*Bioinformatics*, 31(10), 2015, 1648–1654
doi: 10.1093/bioinformatics/btv032
Advance Access Publication Date: 20 January 2015
Original Paper

OXFORD

Data and text mining

# GrammR: graphical representation and modeling of count data with application in metagenomics

**Deepak Nag Ayyala and Shili Lin***

Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on August 12, 2014; revised on December 16, 2014; accepted on January 14, 2015

## Abstract

**Motivation**: Microbiota compositions have great implications in human health, such as obesity and other conditions. As such, it is of great importance to cluster samples or taxa to visualize and discover community substructures. Graphical representation of metagenomic count data relies on two aspects, measure of dissimilarity between samples/taxa and algorithm used to estimate coordinates to study microbiota communities. UniFrac is a dissimilarity measure commonly used in metagenomic research, but it requires a phylogenetic tree. Principal coordinate analysis (PCoA) is a popular algorithm for estimating two-dimensional (2D) coordinates for graphical representation, although alternative and higher-dimensional representations may reveal underlying community substructures invisible in 2D representations.

**Results**: We adapt a new measure of dissimilarity, penalized Kendall's $\tau$-distance, which does not depend on a phylogenetic tree, and hence more readily applicable to a wider class of problems. Further, we propose to use metric multidimensional scaling (MDS) as an alternative to PCoA for graphical representation. We then devise a novel procedure for determining the number of clusters in conjunction with PAM (mPAM). We show superior performances with higher-dimensional representations. We further demonstrate the utility of mPAM for accurate clustering analysis, especially with higher-dimensional MDS models. Applications to two human microbiota datasets illustrate greater insights into the subcommunity structure with a higher-dimensional analysis.

**Availability and implementation**: GrammR is implemented as an R-package available at http://www.stat.osu.edu/~statgen/SOFTWARE/GrammR/. It may also be downloaded from http://cran.rproject.org/web/packages/GrammR/.

**Contact**: shili@stat.osu.edu

**Supplementary information**: Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

The underlying subcommunity structure of data provides vital information about the diversity of microbial composition within the communities. It is well established that human microbiota samples collected from different body habitats show differences in bacterial composition (Costello *et al.*, 2009; Turnbaugh, 2009). Microbiota in the same body habitat (say gut) may also vary wildly within and across populations, which may have great implications in human health, such as on obesity and other conditions (Ridaura *et al.*, 2013; Subramanian *et al.*, 2014; Yatsunenko *et al.*, 2012). As such, it is of great importance to cluster samples or taxa to visualize and discover community substructures. Assigning the recorded gene sequences into different taxonomic ranks yields a matrix consisting of non-negative integer values, typically known as the metagenomic count matrix.

Given a measure of distance between samples (e.g. dissimilarity between counts in different rows; see below), principal coordinate

analysis (PCoA) is a standard technique to place the samples on an Euclidean coordinate space. The aim is to represent the samples as points in a coordinate space so that under a specified metric, the distance between any two points is representative of the dissimilarity between the samples which they represent. The amount of information in the dissimilarity that is conveyed by PCoA is through the principal components (PCs), which increases as the dimension of the coordinate space (number of PCs) increases. However, in practice, a vast majority of researches construct only two-dimensional (2D) PCoA models for ease of presentation (Costello *et al*., 2009; Holmes *et al*., 2012; Ley *et al*., 2008; Lozupone and Knight, 2007).

In several studies that used PCoA models, the amount of variance explained by the first two (PCs) is very small and does not exceed 30% (Costello *et al*., 2009; Lozupone and Knight, 2007). This is because the dimension of the distance matrix is equal to the sample size, which is typical high, and thus it is very unlikely that two or three directions can explain a huge proportion of the variability. An alternative to PCoA is metric multidimensional scaling (MDS), which is modeled by optimizing the difference between true dissimilarity and distance between estimated coordinates, where the distance estimate is determined by the metric specified. In fact, when the metric is Euclidean, metric MDS is equivalent to PCoA, thus metric MDS is a general class of dimension reduction models that includes PCoA as a special case. For different metrics, the shape of representation of the data is different. For example, in two dimensions, the shape of the PCoA model is a circle while that of the $l_1$-metric based MDS model is a square. Studying graphical representation of the samples under different norms might reveal features in the data that are not detected by PCoA.

An important aspect that drives PCoA representation of the data is the measure of dissimilarity between samples. A popular measure of dissimilarity used in the metagenomic research community is UniFrac distance (Lozupone and Knight, 2005), which is calculated by placing the samples on a phylogenetic tree and counting the number of unshared branches between the two samples. Calculation of UniFrac distance requires the phylogenetic information in addition to the taxonomic count matrix. As such, UniFrac is the measure of choice in studies where the 16S rRNA gene sequences are recorded because the phylogenetic tree, an important feature of the 16S data, can reveal relevant biological signals when used appropriately. This approach, however, is not practicable for studies in which the phylogenetic tree is not readily available, such as in shotgun sequencing, where one can only construct Operational Taxonomic Units (OTUs) that best represent the samples, for example, through matching the shotgun sequences to a reference database. This calls for the development of a dissimilarity measure that can be used in the absence of a phylogenetic tree.

Since the data matrix comprises integer-valued counts with some extreme values, regular $l_p$-norms may not be appropriate. As an alternative, we propose to use Kendall's τ-distance as the measure of dissimilarity. While τ-distance is originally proposed to compare two sets of rankings, it can be easily extended to counts. Properties of the Kendall's τ-distance are studied extensively and can be penalized to be applicable even when there are ties (Fagin *et al*., 2004). Depending on the penalty imposed, the Kendall's distance may also be a metric. Hence, metric MDS models can be constructed without compromising the information.

Although, MDS models are used mainly for graphical representation of data, they can also be used to study clustering. Given the true community membership of samples, it is of interest to see if the graphical models can visually separate the clusters. For example, in the metagenomic count data described in Costello *et al*. (2009), body habitats from which the samples are collected can be used to aid visualization of the aptness of the estimated clusters. For the gut microbiome data of obese twins (Holmes *et al*., 2012; Turnbaugh, 2009), as another example, attributes of the individuals such as ethnicity, zygosity and obesity can be used to cluster samples to visually correlate with estimated clusters based on metagenomic count data. An interesting question that arises is to see how accurately clustering techniques such as partitioning around medoids (PAM) perform as a follow-up step after metric MDS modeling.

When clustering using PAM, the optimal number of clusters is determined using the average silhouette width. Standard practice is to use the number of clusters that results in maximum silhouette width (Rousseeuw, 1987). This, however, can result in introduction of uninformative clusters with minute increase in the silhouette width. As an illustration, consider Supplementary Figure S1 which shows the silhouette width plot constructed for a dataset, whose generating model was constructed based on the Costello data as detailed in Section 3. The graph shows that maximum silhouette width is not the best representative of the optimal number of clusters for this dataset. For example, consider the 4D $l_1$-MDS model in the graph. The maximum silhouette width (0.8709) is obtained at 18 clusters. However, the model attains a width of 0.8206 for 10 clusters (the last big jump before the increases become incremental), gaining only 5.02% in width while creating 8 additional clusters. Similar phenomenon is observed for all the PCoA models.

In this article, we propose to use the Kendall's τ-distance as a measure of dissimilarity for metagenomic count data when a phylogenetic tree is not readily available. We then construct metric MDS models for varying dimensions (focusing on dimensions 2–4 for feasibility of visualization) to study their aptness for representing metagenomic community substructures. A modified silhouette width plot in conjunction with PAM, mPAM, is proposed as a novel measure for determining the number of clusters that discounts incremental increases. These results are compared with their counterparts based on the PCoA models. We would like to emphasize that the primary aim of this article is not to determine which of these methods is the best unequivocally, but to carry out a comparison between them to make informed recommendations. As an illustration to demonstrate the methods, especially the outcomes when a higher dimension is used, we applied the methods to analyzing two datasets.

## 2 Methods

### 2.1 Metagenomic count matrix

For a dataset with $n$ samples, let $\mathbf{X} = [X_{ij}]_{1 \le i \le n, 1 \le j \le K}$ denote the metagenomic count matrix with $K$ 'effective' OTUs. That is, $X_{ij}$ represents the number of gene sequences from the $i$th sample that are classified as belonging to the $j$th OTU. To reduce dimension of the count matrix, OTUs that have no recorded observations for all the samples can be ignored, resulting in $K$ effective OTUs that have at least one sample with non-zero count. For ease of presentation and focused simulation study, we describe the notation and proposed methodologies for clustering samples. Of equal interest is clustering OTUs to reveal microbiota community substructures. This can be accomplished by simply transposing the $X$ matrix; the proposed methodologies apply as we demonstrate in an example in Section 4.1.

### 2.2 Kendall's distance

Kendall's τ-distance is a measure of dissimilarity between two samples when the observations within samples are integer valued.

Originally proposed as a measure of dissimilarity between two ranking systems with no ties (Kendall, 1938), the distance can be extended to samples having non-integer values. Further, several extensions have been proposed for calculation of the τ-distance when there are ties (Bansal and Fernández-Baca, 2009). We use the version in Fagin *et al.* (2004), which defines the τ-distance between any two samples $X_i$ and $X_j$ as

$$\tau(X_i,X_j)=\frac{1}{\binom{K}{2}}\sum_{1\le u<v\le K}\Big[\mathcal{I}\Big((X_{iu}-X_{iv})(X_{ju}-X_{jv})<0\Big)+\lambda(\mathcal{I}(X_{iu}=X_{iv})\mathcal{I}(X_{ju}\neq X_{jv})+\mathcal{I}(X_{iu}\neq X_{iv})\mathcal{I}(X_{ju}=X_{jv}))\Big] \quad (1)$$

where $0\le\lambda\le 1$ is a tuning parameter and $\mathcal{I}$ is the usual indicator function. The τ-distance measures the number of pairs that do not agree in their ordering in the samples (first term) and, with a penalty, the number of pairs that are tied in only one of the samples. Properties of the τ-distance are dictated by the tuning parameter λ. The τ-distance is a proper distance metric only if $\frac{1}{2}\le\lambda\le 1$ (Fagin *et al.*, 2004). To be able to use metric MDS with a norm that satisfies the triangle inequality, we restrict ourselves to $\lambda\ge\frac{1}{2}$. In our simulation and data analysis, we set $\lambda=1/2$. The resulting $n\times n$ matrix of dissimilarities $\mathbf{D}=(D_{ij})_{1\le i,j\le n}$ is symmetric, with $D_{ij}=\tau(X_i,X_j)$.

## 2.3 Metric MDS

PCoA involves calculating the largest eigenvalues of the distance matrix. To construct a $d$-dimensional PCoA model, the first $d$ eigenvectors are used to determine the coordinates. An alternative to this approach is metric MDS, where the aim is to represent the samples on a $d$-dimensional Euclidean space as vectors $\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_n$ so that for a given norm $||\cdot||$, $||\mathbf{Y}_i-\mathbf{Y}_j||$ approximates the dissimilarity $D_{ij}$. This is achieved by minimizing the stress function

$$\mathcal{S}=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\|\mathbf{Y}_i-\mathbf{Y}_j\|_p-D_{ij}\right)^2. \quad (2)$$

Since $\mathcal{S}$ increases with the number of samples, its standardized form, $\frac{\mathcal{S}}{\sum_{i,j}||\mathbf{Y}_i-\mathbf{Y}_j||_p}$, is used as the convergence criterion. This approach gives the freedom to use different norms to study the representation of the data on the coordinate system. The dimension $d$ is usually set to be small (less than or equal to 4) for ease of graphical representation. Regarding the norm, any $l_p\,(p>0)$ norm can be used for metric MDS, where the $l_p$ norm is given by $||\mathbf{Y}_i-\mathbf{Y}_j||_p=(\sum_{m=1}^{d}|Y_{im}-Y_{jm}|^p)^{1/p}$. Typical choices are Manhattan (i.e. $l_1$) and Euclidean (i.e. $l_2$) norms. In the current study, we focus on metric MDS models constructed using $l_1$ norm ($l_1$-MDS) and $l_2$ norm (PCoA). The fact that metric MDS with $l_2$ norm reduces to the PCoA model was mentioned earlier.

## 2.4 Clustering

Graphical visualization of MDS models is a very basic exploratory tool to study clustering of samples. However, visual demarcation of clusters is not accurate and is statistically invalid. Standard clustering tools such as $k$-means and $k$-nearest neighbors require the number of clusters to be provided prior to constructing the clusters. In some studies, the samples may be clustered prior to analysis using other attributes recorded. For example, when studying the bacterial composition of various body habitats, the samples can be taken as clustered by the sites from which they are collected. However, to study the composition of subcommunities in an unbiased fashion, we need to determine the optimal number of clusters to be used. As such, we propose a modified version of PAM (Kaufman and Rousseeuw, 1987), from which the optimal number of clusters can be determined internally using the silhouette plot (Rousseeuw, 1987).

Average silhouette width is calculated by varying the number of clusters over a finite range and selecting the value for which the maximum average width is attained. This may however result in over-estimation of the number of clusters, due to insignificant amount of increase in the silhouette width contributed by sub-clustering, as we saw in Supplementary Figure S1. To overcome this shortcoming so that the optimum number of clusters is more robust to futile sub-clustering, we propose a novel automated procedure for selection of optimal number of clusters, leading to the mPAM procedure for clustering. The proposed method works as follows. Let $(w_m,w_{m+1},\ldots,w_{c-1})$ denote the average silhouette widths corresponding to the number of clusters being $m,m+1,\ldots,c-1$, where the minimum number of clusters, $m$, and the maximum number of clusters, $c-1$, are problem dependent and can be specified by users. Let $W=(w_{[1]},w_{[2]},\ldots)$ denote the vector of widths sorted in descending order such that all in $W$ are within $\psi\times w_{[1]}$ of the maximum width $w_{[1]}$ and with $s_1,s_2,\ldots$ being the corresponding numbers of clusters such that $s_i<s_1$. We select $s_1$ as the optimal number of clusters only if $w_{[1]}$ is significantly larger than the rest of the silhouette width (i.e. $W$ only contains $w_{[1]}$. Otherwise, we find $w_{[i]}=\min_j\{(w_{[j]}-w_{[1]})/(s_j-s_1)\}$, which represents the minimum loss of information per step, and the corresponding $s_i$ is taken to be the optimal number of clusters. Note that ψ serves as a tuning parameter controlling the size of the number of clusters. We would recommend setting ψ to be a small number (say between 0 and 0.1) to balance over-estimation by PAM (when $\psi=0$) and under-estimation (if ψ is set too large), as we discuss in more detail below.

## 2.5 Misclassification error (MCE) measure

If the true clustering of observations is known, as in a simulation study or based on other attributes of the samples, performance of the optimal clusters determined by mPAM can be assessed using the Rand Index (Hubert and Arabie, 1985). For a dataset with $n$ samples, suppose that the optimal number of clusters determined by mPAM is $b$ and the true number of clusters in the dataset is $a$. Let $\mathcal{C}_1,\mathcal{C}_2,\ldots,\mathcal{C}_a$ be the true clusters that partition the indices $\{1,2,\ldots,n\}$ and $\mathcal{D}_1,\mathcal{D}_2,\ldots,\mathcal{D}_b$ denote the clusters determined using mPAM. The estimated clusters are different from the true if a pair of observations in the same true cluster are in different estimated clusters or vice versa. The following quantity, MCE, essentially 1 - RandIndex, can be used to assess the performance of a clustering scheme:

$$MCE=\frac{2}{n(n-1)}\sum_{1\le i<j\le n}(T_{i,j}+E_{i,j}), \quad (3)$$

where

$$T_{i,j}=\sum_{\substack{1\le p\le a\\1\le q\le b}}\mathcal{I}(\{X_i,X_j\}\subset\mathcal{C}_p)\mathcal{I}(\{X_i,X_j\}\not\subseteq\mathcal{D}_q),$$

$$E_{i,j}=\sum_{\substack{1\le p\le a\\1\le q\le b}}\mathcal{I}(\{X_i,X_j\}\not\subseteq\mathcal{C}_p)\mathcal{I}(\{X_i,X_j\}\subset\mathcal{D}_q).$$

This measure is equal to zero if and only if the true and estimated clusters are equivalent. Both over- and under-estimation of the

number of clusters will result in a positive *MCE* value; the larger the MCE value, the worse the performance.

## 3 Simulation study

We conducted a simulation study to compare the performance of PCoA and $l_1$-MDS with a dimension between two and four for a number of factors: estimation of the number of clusters, classification errors and graphical representation for visualization (i.e. tightness of clusters). We consider the following two models for generating metagenomic count data.

**Model I.** In this model, 8 clusters were generated consisting of 40 samples each. The number of OTUs is set to be 200. Within each cluster, each OTU was randomly assigned a zero or non-zero count with a probability of 0.7 and 0.3, respectively. That is, the counts for the $k$th feature of the $j$th sample in the $i$th cluster were generated as

$$X_{ijk} \sim U_{ik}NB(\lambda_{ik}, 1) + (1 - U_{ik})NB(10, 0.01),$$
$$\lambda_{ik} \sim NB(100, 1), U_{ik} \sim \text{Bernoulli}(0.3),$$

for $1 \le i \le 8, 1 \le j \le 40, 1 \le k \le 200$.

In words, for any given cluster, each OTU is selected to have non-zero count with probability 0.3 and held fixed for all samples within the cluster. The non-zero counts are sampled to have means coming from a negative binomial distribution, and the actual counts are in turn generated from a negative binomial distribution. Further, to incorporate sequencing errors for OTUs with zero counts, the zero counts are modeled as a zero-inflated negative binomial variable, given by the second term in the expression for $X_{ijk}$.

**Model II.** We used the metagenomic count data of Costello *et al.* (2009) to more realistically model the parameters for generating random samples in this construction. The Costello *et al.* study considered bacterial community across body habitats. Variable region V2 of the 16S rRNA gene was amplified to obtain highly classifiable gene sequences which were then matched to a reference genome to generate the metagenomic count data. We used the counts at the family level of taxonomy to model our parameters. The dataset consists of a total of 312 taxa and we modeled counts from these taxa as follows.

We considered the following 10 body sites: hair, gut, external auditory canal, nostril, oral cavity, left/right axilla, palm, popliteal fossa, plantar foot and glans penis or labia minora. For each body site, we assume the counts for each taxa follow a negative binomial distribution and estimate the mean ($\mu$) and dispersion parameter ($\phi$). We screened out taxa with total count less than 5 and after estimating the negative binomial parameters, we selected, for each body site, the 40 taxa with the highest means as representatives for that body site. This leads to a clear distinction between the clusters (body sites). For the remaining 272 taxa, we used a zero-inflated negative binomial with mean 1 and dispersion $10^6$. This is done to reflect the sparse nature of count data, and the non-zero counts add noise into the model and reduce the ability to demarcate the clusters. The model can be expressed as

$$\text{Representative taxa} : X_{ijk} \sim NB(\mu_{ik}, \phi_{ik}), \quad k = i_1, \ldots, i_{40},$$
$$\text{Noise taxa} : X_{ijk} \sim Z_{ijk}NB(1, 10^6),$$
$$Z_{ijk} \sim \text{Bernoulli}(0.3), \quad k = i_{41}, \ldots, i_{312},$$

for $1 \le i \le 10$ body sites and $1 \le j \le 20$ samples. Note that different body sites have different sets of representative taxa, and therefore the indexes for the top forty taxa, $(i_1, \ldots, i_{40})$, differ from body site to body site, hence the notation.

### 3.1 Estimation accuracy and visualization

We calculated the optimal number of clusters using mPAM for both $l_1$-MDS and PCoA, each for dimensions 2, 3 and 4. To quantify the clustering performance of the various models, we also calculated the misclassification errors, *MCE*. Boxplots for both measures based on 1000 randomly generated samples are presented in Figure 1. Under both simulation models, PCoA underestimates the number of clusters, although this downward bias gets smaller as the dimension increases. $l_1$-MDS also underestimates the number of clusters under model II when the dimension is small. Though, we note that the underestimation is not as severe as with PCoA, with 4D $l_1$-MDS correctly recovering the true number (10) as its median over the 1000 replications. Further, PCoA models have much higher MCE than their $l_1$-MDS counterparts. For both methods, the MCE decreases as the number of dimension increases, though.

For a random realization of data generated under Model I, the average silhouette width plots are given in Supplementary Figure S2, from which one can see that PAM and mPAM will lead to the same estimations of the numbers of clusters for all models. 2D models and a snapshot showing the 3D models are presented in Figure 2. The two 2D plots display over- and under-estimation of the true number of clusters under the $l_1$-MDS and the PCoA models, respectively. This is shown in the graphs by encircling the misrepresented clusters. While the two clusters encircled in the $l_1$-MDS model correspond to a single true cluster, the encircled cluster in the PCoA model contains three true clusters. However, this error in estimation is not seen in the 3D models, with both $l_1$-MDS and PCoA producing the correct number of clusters without any misclassification (Supplementary Table S1). However, for both the 2D and 3D representations, the clusters from the $l_1$-MDS models are much tighter than their PCoA counterparts. Other graphical representations for the 2D and 3D models as well as

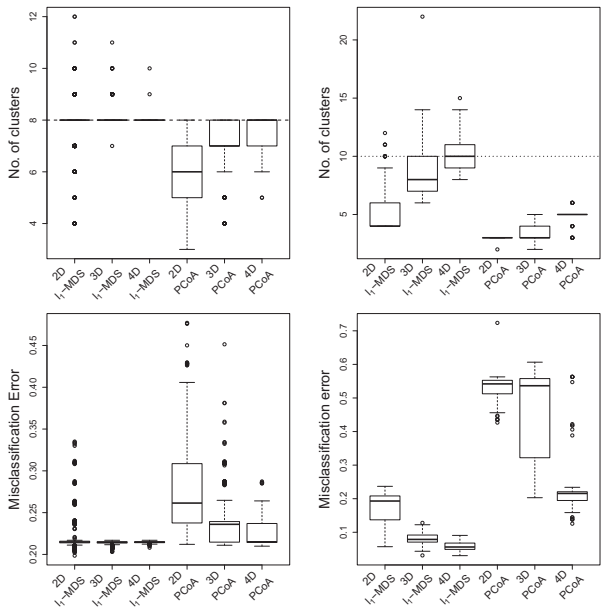


**Fig. 1.** Comparative boxplots showing the number of estimated clusters (top row) and the misclassification error (bottom row) for $l_1$-MDS and PCoA constructed for dimensions 2-4. The dotted line in the boxplots portaiting number of clusters corresponds to the true number of clusters used for data generation. The left and right columns are for Models I and II, respectively

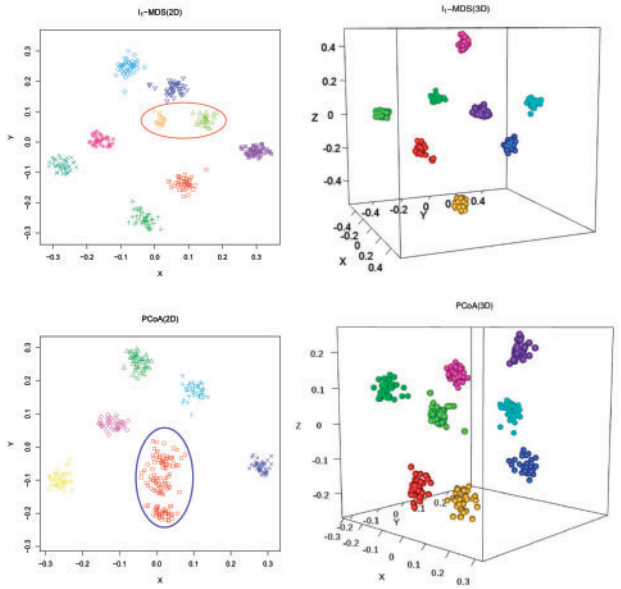


**Fig. 2.** Graphical representations of a random realization of data generated using Model I with 8 clusters. The first row presents the mPAM clustering results from the 2D and 3D $l_1$-MDS models, while the second row contains the corresponding results from the PCoA models. The *x*, *y*, and *z* (if exists), are the coordinates from the $l_1$-MDS or PCoA models. The different gray levels (and colors in the online version) represent different clusters predicted by mPAM. Note that the 3D plot is a snapshot, with interactive view available from the software website provided

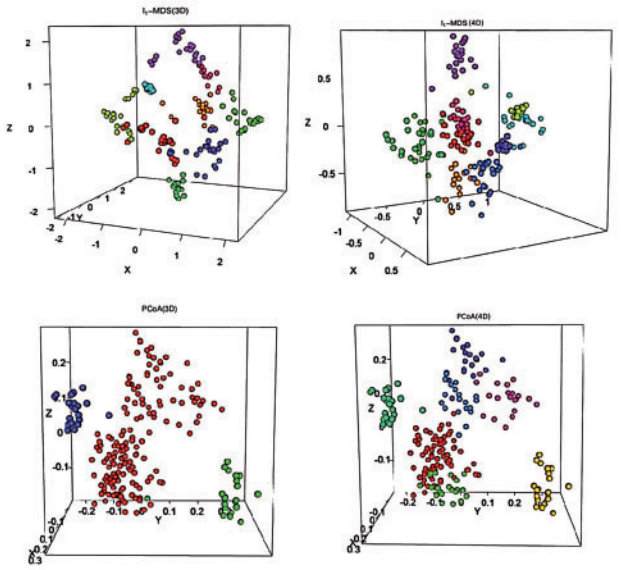


**Fig. 3.** Graphical representations of a random realization of data generated using Model II with 10 clusters. The first row presents the mPAM clustering results from the 3D and 4D $l_1$-MDS models, while the second row contains the corresponding results from the PCoA models. The *x*, *y*, and *z*, are the coordinates from the $l_1$-MDS or PCoA models. The different gray levels (and colors in the online version) represent different clusters predicted by mPAM. Note that the plots are snapshots (and only a single 3D projection for the 4D model), with interactive views available from the software website provided

those for the 4D models lead to similar observations (Supplementary Materials S1 and Figs. S3–S10).

The 3D and 4D visualization for a randomly generated dataset under model II are provided in Figure 3. While the corresponding silhouette width plots are given in Supplementary Figure S11. Additional information and plots are also available in Supplementary Materials S2 and Figures S12–S19. For $l_1$-MDS, the 2D model with mPAM underestimates the number of clusters, while PAM overestimates. In contrast, the 3D and 4D models provide good estimates of the number of clusters under mPAM (9 and 10 for 3D and 4D, respectively), whereas PAM still leads to an overestimation for the 3D model (13 clusters). Conversely, the PCoA models grossly underestimate the number of clusters, even for 3 and 4 dimensions, leading to large MCE (see Supplementary Table S2 for details). Note that for PCoA, as we can see from Supplementary Figure S11, PAM would provide the same estimates as mPAM, indicating underestimation is an inherent feature of PCoA, not the measure for cluster estimation.

The results for both simulation models demonstrate that $l_1$-MDS outperforms PCoA when the measure of dissimilarity is the Kendall's $\tau$-distance. To see if these results are invariant of the distance measure used, we used a different measure, the Jensen-Shannon divergence metric, as an alternative. The Jensen-Shannon divergence is commonly used for count data in bioinformatics (Romero *et al.*, 2014) and genome-comparison (Sims *et al.*, 2009). The results, presented in Supplementary Materials S3 and Figure S20, are similar to those shown in Figure 1. This substantiates our conclusion that $l_1$-MDS outperforms its PCoA counterpart in terms of MCE and estimation of number of clusters.

### 3.2 mPAM versus PAM

The proposed method for selecting the optimal number of clusters, mPAM, is driven by the fact that the maximum average silhouette width, in some cases, is not the best measure for selecting the number of clusters (see Supplementary Figs. S1, S11), in which PAM may (when PAM and mPAM do not agree) over estimate the number of clusters due to insignificant amount of increase in silhouette width. Hence, we expect the mPAM estimates to be smaller than those of PAM. To gain a better understanding of the relative performance of those two measures, we repeated the analysis of model II, but this time using PAM for estimating the number of clusters. Side-by-side barplots of the relative frequencies of biases in the estimated numbers of clusters (mPAM-10 and PAM-10) are given in Figure 4 (for 4D $l_1$-MDS with two different $\psi$ values, 0.01 and 0.1). As we can see from the figure, mPAM clearly outperforms PAM (which can lead to large positive bias), and the results are not very sensitive to the selection of the tuning parameter $\psi$ for 4D $l_1$-MDS. Additional details for comparison between mPAM and PAM and plots with various $\psi$ values and dimensions are provided in Supplementary Materials S4 and Figures S21 and 22. Although Figure 4 shows little sensitivity of the results to $\psi$, in general, as mentioned earlier, $\psi$'s specification is related to the selection of the optimal number of clusters. More specifically, $\psi = 0$ leads to the PAM estimator (potential overestimation); as $\psi$ increases, the estimated number of clusters decreases in general (Fig. S23), leading to the recommendation of a small $\psi$ to balance over- and under-estimation.

## 4 Two data analysis examples

To demonstrate the utility of GrammR, we applied it to two datasets: the human body habitats of Costello *et al.* (2009) and the fecal data of Yatsunenko *et al.* (2012). For the Costello data, we will not only illustrate clustering of samples, but also clustering of taxa.

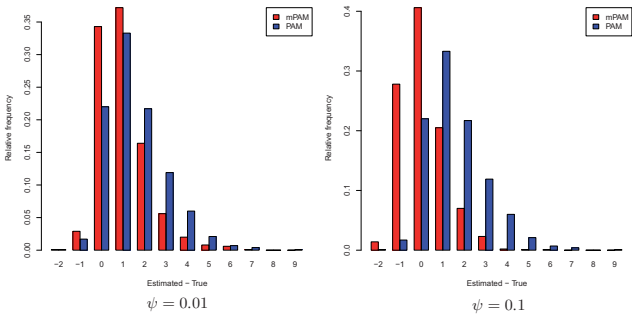


**Fig. 4.** Side-by-side barplots of the relative frequencies of biases in the estimated numbers of clusters (mPAM-10 (left bar) and PAM-10 (right bar)) for 4D $l_1$-MDS

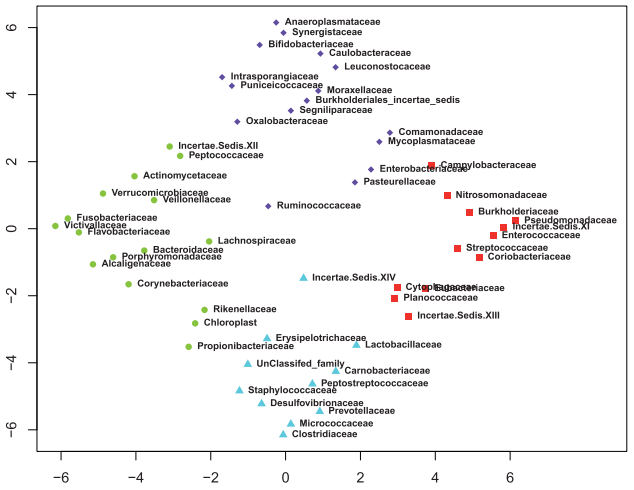


**Fig. 5.** Two dimensional $l_1$-MDS models displaying the four estimated clusters for taxa using the gut samples. Name of the taxa corresponding each sample point are provided as labels

### 4.1 Analysis of Costello data

The data from Costello *et al.* (2009) contain counts over various body sites of the 16s rRNA, and thus Unifrac with 2D PCoA were used in the original analysis to study the bacteria composition of body habitats given the existance of the phylogenetic tree. Our goal of this reanalysis is to demonstrate the importance of higher-dimensional analysis, and to compare PCoA with $l_1$-MDS. The 4D $l_1$-MDS analysis reveals an interesting and significant observation that was not apparent in the lower-dimensional analysis or using PCoA. That is, the oral cavity samples that were directly collected from the mouth and tongue were clustered together while those that are in fact transplant samples made up a different cluster (Supplementary Materials S5 and Table S3).

In the following, we also illustrate the ability of GrammR for clustering taxa with OTUs using the 45 gut samples. Of the 312 taxa at the bacteria family level, 257 did not have any recorded sequences classified to them, resulting in zero counts for all the samples for these taxa and were thus excluded from consideration. We analyzed the remaining 55 taxa by constructing $l_1$-MDS and PCoA models of dimensions 2, 3 and 4. The 2D $l_1$-MDS results showing four estimated clusters with 12, 16, 11, and 16 taxa are presented in Figure 5. In contrast, for 2D PCoA, only three taxa—Comamonadaceae, Moraxellaceae and Caulobacteraceae—are observed to belong to a separate cluster, with the rest make up another cluster (Supplementary Fig. S24). Apart from having fewer number of clusters and a lopsided membership, another notable difference is the multidimensional model axes. We can see that clusters in the PCoA model have much smaller spread compared to those in the $l_1$-MDS model.

### 4.2 Analysis of the Yatsunenko data

The Yatsunenko *et al.* (2012) data were obtained from the fecal samples of 252 adults and 176 infants living in three countries—USA, Venezuela and Malawi. The differences in bacterial composition among samples from the different countries is of interest. This is used as another example to demonstrate the benefits of higher dimensional analysis using GrammR using OTU information. After preprocessing and filtering (see Supplementary Materials S6 for detail), we were left with 4899 OTUs. Our analysis of the data from the 252 adults show that the 4D models (but not the 2D or 3D models) are capable of separating the samples from the different countries. For example, for $l_1$-MDS, the 2D models lead to the identification of three clusters, but the samples from Malawi and Venezuela are being grouped together, whereas the samples from USA are split into two clusters. In contrast, when 4D models are constructed, one can see a clear separation of samples from the three countries. This is true for both the $l_1$-MDS model (Supplementary Fig. S25) and the PCoA model (Supplementary Fig. S26), although the clusters are much tighter for $l_1$-MDS than PCoA, a phenomenon that we have seen in the other examples.

## 5 Discussion

In this article, we present a novel tool for graphical representation and clustering of metagenomic count data. In particular, we propose mPAM, a modified silhouette width measure for determining the number of clusters in conjunction with the PAM clustering method. In addition, we investigated and compared two methods for dimension reduction with different number of dimensionality in the projected space. Results from simulation clearly indicate the need to explore projections to a dimension greater than two. We further demonstrate the use of the novel tool for two datasets that would benefit from being considered from a higher-dimensional perspective.

While the original distance matrix (without being projected into a lower dimensional space) can be used directly for clustering, mPAM clustering based on the projected data is in fact advantageous due to better separation obtained from minimizing the stress function in $l_1$-MDS, especially if the projected dimension is relatively large (say 4D). We illustrate this using data generated from model II (Supplementary Materials S7 and Fig. S27).

Kendall's τ-distance is proposed to be used as a measure of dissimilarity for metagenomic count data, which is a useful tool for multidimensional modeling when the phylogenetic tree is not available. However, calculation of the τ-distance is computationally intensive, with the computation time being $O(n^2k^2)$, where $n$ is the number of samples and $k$ is the number of OTUs. A study on the computational time taken to compute the Kendall's τ-distance using GrammR for various sample sizes and OTUs is provided in Supplementary Materials S8 and Table S4. For example, for a dataset with 100 samples and 2000 OTUs, it took a little bit over 2 min to complete the computation. We note that there are several potential methods (Bansal and Fernández-Baca, 2009; Dietz, 1989) that can lead to speed up of the computation, which will be considered in a future release of GrammR.

A general class of metric MDS models is proposed for dimension reduction, enabling informative graphical visualization of data.

The popular PCoA model with two-dimensional projection is in fact a special case of this class. Comparison of PCoA and another metric MDS model, $l_1$-MDS, for two simulation models appears to show that PCoA frequently underestimate the true number of clusters, although the biases tend to get smaller with a higher-dimensional projection. Regardless of which metric MDS model is used, the benefit of using a dimension greater than two is clearly demonstrated in both the simulation and the real data applications. While two-dimensional models are easier to present on paper and thus popular, we would like to emphasize the benefit of higher dimensional representations, and that our method and accompanying software provide an interactive visualization tool for examining results for dimensions up to four. As seen in simulation studies, higher dimensional models reduce error of misclassification. Also, the optimal number of clusters used in these models gets closer to the true number of clusters with an increase in dimension. While a single graphical representation of a dimension greater than three is not feasible, one can construct representations selecting three directions at a time to get a series of rotatable 3D plots for closer examination. This, however, is not recommendable for dimensions much greater than 4, as the number of such representations increases at a polynomial rate.

## Funding

This work was partially supported by the United States National Science Foundation grant DMS-1220772.

*Conflict of Interest:* none declared.

## References

Bansal,M.S. and Fernández-Baca,D. (2009) Computing distances between partial rankings. *Inf. Process. Lett.*, **109**, 238–241.

Costello,E.K. *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

Dietz,P.F. (1989) Optimal algorithms for List Indexing and Subset Rank. *Algorithms Data Struct. Lect. Notes Comput. Sci.*, **382**, 39–46.

Fagin,R. *et al.* (2004) Comparing partial rankings. *SIAM J. Discret. Math.*, **20**, 47–58.

Holmes,I. *et al.* (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, **7**, e30126.

Huang,Z. (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl. Discov.*, **2**, 283–304.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Class.*, **2**, 193–218.

Kaufman,L. and Rousseeuw,P. (1987) Clustering by means of medoids. *Stat. Data Anal. Based $l_1$-Norm Relat. Methods*, **20**, 53–65.

Kendall,M.G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Ley,R.E. *et al.* (2008) Evolution of mammals and their gut microbes. *Science*, **320**, 1647–1651.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Lozupone,C.A. and Knight,R. (2007) Global patterns in bacterial diversity. *Proc. Natl Acad. Sci. USA*, **104**, 11436–11440.

Ridaura,V.K. *et al.* (2013) Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, **341**, 1079–U49.

Romero,R. *et al.* (2014) The composition and stability of the vaginal microbiota of normal and pregnant women is different from that of non-pregnant women. *Microbiome*, **2**, 4.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Sims,G.E. *et al.* (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl Acad. Sci. USA*, **106**, 2677–2682.

Subramanian,S. *et al.* (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*, **509**, 417–421.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Yatsunenko,T. *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.