principal components in the clr-space serve as ordination scores, meaning that in case the interpretation is required, that needs to take place also in the clr-space.

The popularity of classical algorithmic-based ordination methods arises from their computational simplicity. A notable drawback is however the absence of a probabilistic model for the observed data. By directly modeling data, we can better account for key statistical properties (e.g., mean-variance relationship and sparsity) of data at hand (Warton et al. 2012). The use of probabilistic formulation also allows us to use standard statistical techniques for model selection, inference, and diagnostics. In the next section, we review two model-based approaches for unconstrained ordination.

2.2. Ordination based on latent variable models

Generalized linear latent variable models (GLLVMs, Skrondal and Rabe-Hesketh 2004) offer a flexible framework for specifying a joint model for microbial count data. GLLVMs are extensions of generalized linear models to multivariate response case where correlation within responses is taken into account using a factor-analytic approach. This allows us to use GLLVMs for model-based ordination for any response type as shown in Hui et al. (2015) and reviewed next. For more examples of GLLVMs in the analysis of any multivariate abundance data (e.g., presence-absence data, counts, biomass, cover data), see Warton et al. (2015), and references therein.

Assume now that $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})^\top$ are d-dimensional latent variables that are assumed to follow a standard multivariate normal distribution, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$. In

GLLVMs, we assume that, conditional on latent variables $\boldsymbol{u}_i$, the responses y_{ij} are distributed independently according to some distribution characterized by its mean and possibly some nuisance parameters. To be more specific, we assume that $y_{ij}|\boldsymbol{u}_i \sim F(\mu_{ij}, \boldsymbol{\phi})$, where $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{u}_i)$ is the conditional mean and vector $\boldsymbol{\phi}$ includes possible response-specific parameters for modeling e.g., dispersion and zero-inflation in count data models. When GLLVMs are used to perform model-based unconstrained ordination, we link μ_{ij} to the linear predictor via

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{\lambda}_j^\top \boldsymbol{u}_i, \tag{2}$$

where $g(\cdot)$ is a known link-function (usually log-link for count data), β_{0j} is a column-specific intercept to account for differences in column totals, and $\boldsymbol{\lambda}_j = (\lambda_{j1}, \ldots, \lambda_{jd})^\top$ are d-dimensional factor loadings. To ensure identifiability of the model, the upper triangular of the loading matrix $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \cdots \boldsymbol{\lambda}_m]^\top$ are set to zero to avoid rotational invariance, and the diagonal elements of it are set positive to avoid sign switching (Huber et al. 2004; Niku et al. 2017). The row-specific α_i adjusts for row total abundance and allows us to model relative abundance or composition instead of absolute abundance. Here we assume that α_i is a fixed effect (with an identifiability constraint $\alpha_{i1} = 0$), but in practice it can also be included in model as a random effect (Niku et al. 2019).

As in factor analysis, the latent variables and related factor loadings capture the correlation across study units and the number d of latent variables controls the model complexity. The model (2) can also be seen as a rank-reduced version of generalized linear mixed model (GLMM) with general residual covariance structure $\boldsymbol{\Sigma}$. Now $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, where $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \cdots \boldsymbol{\lambda}_m]^\top$ is a $m \times d$ matrix. If GLLVMs are used for model-based ordination analysis (Hui et al. 2015), the predicted latent variables, $\hat{\boldsymbol{u}}_i$, (usually with $d = 2$) are plotted to illustrate how different study units differ in terms of the microbiota composition. For comparisons of model-based and classical unconstrained ordination methods, see Hui et al. (2015) and Niku et al. (2017), for example. For constrained and concurrent model-based ordination methods utilizing GLLVMs, one can introduce covariates related study units in model (2) as shown in van der Veen et al. (2023).

GLLVMs can be estimated efficiently using maximum likelihood estimation. Collect now into vectors $\boldsymbol{\Psi}$ and $\boldsymbol{u}$ all the parameters and latent variables in the model, respectively. The marginal log-likelihood function to be maximized is obtained by integrating over the missing latent variables, that is,

$$\ell(\boldsymbol{\Psi}) = \log \mathcal{L}(\boldsymbol{\Psi}) = \log \int_{\mathbb{R}^{nd}} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) \right) f(\boldsymbol{u}_i) \right] d\boldsymbol{u}, \tag{3}$$

where $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$ is the conditional distribution of y_{ij} and $f(\boldsymbol{u}_i)$ is the distribution of a latent variable. The above log-likelihood has however a closed-form expression only in the normal-response identity-link (i.e., factor analytic) model. For other response types, a variety of approximation approaches have been proposed. We use the variational approximation (VA) method that allows us to derive a closed-form lower bound for (3) at models suitable for sparse, overdispersed count data. The computational tools are developed in Hui et al. (2017) and Niku et al. (2019). For a recent review of several other computational approaches (including VA), see Korhonen et al. (2024).

2.3. Ordination based on copulas

In high-dimensional data settings, where the number of variables can be counted in thousands, GLLVMs can be computationally intensive even if methods that approximate marginal likelihood in closed form are used. To overcome this, Popovic et al. (2022) proposed a model-based ordination method that uses copulas. The copula model couples a marginal model for the data and a multivariate model that accounts for covariance across responses. When applied to unconstrained ordination, we couple marginal GLMs suitable for sparse, overdispersed count data with a factor analysis as described below. For other recent examples of copula modeling in the analysis of multivariate abundance data, see Popovic et al. (2019) and Anderson et al. (2019).

To specify the marginal model for the counts, we assume that $y_{ij} \sim F_j(\mu_{ij}, \boldsymbol{\phi})$, where $\mu_{ij} = \mathbb{E}(y_{ij})$ and the vector $\boldsymbol{\phi}$ includes all the nuisance parameters as before. We use GLMs to link μ_{ij} to the linear predictor via

$$g(\mu_{ij}) = \alpha_i + \beta_{0j}, \tag{4}$$

where $g(\cdot)$ is again a known link function, and α_i and β_{0j} are row-specific and column-specific intercepts to account for differences in row and column totals respectively. Notice that in Popovic et al. (2022), only a column-specific intercept was included in model (4). We, however, also include a row-specific intercept to account for compositional nature of data. In the Gaussian copula model, counts y_{ij} are mapped to copula values z_{ij} that have a multivariate normal distribution as follows

$$F_j(y_{ij} - 1) \leq \Phi(z_{ij}) < F_j(y_{ij}).$$

Here $F_j(\cdot)$ denotes the cumulative distribution function (cdf) assumed for jth column in data matrix under the marginal GLM, $\Phi(\cdot)$ is the cdf of the standard normal distribution, and $F_j(y_{ij} - 1)$ is the left limit of F_j at y_{ij}. When applying copula model for unconstrained ordination, we assume a factor analytic formulation for copulas, that is,

$$z_{ij} = \boldsymbol{\lambda}_j^\top \boldsymbol{u}_i + \epsilon_{ij}, \tag{5}$$

where, as in GLLVMs, $\boldsymbol{u}_i$ is a d-dimensional latent variable related to study unit and $\boldsymbol{\lambda}_j$ is a d-vector of factor loadings. As in factor analysis, ϵ_{ij} are independent Gaussian errors, with variances σ_j^2 and $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$. Again, predicted latent variables, $\hat{\boldsymbol{u}}_i$, (usually with $d = 2$) serve as ordination points in unconstrained ordination. If one wants to apply copula modeling for constrained and concurrent ordination, covariates can be included in (5) in a similar fashion as in van der Veen et al. (2023).

Popovic et al. (2022) applied a two-step procedure proposed by Joe (2005) for estimating the parameters of the copula model. In the first step, marginal distributions $F_j(\cdot)$ are estimated using GLMs suitable for sparse, overdispersed count data. After that a Monte Carlo Expectation Maximization (MCEM, Wei and Tanner 1990) is applied to estimate the covariance parameters in (5), i.e., $\sigma_1^2, \ldots, \sigma_m^2$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m$. In the E-step, sampling from distribution of the copula values is done using an importance sampling approach based on Dunn and Smyth (1996) residuals, and in the M-step, maximization can be done efficiently by applying algorithms for Gaussian factor analysis.

3. Simulation Studies

Let us next compare the unconstrained ordination methods introduced in Section 2 using synthetic data inspired by the dataset `microbialdata` provided in the R package **gllvm** and introduced in Section 1. In the next, we introduce the simulations setups used in the comparisons.

3.1. Ordination score recovery

To generate data with similar overdispersion and sparsity properties, we used the following two approaches:

1. Data were generated using the generalized linear latent variable model (GLLVM) in (2) with $d = 2$, assuming negative binomial (NB) or zero-inflated negative binomial (ZINB) distribution for counts.

2. Data were generated using the Gaussian copula latent variable model (GCLVM) in (5) with $d = 2$, assuming negative binomial (NB) or zero-inflated negative binomial (ZINB) distribution for counts.

To generate data from GLLVM, we used the R package **gllvm**. For generating data from GCLVM, the code in Popovic (2021) accompaying Popovic et al. (2022) was modified accordingly. In the case of copula model, fitting marginal GLMs with row-specific and column-specific intercepts was performed using the R package **gllvm** with $d = 0$ as the package allows fitting model (4) for ZINB -distributed responses.

To mimic the properties of sparse, overdispersed count data, we obtained the true parameters for the simulation model by fitting either (zero-inflated) NB-GLLVM or (zero-inflated) NB-GCLVM to the microbial data. To study the effect of sparsity on ordination results, we first ordered the bacterial species (columns) in the data based on the number of zeros in columns so that the bacteria that was present in most samples was given in the first column. We then used subsets of the original data, retaining the same rows but including different columns. Since the columns were pre-ordered based on the number of zeros, the proportion of zeros increased as more columns were included. In our simulation, we selected the first $m = 50$ (12.5% zeros), 100 (26.7% zeros), 200 (44.6% zeros), and 400 (64.2% zeros) columns from the original data as shown in Figure 1. For each value of m and for each of the four different simulation models, we generated 500 repetitions of data. Note that we did not extend beyond $m = 400$ because the proportion of zeros in the full dataset reached 91%, meaning high sparsity and thus insufficient information in the counts (Figure 1).

We applied seven competing methods to the simulated data to compare their performance in estimating ordination scores. Four model-based approaches included GLLVM and GCLVM, each with negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions and with $d = 2$. In addition, three algorithm-based methods were included in comparisons: PCA for clr-transformed data, implemented using the R package **robCompositions** (Templ et al. 2011; Filzmoser et al. 2018); nMDS with the Aitchison distance, and nMDS with Bray-Curtis distance, computed using the R package **vegan** (Oksanen et al. 2018), which provides a fast algorithm for nMDS.

The performance of methods was assessed by computing the Procrustes distance (Hurley and Cattell 1962) between the estimated ordination scores and the true ordination scores. The Procrustes distance commonly used for assessing ordination accuracy, is the minimal squared difference (e.g., Euclidean distance) between the estimated and true ordination scores.

$$\text{Procrustes Distance} = \sum_{i=1}^{n}\sum_{r=1}^{d}\left(\tilde{u}_{ir,\text{fitted}} - u_{ir,\text{true}}\right)^2,$$

where $u_{ir,\text{true}}$ denotes the corresponding true ordination coordinate for site i and latent variable r, and $\tilde{u}_{ir,\text{fitted}}$ denotes the estimated ordination scores after the Procrustes transformation, that is, the transformation that minimizes the squared differences between the fitted scores and the true scores by employing a combination of translation, rotation, reflection, and uniform scaling. At the same time, the transformation preserves the relative geometric relationships of the scores, ensuring meaningful alignment between the configurations. In the case of GLLVMs and GCLVMs, the ordination scores are obtained as the predicted latent variables.

3.2. Simulation results

In this section, we present the simulation results only for the zero-inflated NB distribution case. The results when simulating from NB distribution are shown in Figures 6 and 7 in Appendix. Figures 3 and 4 show the Procrustes errors for each of the seven methods for each subsamples of sizes $m = 50$, 100, 200 and 400, when data are generated from GLLVM and GCLVM, respectively.

As seen in Figure 3, when the data are generated from ZINB-GLLVM model, all model-based ordination methods outperform the algorithm-based classical approaches for all sizes m. The differences between the methods are small when the number of columns m is small which means proportion of zeros is modest. However, as m increases—leading to greater sparsity in the dataset—the gap between model-based ordination methods and algorithm-based approaches becomes more pronounced. The GLLVM method outperforms the GCLVM method, however, ZINB-GLLVM outperforms NB-GLLVM only when m is small. When m is large, NB-GLLVM is slightly better than the method that was used to generate the data. This may be due to two reasons: either the zero-inflation in data is not very extreme, or ZINB-GLLVM has in high-dimensional settings just too many parameters to be estimated efficiently. When data are generated from the ZINB-GCLVM model, the ordination based on the corresponding model yields the lowest Procrustes errors (Figure 4). The differences between all four model-based ordination methods are however minimal. Interestingly, the clr-transformation followed by PCA performs almost equally well as the model-based approaches.

To conclude, model-based methods seem to maintain their robust performance even when sparsity increases, whereas the algorithm-based methods (especially those using nMDS) struggle significantly with higher proportions of zeros in the dataset. The differences between all model-based ordination methods are very small. Thus, for data with similar properties as in microbiome data used here, the NB-GLLVM seems to be a safe choice for ordination. This result is also supported by the simulation results in Figures 6 and 7 in Appendix showing a stable performance of NB-GLLVM

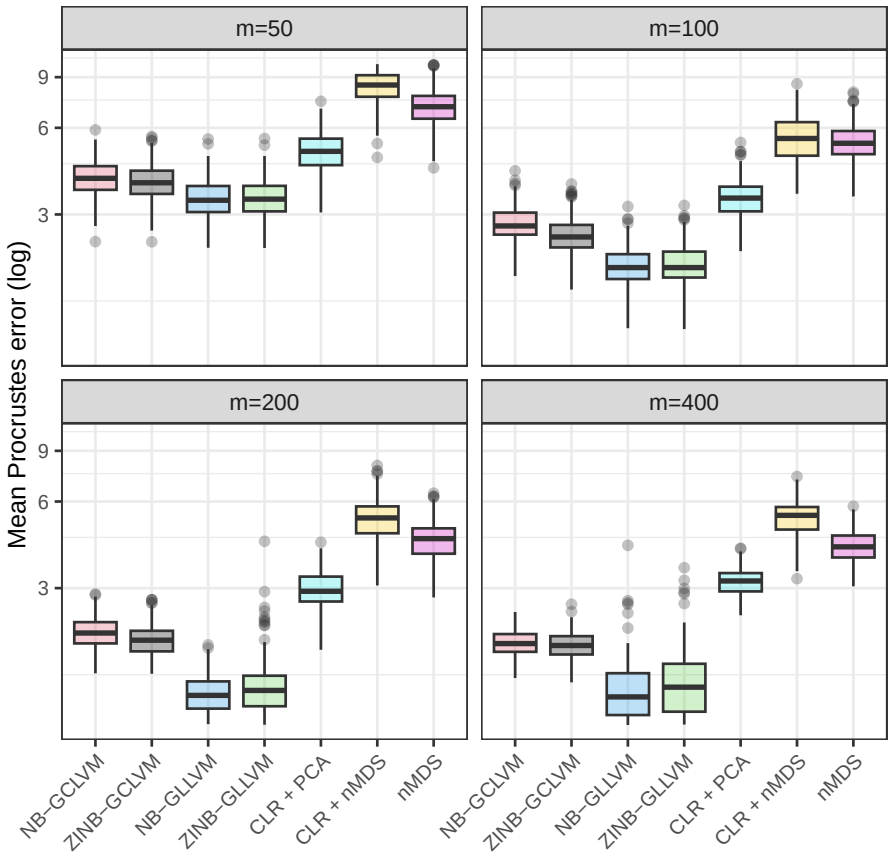

Figure 3: Comparative boxplots of Procrustes errors between the true and estimated ordination scores. The true models were GLLVMs assuming zero-inflated NB responses with $d = 2$ fitted to subsets of microbiome data of dimensions $m = 50$, 100, 200 and 400. We compared the GCLVM model assuming NB distributed responses (NB-GCLVM) and zero-inflated NB distributed responses (ZINB-GCLVM), the GLLVM model assuming NB distributed responses (NB-GLLVM) and zero-inflated NB distributed responses (ZINB-GLLVM), clr-transformation followed by PCA (CLR+PCA) and nMDS (CLR+nMDS), and nMDS without transformation.

method when data are generated from NB-GLLVM or NB-GCLVM model. What is notable in all simulation results is that, when the data dimension increases, the variation in Procrustes errors based on GLLVMs increase. This is likely attributable to the inherent computational challenges of fitting joint models in high-dimensional settings—particularly when the number of samples (n) is much smaller than the number of variables (m), i.e., $n \ll m$. As the number of variables increases, the number of parameters to be estimated grows, making it significantly more difficult to identify the global maximum of the log-likelihood. Consequently, the model-fitting process becomes increasingly complex and unstable, resulting in greater fluctuations in the Procrustes errors.

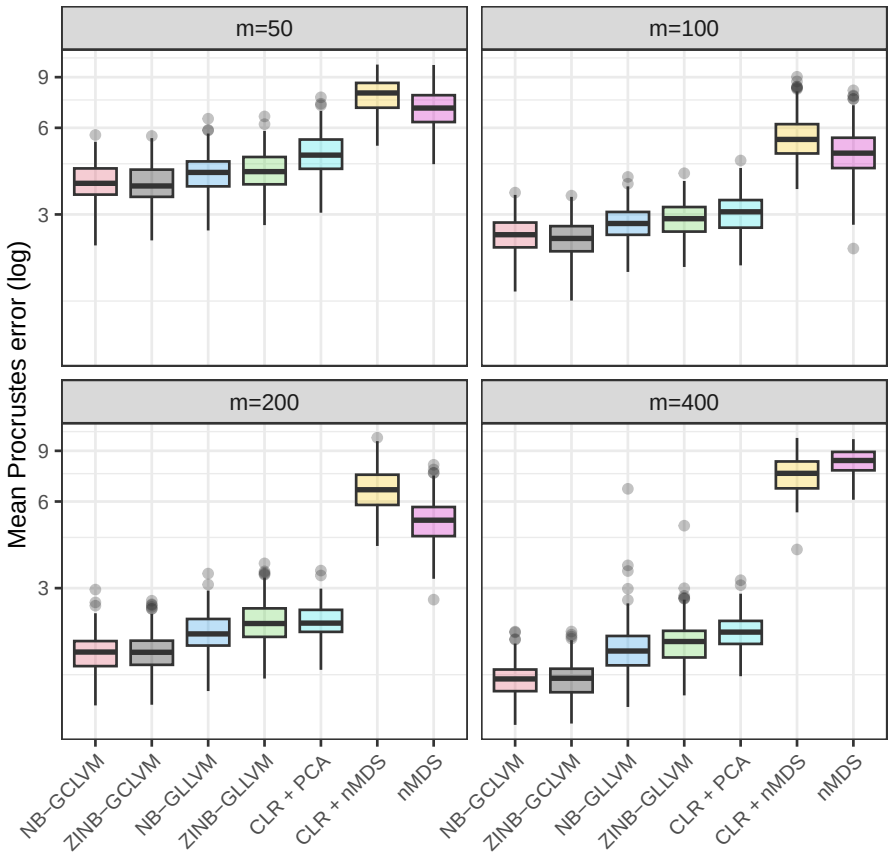

Figure 4: Comparative boxplots of Procrustes errors between the true and estimated ordination scores. The true models were GCLVMs assuming zero-inflated NB responses with $d = 2$ fitted to subsets of microbiome data of dimensions $m = 50$, 100, 200 and 400. We compared the GCLVM model assuming NB distributed responses (NB-GCLVM) and zero-inflated NB distributed responses (ZINB-GCLVM), the GLLVM model assuming NB distributed responses (NB-GLLVM) and zero-inflated NB distributed responses (ZINB-GLLVM), clr-transformation followed by PCA (CLR+PCA) and nMDS (CLR+nMDS), and nMDS without transformation.

3.3. Goodness-of-fit

In our second empirical study, we compared the goodness-of-fit of the four model-based methods under similar data scenarios as in simulation studies above. The data were generated from different models and the goodness-of-fit was measured using the following three statistics (Meynard and Quinn 2007; Liu et al. 2011).

The mean absolute range normalized error (MARNE) is defined as

$$\text{MARNE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}\frac{|y_{ij} - \hat{y}_{ij}|}{\max_j(y_{ij}) - min_j(y_{ij})}\right)$$

where y_{ij} represents the observed count, $\hat{y}_{ij}$ denotes the count predicted by the model, n is the total number of observational units (here samples) and m is the number of variables (here bacteria). A smaller MARNE value indicates a better model fit.

The global correlation (gCOR) between the predicted values and the observed values across all data points is simply calculated as the Pearson correlation coefficient between the observed and predicted counts, that is,

$$\text{gCOR} = \text{Cor}(\hat{\boldsymbol{y}}, \boldsymbol{y}),$$

where $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$, with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\top$, is a stacked response vector and $\hat{\boldsymbol{y}}$ is the corresponding predicted response vector.

The average correlation between the predicted and observed values for each individual species (or column in the dataset) (mCOR) is given by

$$\text{mCOR} = \frac{1}{m}\sum_{j=1}^{m} \text{Cor}(\hat{\boldsymbol{y}}^j, \boldsymbol{y}^j),$$

where $\boldsymbol{y}^j = (y_{1j}, \ldots, y_{nj})^\top$ are the species-wise counts and $\hat{\boldsymbol{y}}^j$ includes corresponding predictions. For both gCOR and mCOR, higher values indicate better model performance.

The goodness-of-fit values for each method in each data scenario are shown in Table 1. The results indicate GLLVMs consistently exhibit the smallest MARNE values across all values of m. When $m = 50$, the global correlation between predicted and observed counts (gCOR) is highest for copula-based methods (GCLVMs). However, when m is large, the differences in gCOR values for GLLVMs and GCLVMs are very small. When looking at mean of species-wise correlations (mCOR), the GLLVMs outperform GCLVMs in all considered cases. ZINB-GCLVM and NB-GCLVM models seem to exhibit some performance fluctuations across different data dimensions, whereas ZINB-GLLVM and NB-GLLVM models demonstrate more stable performance. A more detailed examination reveals very minimal differences between the two GLLVMs.

4. Real Data Analysis

Let us then illustrate the model-based ordination methods using the microbiome dataset `microbialdata` in the R package **gllvm** with $n = 56$ samples and $m = 986$ bacteria (Niku et al. 2024). In this example, we use the methods based on GLLVMs only as they were shown to be a bit more stable in our comparisons in Section 3.

Table 2 shows the information criteria (AIC and BIC) based on the negative binomial and ZINB models (2) with $d = 2$ and without any covariates, that is, for the purpose of unconstrained ordination. Based on the information criteria, we can say that the negative binomial model fitted data best. Figure 5 (left) shows the two unconstrained ordination scores based on the NB-GLLVM model labeled according to sampling sites (i.e., Mayrhofen, Kilpisjärvi and NyÅlesund). As seen in the figure, we can state that especially the samples taken from Kilpisjärvi are more separated from the other two locations indicating that these sites differed from Mayrhofen and NyÅlesund in terms of bacteria species composition.

As an extension of model-based unconstrained ordination method, let us next illustrate how GLLVM in (2) can be extended to perform so-called concurrent ordination that aims to infer which environmental covariates affect the community composition (Ter

Table 1: Goodness-of-fit of GLLVM and GCLVM models with $d = 2$ assuming (zero-inflated) NB responses measured using the mean absolute range normalized error (MARNE), global correlation between predicted and observed counts (gCOR) and mean of species-wise correlations (mCOR).

		MARNE	gCOR	mCOR
$m = 50$	ZINB-GCLVM	0.145	0.903	0.639
	NB-GCLVM	0.147	0.902	0.638
	ZINB-GLLVM	0.115	0.847	0.660
	NB-GLLVM	0.115	0.846	0.659
$m = 100$	ZINB-GCLVM	0.148	0.898	0.628
	NB-GCLVM	0.144	0.901	0.628
	ZINB-GLLVM	0.117	0.885	0.648
	NB-GLLVM	0.118	0.883	0.646
$m = 200$	ZINB-GCLVM	0.155	0.888	0.609
	NB-GCLVM	0.137	0.890	0.610
	ZINB-GLLVM	0.113	0.889	0.628
	NB-GLLVM	0.114	0.878	0.628
$m = 400$	ZINB-GCLVM	0.172	0.883	0.558
	NB-GCLVM	0.138	0.886	0.560
	ZINB-GLLVM	0.115	0.898	0.593
	NB-GLLVM	0.115	0.886	0.595

Table 2: The AIC and BIC values for negative binomial (NB) and zero-inflated NB GLLVMs without environmental covariates (unconstrained) and with environmental covariates (concurrent).

	unconstrained		concurrent	
	AIC	BIC	AIC	BIC
ZINB-GLLVM	130,535.4	183,722.3	130,441.7	183,673.2
NB-GLLVM	126,965.6	162,584.1	126,647.7	162,310.8

Braak and Prentice 1988). Now as environmental covariates, pH, soil organic matter (SOM), and available phosphorus (P) were recorded from each sample (Kumar et al. 2017). We proceed as in van der Veen et al. (2023) and fit a latent variable model where latent variables $\boldsymbol{u}_i$ are driven by the measured covariates with an additional set of "residual" variables that account for unmeasured drivers of species covariation. To be more specific, the model is defined as

$$\begin{aligned} g(\mu_{ij}) &= \alpha_i + \beta_{0j} + \boldsymbol{\lambda}_j^\top \boldsymbol{u}_i \\ \boldsymbol{u}_i &= \boldsymbol{B}^\top \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \end{aligned} \tag{6}$$

where $\boldsymbol{B}$ is a $k \times d$ matrix containing the reduced-rank regression coefficients and $\boldsymbol{\epsilon}_i$ are the additional d-dimensional residuals that are assumed to follow a multivariate normal distribution, $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal $d \times d$ matrix. Here, latent

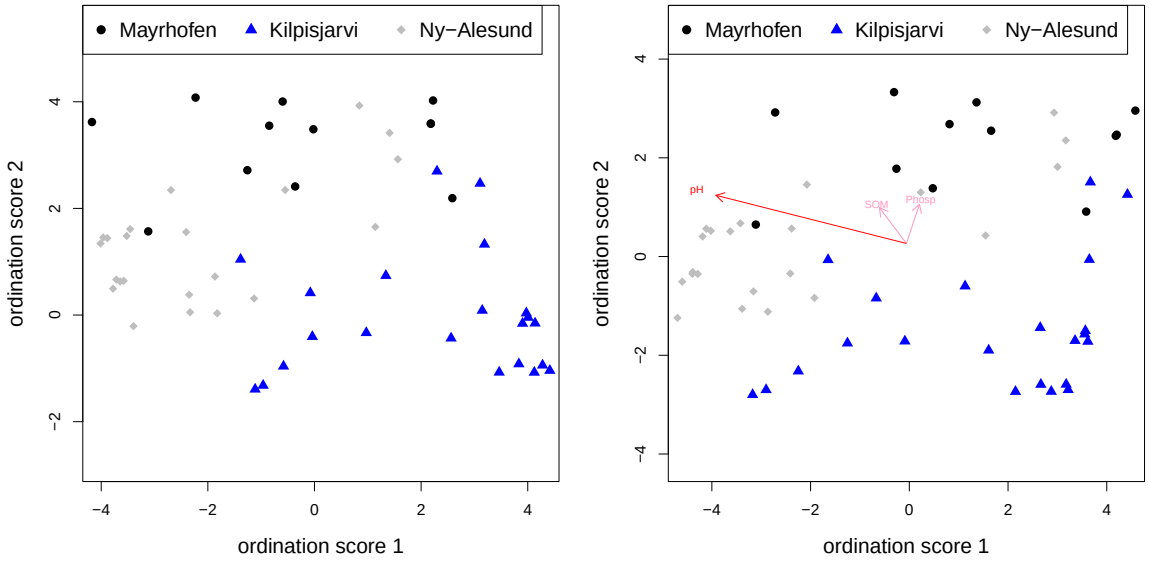

Figure 5: Model-based unconstrained (left) and concurrent (right) ordination plots based on the full microbiome dataset, both fitted with a NB-GLLVM incorporating two latent variables ($d = 2$). In the concurrent ordination, three environmental covariates ($k = 3$) were included in the model. In concurrent ordination plot, longer arrows represent covariates with the largest relative effects, and dark red arrows (here associated to pH) indicate covariates with a significant effect on ordination.

variables $\boldsymbol{u}_i$ are referred to as informed latent variables. For more details on model components, identifiability issues and how inference is performed, see van der Veen et al. (2023). Notice that the model-based concurrent ordination and many of its variants are implemented in the R package **gllvm**.

Table 2 shows the information criteria based on the negative binomial and ZINB models (6) with $d = 2$ and with (scaled) pH, soil organic matter (SOM), and available phosphorus (P) as covariates. Again, the model assuming a negative binomial distribution yields the lowest information criterion. Figure 5 (right) shows the two concurrent ordination scores based on the NB-GLLVM model (6) labeled according to sampling sites. Red arrows in the figure show the effects of environmental covariates with the longest arrow indicating the covariate with largest relative effect (here pH). The dark arrow associated to pH indicates a significant effect, that is, the 95% confidence interval of the associated slope in the matrix of reduced-rank regression coefficients $\boldsymbol{B}$ excludes zero in both dimensions.

Finally, as a diagnostic tool, we plotted the Dunn–Smyth residuals (Dunn and Smyth 1996) against the linear predictors and Q–Q plots of the residuals for NB-GLLVM models with $d = 2$, without covariates (unconstrained) and with three available covariates (concurrent). The plots in Figure 8 in Appendix show that for both models, the Dunn–Smyth residuals given by NB-GLLVMs are uniformly distributed around zero indicating that the models fit the data well. The Q–Q plots show that the residuals do not deviate from normality.

5. Conclusions

In this paper, we compared two recently developed model-based unconstrained ordination methods in the analysis of microbiome data. The first method builds upon generalized linear latent variable models (GLLVMs) and, while very flexible, the method requires fitting a joint model for high-dimensional data to account for correlation across responses (Hui et al. 2015; Warton et al. 2015). The second method combines marginal generalized linear models (GLMs) with a Gaussian copula model to address correlations across responses (Popovic et al. 2022). Our simulation results indicated that the differences between the model-based approaches are small when the data dimension and sparsity in data are moderate. The model-based approaches also perform very similarly to some algorithm-based approaches such as the one that combines the centered log-ratio (clr) transformation (Aitchison 1982) with principal component analysis (PCA). However, when the data dimension increases—leading possibly to greater sparsity in data—the gap between model-based ordination methods and algorithm-based approaches becomes more pronounced.

Between the two model-based approaches, in simulations, GLLVM was more robust in latent variable (LV) recovery under model misspecification, and attained better scores on goodness-of-fit measures. This can, at least in part, be alluded to the fact that the advantages of the copula-based method (as demonstrated e.g., in Popovic et al. 2022), are in some sense reliant on whether proper marginal models can be exploited. With sequencing data we require the observation-level effects α_i to be included in (2), in order to properly account for compositionality in count data models. For the copula approach, this, in turn, posits that the estimation should be based on a joint GLM, or on treating the data in the "long" format; both being options that lead to increased computation or memory overhead, compared to GLMs fitted on each of the species/OTUs separately.

The inherent compositionality of microbiome data—caused by finite capacities of the sequencing instruments—can be controlled for by including the aforementioned site effects α_i, as in models (2) and (5). Another approach would be to base the joint model on some distribution that is tailored for compositional count data, such as the Dirichlet-multinomial (DM) distribution (Wang 2021). Such a model however suffers from two problems. First, the model does not take into account structural zeros meaning that all zeros are assumed to be due to under-sampling, thus the zero-inflated DM distribution (Koslovsky 2023) can be a better option for sparse compositional data. Second, the DM model intrinsically imposes a negative correlation among variables which may not be very realistic assumption for biological or ecological datasets. To allow also positive correlations, models such as zero-inflated generalized DM could be used along the lines of Tang and Chen (2018).

Algorithmic compositional data analysis methods that rely on log-transformations face a significant challenge: the transformation is undefined for zero counts. In this article, we addressed this issue by adding one to each count. While this approach is simple and convenient, it implicitly assumes that none of the zero counts represent structural zeros.

Given the prevalence of zeros in the data, it seems infeasible to determine whether each zero arises from being below the detection limit or represents a structural zero. From

a biological perspective, however, structural zeros should not be disregarded, as they may hold critical importance. For instance, the presence of one bacterial species might exclude the presence of another. Consequently, modeling approaches that explicitly account for structural zeros could be more appropriate in this context.

Computational Details

The results in this paper were obtained using R 4.4.1. R itself and all packages used are either available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/ or on github and listed together with the code to reproduce all results at https://github.com/tangwenq/microbial-data.

Acknowledgments

We acknowledge the support from the Finnish Doctoral Program Network in Artificial Intelligence (AI-DOC), Decision VN/3137/2024-OKM-6, to WT, the support from the Kone Foundation to PK, JN and ST, the support from the Research Council of Finland (363261, 453691) to KN, PK and ST, respectively, and the support from the HiTEc COST Action (CA21163) to KN and ST. We also thank CSC – IT Center for Science, Finland, for providing computational resources.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–160, DOI: 10.1111/j.2517-6161.1982.tb01195.x.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, ISBN: 0412280604.

Anderson, M. J., de Valpine, P., Punnett, A., and Miller, A. E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution*, 9(6):3276–3294, DOI: 10.1002/ece3.4948.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):325–349, DOI: 10.2307/1942268.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18:117–143, DOI: 10.1111/j.1442-9993.1993.tb00438.x.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(13), DOI: 10.1186/s13059-016-0881-8.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244, DOI: 10.2307/1390802.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(15), DOI: 10.1186/2049-2618-2-15.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Compositional Data Analysis. With Worked Examples in R*. Springer, Cham, DOI: 10.1007/978-3-319-96422-5.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8:1–6, DOI: 10.3389/fmicb.2017.02224.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: Analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329, DOI: 10.1016/j.annepidem.2016.03.003.

Greenacre, M., Martínez-Álvaro, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. *Frontiers in Microbiology*, 12, DOI: 10.3389/fmicb.2021.727398.

Huber, P., Ronchetti, E., and Victoria-Feser, M. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66:893–908, DOI: 10.1111/j.1467-9868.2004.05627.x.

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., and Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, 6:399–411, DOI: 10.1111/2041-210X.12236.

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26:35–43, DOI: 10.1080/10618600.2016.1164708.

Hurley, J. R. and Cattell, R. B. (1962). The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262, DOI: 10.1002/bs.3830070216.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, ISSN: 0047-259X, DOI: https://doi.org/10.1016/j.jmva.2004.06.003.

Korhonen, P., Nordhausen, K., and Taskinen, S. (2024). A review of generalized linear latent variable models and related computational approaches. *WIREs Computational Statistics*, 16(6):e70005, DOI: 10.1002/wics.70005.

Koslovsky, M. D. (2023). A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. *Biometrics*, 79(4):3239–3251, DOI: https://doi.org/10.1111/biom.13853.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, DOI: 10.1007/BF02289565.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, DOI: 10.1007/BF02289694.

Kumar, M., Brader, G., Sessitsch, A., Mäki, M., van Elsas, J. D., and Nissinen, R. (2017). Plants assemble species specific bacterial communities from common core taxa in three arcto-alpine climate zones. *Frontiers in Microbiology*, 8:12, DOI: 10.3389/fmicb.2017.00012.

Legendre, P. and Legendre, L. (2012). *Numerical Ecology*. Developments in Environmental Modelling. Elsevier, Oxford. ISBN: 9780444538697.

Liu, C., White, M., and Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence—absence data. *Ecography*, 34(2):232–243, DOI: 10.1111/j.1600-0587.2010.06354.x.

Lubbe, S., Filzmoser, P., and Templ, M. (2021). Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, 210:104248, DOI: 10.1016/j.chemolab.2021.104248.

Lutz, K. C., Jiang, S., Neugent, M. L., De Nisco, N. J., Zhan, X., and Li, Q. (2022). A survey of statistical methods for microbiome data analysis. *Frontiers in Applied Mathematical Statistics*, 8:884810, DOI: 10.3389/fams.2022.884810.

Meynard, C. N. and Quinn, J. F. (2007). Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34(8):1455–1469, DOI: 10.1111/j.1365-2699.2007.01720.x.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Korhonen, P., Taskinen, S., van der Veen, B., and Warton, D. I. (2024). ***gllvm**: Generalized Linear Latent Variable Models*, https://CRAN.R-project.org/package=gllvm. R package version 2.0.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PLOS One*, 14:e0216129, DOI: 10.1371/journal.pone.0216129.

Niku, J., Warton, D. I., Hui, F. K. C., and Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 22:498–522, DOI: 10.1007/s13253-017-0304-7.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2018). ***vegan**: Community Ecology Package*, http://CRAN.R-project.org/package=vegan. R package version 2.5-2.

Peterson, C. B., Saha, S., and Do, K.-A. (2024). Analysis of microbiome data. *Annual Review of Statistics and Its Application*, 11:483–504, DOI: 10.1146/annurev-statistics-040522-120734.

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. *Applied and Environmental Microbiology*, 84(7):e02627–17, DOI: 10.1128/AEM.02627-17.

Popovic, G. (2021). *Fast Model-Based Ordination with Copulas: Simulation Code (v1.0.0)*, DOI: 10.5281/zenodo.5525716.

Popovic, G. C., Hui, F. K. C., and Warton, D. I. (2022). Fast model-based ordination with copulas. *Methods in Ecology and Evolution*, 13(1):194–202, DOI: 10.1111/2041-210X.13733.

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C., and Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583, DOI: 10.1111/2041-210x.13247.

Ricotta, C. and Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 31:201–205, DOI: 10.1016/j.ecocom.2017.07.003.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton, DOI: 10.1201/9780203489437.

Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., and Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. *WIREs Computational Statistics*, 15(1):e1586, DOI: 10.1002/wics.1586.

Tang, Z.-Z. and Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20:698–713, DOI: 10.1093/biostatistics/kxy025.

Templ, M., Hron, K., and Filzmoser, P. (2011). **robCompositions**: An R-package for robust statistical analysis of compositional data. *Compositional Data Analysis: Theory and Applications*, pages 341–355. John Wiley & Sons, Ltd, DOI: 10.1002/9781119976462.ch25.

Templ, M., Hron, K., Filzmoser, P., and Gardlo, A. (2016). Imputation of rounded zeros for high-dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems*, 155:183–190, DOI: 10.1016/j.chemolab.2016.04.011.

Ter Braak, C. J. and Prentice, I. C. (1988). A theory of gradient analysis. *Advances in Ecological Research*, 18:271–317, DOI: 10.1016/S0065-2504(08)60183-X.

van der Veen, B., Hui, F. K. C., Hovstad, K. A., and O'Hara, R. B. (2023). Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling. *Methods in Ecology and Evolution*, 14(2):683–695, DOI: 10.1111/2041-210X.14035.

Wang, G. (2021). Bayesian and frequentist approaches to multinomial count models in ecology. *Ecological Informatics*, 61:101209, DOI: 10.1016/j.ecoinf.2020.101209.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30:766–779, DOI: 10.1016/j.tree.2015.09.007.

Warton, D. I., Wright, S. T., and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, 3:89–101, DOI: 10.1111/j.2041-210X.2011.00127.x.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704. 10.1080/01621459.1990.10474930.

Zeng, Y., Pang, D., Zhao, H., and Wang, T. (2023). A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, 118(544):2356–2369, DOI: 10.1080/01621459.2022.2044827.

Appendix: Additional Simulations Results

Below, we present the from the simulation settings 1 and 2 in Section 3.1 for negative binomial GLLVM. Figure 8 shows the Dunn–Smyth residuals against linear predictors and Q-Q plots for the unconstrained and constrained NB-GLLVM models applied to microbial data in Section 4.

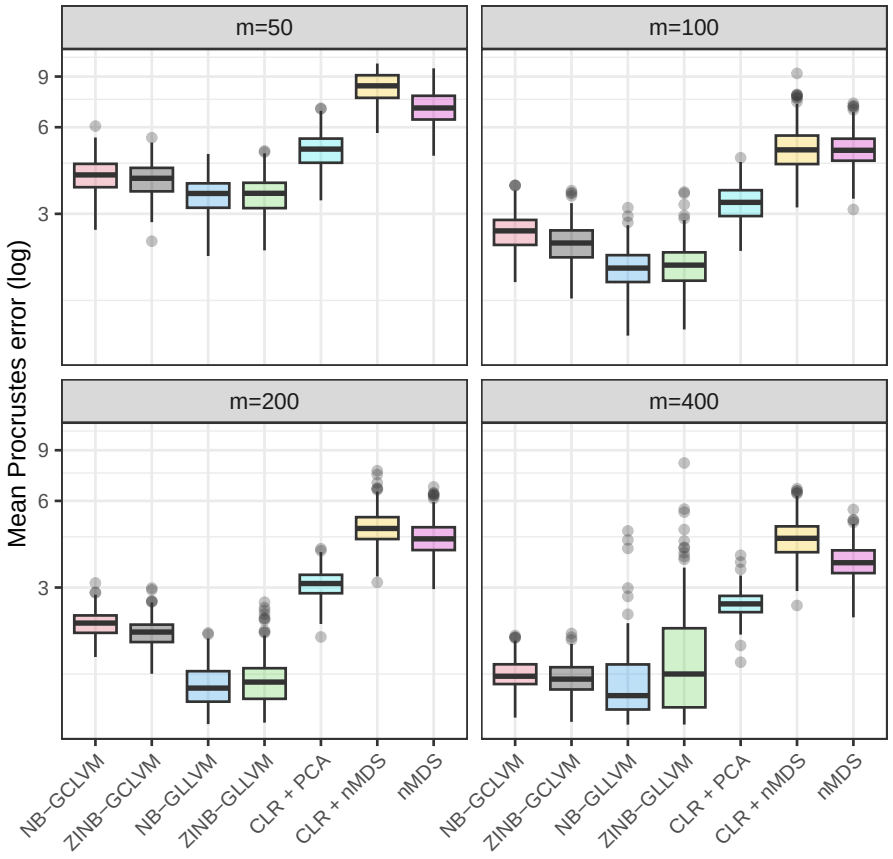

Figure 6: Comparative boxplots of Procrustes errors between the true and estimated ordination scores. The true models were GLLVMs assuming negative binomial responses with $d = 2$ fitted to subsets of microbiome data of dimensions $m = 50$, 100, 200 and 400. We compared the GCLVM model assuming NB distributed responses (NB-GCLVM) and zero-inflated NB distributed responses (ZINB-GCLVM), the GLLVM model assuming NB distributed responses (NB-GLLVM) and zero-inflated NB distributed responses (ZINB-GLLVM), clr-transformation followed by PCA (CLR+PCA) and nMDS (CLR+nMDS), and nMDS without transformation.

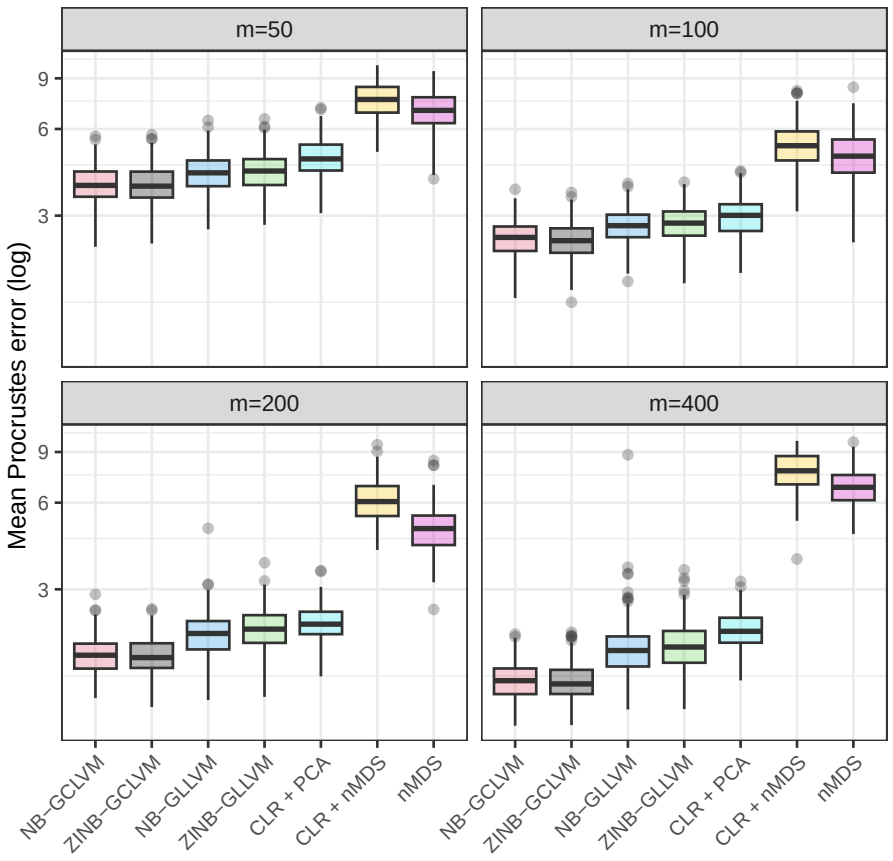

Figure 7: Comparative boxplots of Procrustes errors between the true and estimated ordination scores. The true models were GCLVMs assuming zero-inflated negative binomial responses with $d = 2$ fitted to subsets of microbiome data of dimensions $m = 50$, 100, 200 and 400. We compared the GCLVM model assuming NB distributed responses (NB-GCLVM) and zero-inflated NB distributed responses (ZINB-GCLVM), the GLLVM model assuming NB distributed responses (NB-GLLVM) and zero-inflated NB distributed responses (ZINB-GLLVM), clr-transformation followed by PCA (CLR+PCA) and nMDS (CLR+nMDS), and nMDS without transformation.

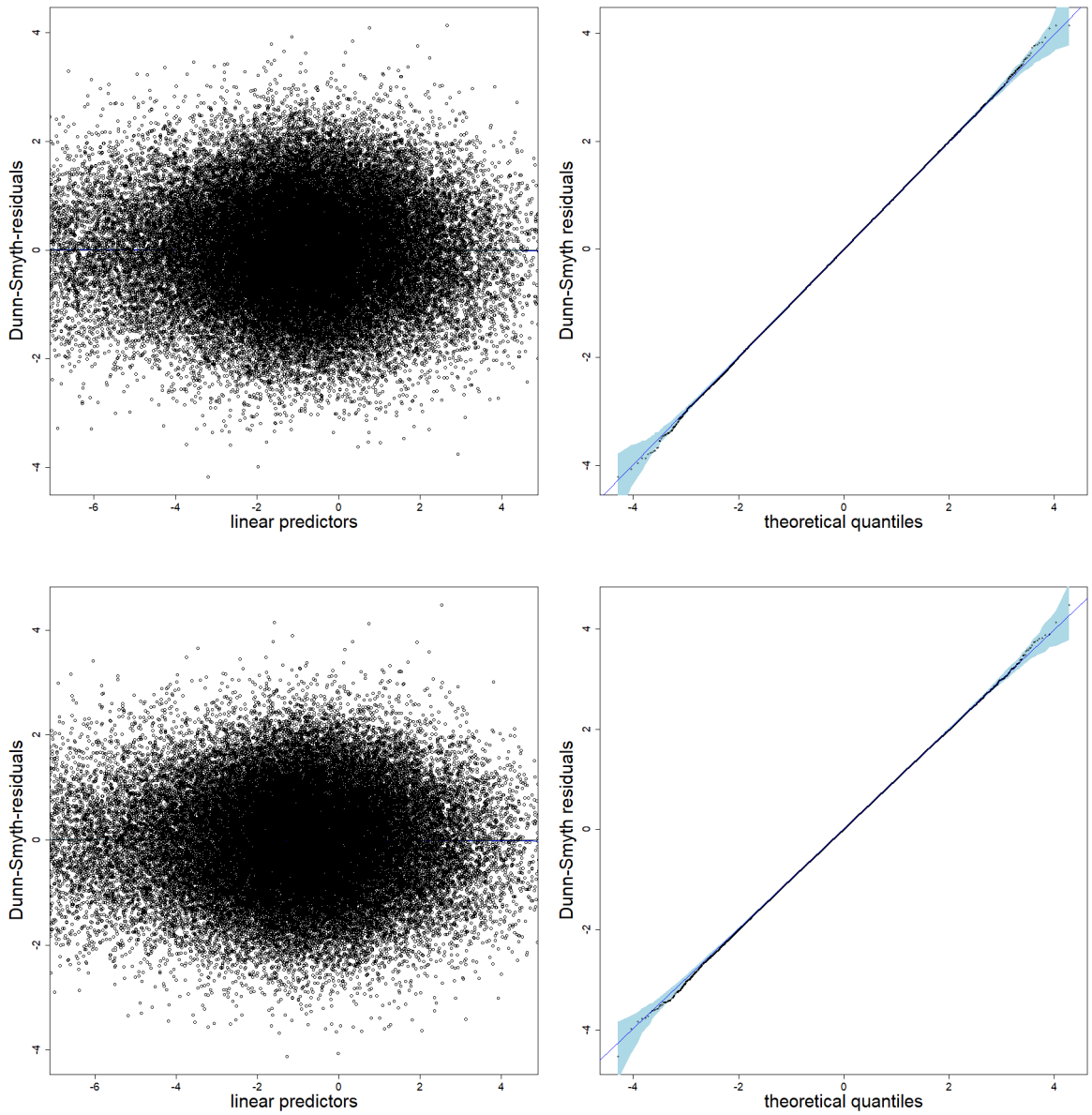

Figure 8: Diagnostic plots for the model-based unconstrained (top row) and concurrent (bottom row) ordination based on the full microbiome dataset, both fitted with a NB-GLLVM incorporating two latent variables ($d = 2$). In the concurrent ordination three environmental covariates ($k = 3$) were included in the model. The left panel displays Dunn-Smyth residuals plotted against linear predictors, and the right panel features a Q-Q plot of the Dunn-Smyth residuals, with the blue region representing the 95% confidence interval for evaluating residual distribution.

Affiliation:

Wenqi Tang, Pekka Korhonen, Jenni Niku,
Department of Mathematics and Statistics
FI-40014 University of Jyväskylä, Finland
E-mail: wenqi.t.tang@jyu.fi

Pekka Korhonen
Department of Mathematics and Statistics
FI-40014 University of Jyväskylä, Finland
E-mail: pekka.o.korhonen@jyu.fi

Jenni Niku
Department of Mathematics and Statistics
FI-40014 University of Jyväskylä, Finland
E-mail: jenni.m.e.niku@jyu.fi

Klaus Nordhausen
Department of Mathematics and Statistics
FI-00014 University of Helsinki, Finland
E-mail: klaus.nordhausen@helsinki.fi

Sara Taskinen
Department of Mathematics and Statistics
FI-40014 University of Jyväskylä, Finland
E-mail: sara.l.taskinen@jyu.fi

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
May 2025, Volume V, Issue VI *Submitted:* 2024-12-20
doi:10.52933/jdssv.v5i6.133 *Accepted:* 2025-04-07