



The Generalized Matrix Decomposition Biplot and Its Application to Microbiome Data

Yue Wang,^{a,b} Timothy W. Randolph,^{a,c} Ali Shojaie,^b Jing Ma^{a,d}

^aPublic Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

^bDepartment of Biostatistics, University of Washington, Seattle, Washington, USA

^cClinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

^dDepartment of Statistics, Texas A&M University, College Station, Texas, USA

ABSTRACT Exploratory analysis of human microbiome data is often based on dimension-reduced graphical displays derived from similarities based on non-Euclidean distances, such as UniFrac or Bray-Curtis. However, a display of this type, often referred to as the principal-coordinate analysis (PCoA) plot, does not reveal which taxa are related to the observed clustering because the configuration of samples is not based on a coordinate system in which both the samples and variables can be represented. The reason is that the PCoA plot is based on the eigen-decomposition of a similarity matrix and not the singular value decomposition (SVD) of the sample-by-abundance matrix. We propose a novel biplot that is based on an extension of the SVD, called the generalized matrix decomposition biplot (GMD-biplot), which involves an arbitrary matrix of similarities and the original matrix of variable measures, such as taxon abundances. As in a traditional biplot, points represent the samples, and arrows represent the variables. The proposed GMD-biplot is illustrated by analyzing multiple real and simulated data sets which demonstrate that the GMD-biplot provides improved clustering capability and a more meaningful relationship between the arrows and points.

IMPORTANCE Biplots that simultaneously display the sample clustering and the important taxa have gained popularity in the exploratory analysis of human microbiome data. Traditional biplots, assuming Euclidean distances between samples, are not appropriate for microbiome data, when non-Euclidean distances are used to characterize dissimilarities among microbial communities. Thus, incorporating information from non-Euclidean distances into a biplot becomes useful for graphical displays of microbiome data. The proposed GMD-biplot accounts for any arbitrary non-Euclidean distances and provides a robust and computationally efficient approach for graphical visualization of microbiome data. In addition, the proposed GMD-biplot displays both the samples and taxa with respect to the same coordinate system, which further allows the configuration of future samples.

KEYWORDS data visualization, clustering, dimension reduction, structured data, non-Euclidean distances


A biplot simultaneously displays, in two dimensions, rows (samples) of a data matrix as points and columns (variables) as arrows. Based on a matrix decomposition of the data matrix, the biplot is a useful graphical tool for visualizing the structure of large data matrices. It displays a dimension-reduced configuration of samples, as in a principal-coordinate analysis plot, and the variables with respect to the same set of coordinates. If meaningful sample groupings are observed, this allows visualizing which variables contribute most to the separation. The traditional biplot, as first introduced in reference 1, displays the first two left and right singular vectors of the singular value

Citation Wang Y, Randolph TW, Shojaie A, Ma J. 2019. The generalized matrix decomposition biplot and its application to microbiome data. *mSystems* 4:e00504-19. <https://doi.org/10.1128/mSystems.00504-19>.

Editor J. Gregory Caporaso, Northern Arizona University

Copyright © 2019 Wang et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Jing Ma, jingma@fredhutch.org.

 The GMD-biplot: a novel tool for constructing biplot when you have two-way structured data

Received 19 August 2019

Accepted 13 November 2019

Published 17 December 2019

decomposition (SVD) of the data matrix as points and arrows, respectively. This biplot, which we hereafter refer to as the SVD-biplot, uses the SVD to find the optimal least-square representation of the data matrix in a low-dimensional space. The SVD-biplot can show Euclidean distances between samples and also display approximate variances and correlations of the variables. It also has the appealing property that the singular values obtained from the SVD are nonincreasing, indicating that the decomposition of the total variance of the data matrix into each dimension is nonincreasing.

In many scenarios, the Euclidean distance may not be optimal for characterizing dissimilarities between samples. An important example arises in the analysis of microbiome data, in which marker gene sequences (e.g., 16S rRNA) are often grouped into taxonomic categories using bioinformatic pipelines such as QIIME (2) or mothur (3). These taxon counts can be summarized into a data matrix with rows and columns representing samples and taxon abundances, respectively. A variety of non-Euclidean distance measures, including nonlinear measures, are then used to quantify the similarity between samples. One common measure of dissimilarity is the UniFrac distance (weighted or unweighted), which is a function of the phylogenetic dissimilarity of a pair of samples (4, 5). Other nonphylogenetic, non-Euclidean dissimilarities include Jaccard or Bray-Curtis distances (see, e.g., reference 6 and the references therein). Plotting the samples in the space of the first few principal components (PCs) of the similarity matrix obtained from such non-Euclidean distance matrices—often referred to as principal-coordinate analysis (PCoA)—may reveal an informative separation between samples. However, the configuration of samples yielded by PCoA keeps only pairwise distances between samples and lacks a coordinate system that relates to the taxa that constitute each sample. Hence, it does not shed any light on which taxa may play a role in this separation. One approach for addressing this problem is to simply plot an arrow for each taxon based on its correlation with the first two PCs of the non-Euclidean similarity matrix (7). However, in such a “joint plot” (8), the direction and length of an arrow does not represent the taxon’s true contribution to the dissimilarity between samples. In addition, due to the lack of a coordinate system, one cannot add sample points for future observations into this “joint plot.”

Three main approaches have been recently proposed to extend the SVD-biplot to more general distances defined on the samples. The R package “ade4” (9) provides a biplot that can handle weighted Euclidean distances but it cannot handle non-Euclidean distances. The second approach, proposed by Greenacre (10), aims to approximate the non-Euclidean distance by a weighted Euclidean distance. Weights are estimated for variables, and the biplot can subsequently be constructed using weighted least-square approximation of the matrix. This approach has a straightforward interpretation. However, the estimated weighted Euclidean distance may not capture all the information from the original non-Euclidean distance. A recent proposal in reference 11 appears to be the first to address the lack of mathematical duality between the samples’ locations (points) and the variables’ contribution (arrows) to those locations. This approach seeks an approximate SVD-like decomposition of the data matrix, which directly takes the non-Euclidean distance into consideration. This SVD-like decomposition has the following two advantages. First, the left singular vectors are the eigenvectors of the similarity measure derived from the non-Euclidean distance, which preserve the role of the non-Euclidean distance in classifying the samples. Second, an approximate matrix duality (AMD) between the left and right singular vectors is restored, which simply means that each set of vectors can be immediately obtained from the other. To emphasize this connection, we hereafter refer to this decomposition as the AMD. Unfortunately, the AMD also suffers from two drawbacks. First, the AMD is only an approximate decomposition of the data matrix, and hence may not capture all the variation of the original data. In particular, the configuration of samples displayed in an AMD-biplot is independent of the data matrix, since the left singular vectors of the AMD depend only on the non-Euclidean distances. Ignoring the data matrix for classifying samples seems nonintuitive since the data matrix is typically assumed to contain some information on the sample similarities.

Second, the AMD may result in nondecreasing “singular values.” While these seem like minor technical issues, the second drawback can have important practical implications: which of the left and right singular vectors should be displayed in the resulting biplot? The authors of reference 11 suggest constructing the AMD-biplot based on the two left and right singular vectors that correspond to the two largest singular values. This AMD-biplot assures that the arrows for variables are as meaningful as possible, but they may fail to reveal meaningful sample clusters if the information of sample clusters is associated only with the first several left singular vectors. An alternative approach may be to simply display the first and second left and right singular vectors of the AMD (as done for the SVD). Unfortunately, this strategy does not solve the problem either: although we may observe meaningful sample clusters, the arrows may not be meaningful due to the small singular values. There is thus a lack of clarity regarding which singular vectors should be used to construct the AMD-biplot.

The drawbacks of the AMD-biplot motivate our proposal which is based on the generalized matrix decomposition biplot (GMD-biplot) (12). The GMD-biplot is a direct generalization of the SVD-biplot that accounts for structural dependencies among the samples and/or variables. This approach has several advantages. First, as with the AMD, it directly handles any non-Euclidean distance matrix. Specifically, the full information from that distance matrix is used. Second, unlike the AMD, which provides an approximate decomposition of the data matrix, the GMD provides an exact decomposition of the original data matrix without losing any information. Third, the GMD restores the matrix duality in a mathematically rigorous manner, unlike the approximate matrix duality obtained with the AMD; it naturally extends the duality inherent in the SVD and allows one to plot both the configuration of samples and the contribution of individual variables with respect to a new coordinate system. Fourth, the GMD gives nonincreasing GMD values, so the resulting GMD-biplot can be directly constructed based on the first two left and right GMD vectors. Last, unlike the AMD-biplot whose sample clusters depend only on distance, the GMD-biplot uses both the non-Euclidean distance and the data matrix for classifying samples, which more directly connects the contribution of the individual variables to the configuration of samples. Additionally, besides accounting for the non-Euclidean distances between samples, the GMD can also account for auxiliary information on (dis)similarities between the variables.

The remainder of this paper is organized as follows. We first illustrate the GMD-, AMD-, and SVD-biplots in three numerical studies. We then discuss advantages of the proposed GMD-biplot and further extensions. In Materials and Methods, we present detailed description of the GMD-biplot framework.

RESULTS

In the results below, we compare the GMD-, AMD-, and SVD-biplots on three data sets in the manner that each has been proposed recently for microbiome data. In particular, in reference 11, the AMD-biplot is advocated specifically for relative abundance data, while in reference 13, the SVD-biplot is advocated for data that have been scaled by the centered log ratio (CLR) transformation. The GMD-biplot is constructed using the CLR-transformed data. We first examine the performance of all biplots using the smokeless tobacco data set explored in reference 11. In the second study, we compare their performances using the human gut microbiome data from reference 14. In the third analysis, we simulate a data set based on the smokeless tobacco data to illustrate a dilemma that the AMD-biplot may face.

Analysis of the smokeless tobacco data. This data set includes 15 smokeless tobacco products: 6 dry snuff samples, 7 moist snuff samples, and 2 toombak samples from Sudan. Three separate (replicate) observations (starting with sample preparation) were made of each product, so that a total of 45 observations are available. Each observation has a 271×1 vector of taxon counts, and thus, the data set can be formed into a 45×271 matrix, denoted by X . The squared weighted UniFrac distance, denoted by $\Delta \in \mathbb{R}^{45 \times 45}$, was used to measure the distance between samples. The corresponding

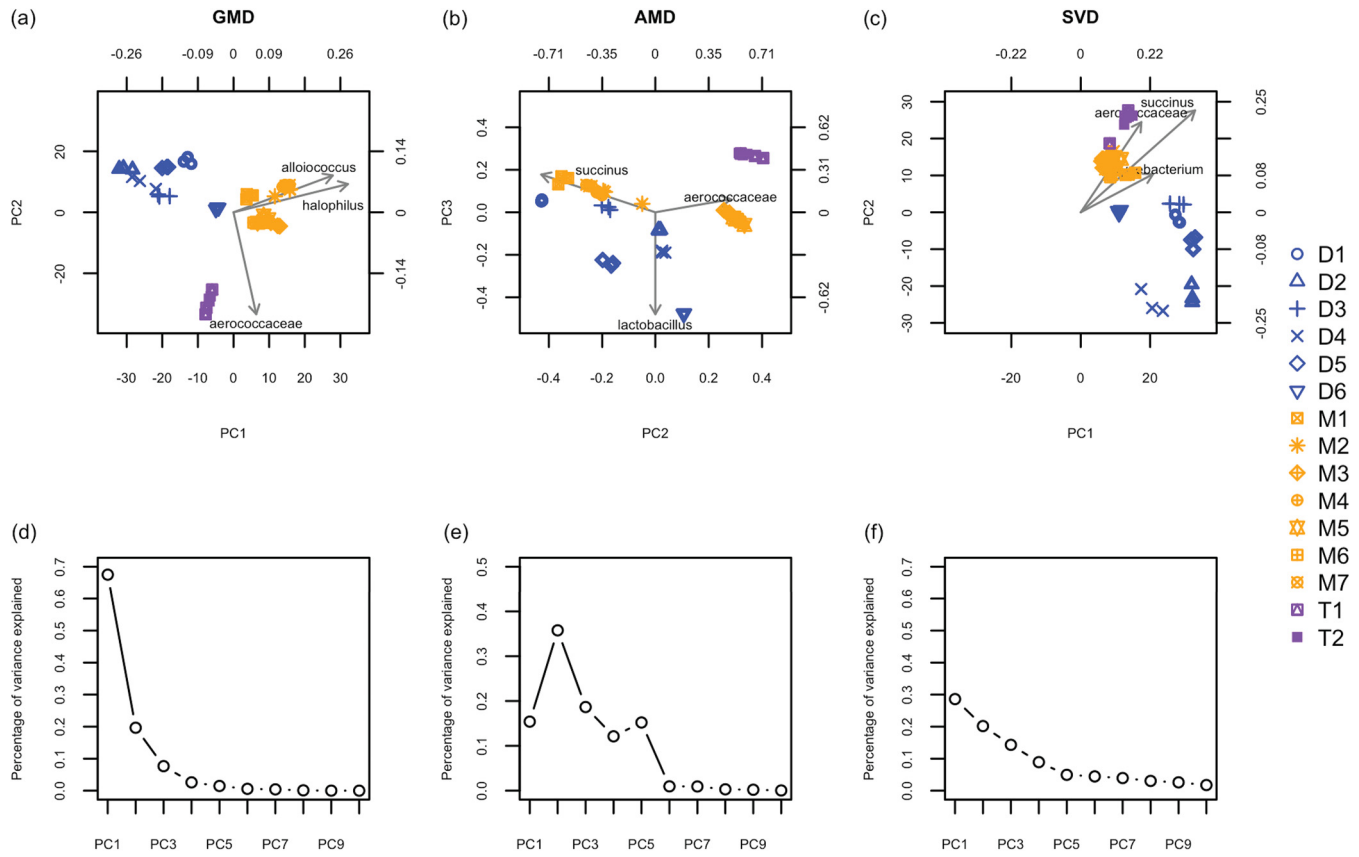


FIG 1 Biplots and scree plots for the analysis of smokeless tobacco data. (a) GMD-biplot based on the first and second components; (b) AMD-biplot based on the second and third components; (c) SVD-biplot based on the first and second components; (d) GMD scree plot; (e) AMD scree plot; (f) SVD scree plot. The biplots in panels a, b, and c display the top taxa with the longest arrows. The sample points are colored by sample type (dry snuff [blue], moist snuff [orange], and toombak [purple]), and samples corresponding to replicates of the same product are plotted with the same symbol (dry snuff [D], moist snuff [M], and toombak [T]). The scree plots in panels d, e, and f display the contributions of the top 10 components. (Panel e is adapted from reference 11.)

similarity kernel \mathbf{H} was calculated as $\mathbf{H} = -\frac{1}{2}\mathcal{F}\Delta\mathcal{F}$, where $\mathcal{F} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix and $\mathbf{1}_n$ is an $n \times 1$ vector. Since \mathbf{H} is not positively semidefinite, we forced it to be positive semidefinite by removing its negative eigenvalues and corresponding eigenvectors. The resulting similarity kernel, denoted \mathbf{H}^* , has rank 27.

For the GMD-biplot, we consider the CLR transformation of \mathbf{X} . Specifically, denoting the geometric mean of a vector \mathbf{z} by $g(\mathbf{z}) = (\prod_{k=1}^p z_k)^{1/p}$, the CLR transformation of \mathbf{x}_i ; $i = 1, \dots, 45$ is given by

$$\tilde{\mathbf{x}}_i = \left[\log \frac{x_{i1}}{g(\mathbf{x}_i)}, \dots, \log \frac{x_{ip}}{g(\mathbf{x}_i)} \right].$$

We denote the resulting data matrix by $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{45})^T$. For the AMD-biplot, we converted each row of \mathbf{X} into the empirical frequencies and further centered the rows and columns to have mean 0, as done in reference 11. We denote the resulting data matrix by $\check{\mathbf{X}}$.

We constructed the GMD-biplot and the AMD-biplot based on \mathbf{H}^* using $\tilde{\mathbf{X}}$ and $\check{\mathbf{X}}$, respectively. Figure 1d displays the proportion of variance captured by each GMD component. It can be seen that the first two GMD components capture more than 80% of the total variance of \mathbf{X} , which assures that the resulting GMD-biplot (Fig. 1a) visualizes the data well. As shown in Fig. 1a, the GMD-biplot is perfectly successful at separating the different tobacco products (dry, moist, and toombak). Furthermore, the replicates corresponding to the same product are tightly clustered. By examining the arrows for taxa in Fig. 1a, we see

that moist samples may be characterized by elevated levels of *Alloiococcus* and *Halophilus*, while *Aerococcaceae* appears elevated in toombak samples.

Figure 1e, which is the same as the right bottom panel of Fig. 1 in reference 11, shows that the AMD singular values are not necessarily decreasing. It should be noted that Fig. 1b is slightly different from Fig. 3 in reference 11; this difference may be due to the use of \mathbf{H}^* here as opposed to \mathbf{H} in reference 11. This is because we wanted the AMD-biplot to be directly comparable to the GMD-biplot, since the GMD requires both \mathbf{H} and \mathbf{R} to be positive semidefinite. From Fig. 1b, it can be seen that the AMD-biplot successfully separates toombak samples (purple points) from dry (blue) and moist (orange) snuff samples, although the separation between dry and moist snuff samples in the AMD-biplot is not as definitive as that in the GMD-biplot (Fig. 1a).

Additionally, we included the SVD-biplot and its corresponding scree plot in Fig. 1c and f, respectively. As the SVD-biplot assumes Euclidean distances between samples, it is more appropriate to construct the SVD-biplot using the CLR-transformed data $\tilde{\mathbf{X}}$ than the relative abundance data $\check{\mathbf{X}}$ (13). It can be seen from Fig. 1c that although the SVD-biplot successfully separates dry snuff from moist and toombak samples, it does not give a clear separation between moist snuff and toombak samples.

It is worth noting that the three biplots identify different top taxa, i.e., the taxa with the longest arrows. Although a biplot is not a rigorous statistical method to detect important taxa, it may shed light on which taxa are important to the observed sample clustering. To see this, we performed a univariate linear regression of each taxon (each column of $\tilde{\mathbf{X}}$) on the tobacco groups (dry, moist, and toombak) and obtained P values representing the strength of association between each taxon and the tobacco groups. We then sorted these P values in a nondecreasing order and obtained the rank of each taxon based on the sorted P values. Hence, it is desirable that the taxa with the lowest ranks can be identified by the biplots. Table S1 in the supplemental material summarizes the ranks of the top 10 taxa identified by each biplot. It can be seen that the top 10 taxa identified by the GMD-biplot have lower ranks on average than those identified by the AMD and SVD biplots, indicating that the GMD-biplot may identify more meaningful taxa with respect to the separation of the samples than the AMD and SVD biplots.

Analysis of human gut microbiome data. We consider the human gut microbiome data collected in a study of healthy children and adults from the Amazonas of Venezuela, rural Malawi, and U.S. metropolitan areas (14). The original data set \mathbf{X} consists of counts for 149 taxa for 100 samples. The squared unweighted UniFrac distance matrix $\Delta \in \mathbb{R}^{100 \times 100}$, computed using the R package phyloseq (15), was used to measure the distance between samples. Here, the distance between two samples is based entirely on the number of branches they share on a phylogenetic tree. The distance hence accounts only for the presence/absence of each taxon (not its abundance). The corresponding similarity kernel \mathbf{H} was then derived as $\mathbf{H} = -\frac{1}{2}\mathcal{J}\Delta\mathcal{J}$, which is a positive semidefinite matrix with rank 99. Let $\tilde{\mathbf{X}}$ and $\check{\mathbf{X}}$, respectively, denote the CLR-transformed data and the relative abundance data. Similar to the first study, the GMD-biplot and the AMD-biplot were constructed based on the similarity kernel \mathbf{H} using $\tilde{\mathbf{X}}$ and \mathbf{X} respectively, and the SVD-biplot was constructed based on the SVD of $\tilde{\mathbf{X}}$.

As concluded in reference 14, shared features of the functional maturation of the gut microbiome are identified during the first 3 years of life. We thus define a binary outcome h_i based on the age of the individual (in years) when each sample was taken as:

$$h_i = \begin{cases} 0 & \text{age}_i < 3 \\ 1 & \text{age}_i \geq 3, \end{cases}$$

for $i = 1, \dots, 100$. Approximately 70% of the samples are assigned to group 0, and the remaining 30% are assigned to group 1.

In all biplots, the i th sample is colored by age, and symbolized by h_i . Figure 2d indicates that more than 80% of the total variance is explained by the GMD-biplot in

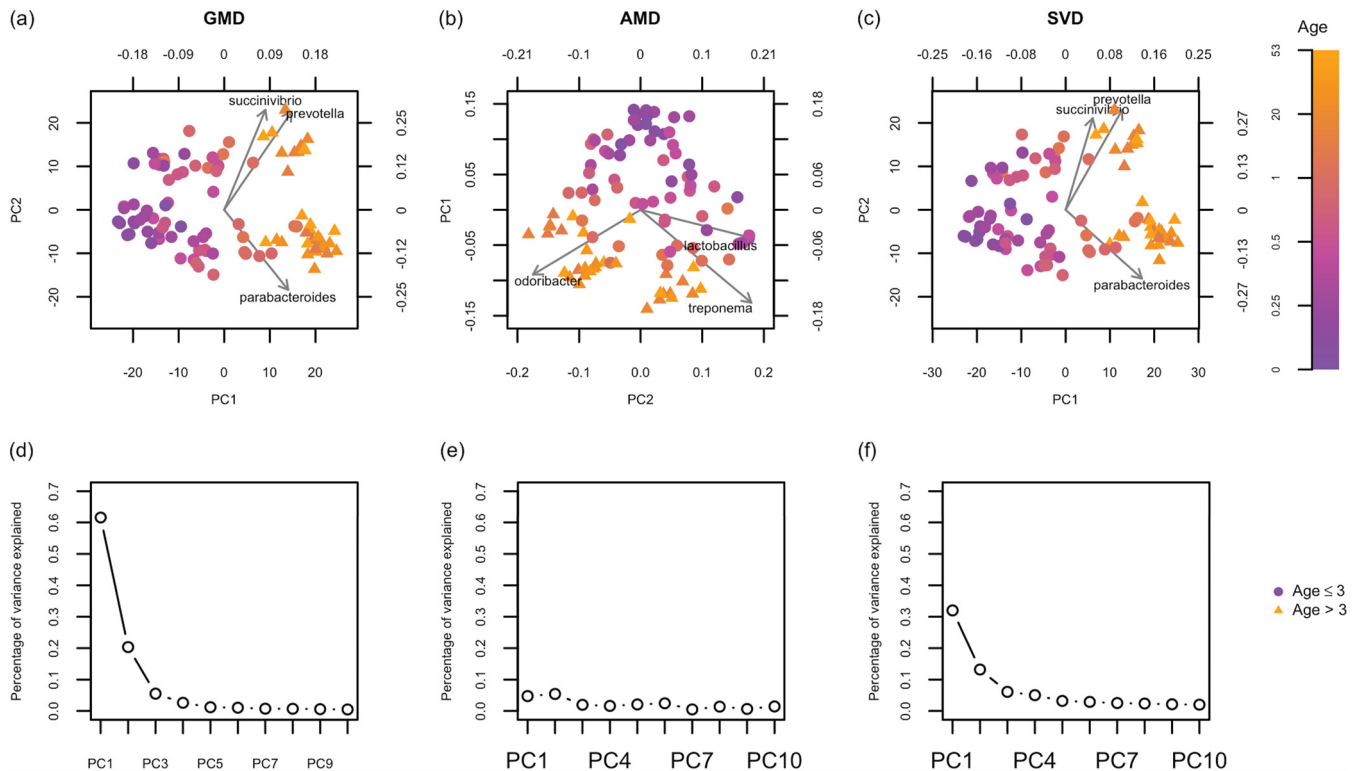


FIG 2 Biplots and scree plots for the analysis of the human gut microbiome data. (a) GMD-biplot based on the first and second GMD components; (b) AMD-biplot based on the first and second components; (c) SVD-biplot based on the first and second SVD components; (d) GMD scree plot; (e) AMD scree plot; (f) SVD scree plot. The biplots in panels a, b, and c display the top three taxa with the longest arrows. Symbols of sample points are based on the ages of individuals when the samples were collected (age \leq 3 years indicated by circles, and age $>$ 3 years indicated by triangles). The scree plots in panels d, e, and f display the contributions of the top 10 components.

Fig. 2a, which provides a good visualization of sample clusters across age. By examining the relationship between the arrows and the color of the sample points in Fig. 2a, we see that *Prevotella* may be elevated in adults, while *Parabacteroides* appears to be elevated in infants. In contrast, Fig. 2e shows that less than 15% of the total variance is explained by the AMD-biplot in Fig. 2b and the AMD values are not decreasing. As shown in Fig. 2b, the AMD-biplot also displays potential clusters across age, but the sample points are not as tightly clustered as those in Fig. 2a. *Odoribacter* appears to be elevated in adults in Fig. 2b, while *Lactobacillus* appears associated with infants. As a reference, Fig. 2c shows the SVD-biplot of \bar{X} , which looks very similar to Fig. 2a.

To further quantify the classification accuracy, for each biplot, we predicted the probability that each sample belongs to group 1 based on leave-one-out cross validation using the binary logistic regression of the group index h_i on the two selected components. We then plotted a receiver operating characteristic (ROC) curve for each biplot based on the predicted probabilities (see Fig. S1 in the supplemental material) and calculated the area under the ROC curve (AUC): the GMD-, AMD-, and SVD-biplots yield an AUC of 0.989, 0.976, and 0.990, respectively. The AUC results indicate that the GMD-biplot provides a better separation of age groups than the AMD-biplot, but there is not a clear difference between the GMD-biplot and the SVD-biplot. This may be because, for the CLR-transformed data \bar{X} , the unweighted UniFrac distance is not as informative with respect to age as the weighted UniFrac distance was in the tobacco data with respect to product groups.

We emphasize that both the GMD-biplot and the SVD-biplot identify *Prevotella* and *Parabacteroides* as top taxa, while the AMD-biplot identifies completely different ones. As reference 14 confirms that the trade-off between *Prevotella* and *Bacteroides* (including *Parabacteroides*) considerably drives the variation of microbiome abundance in

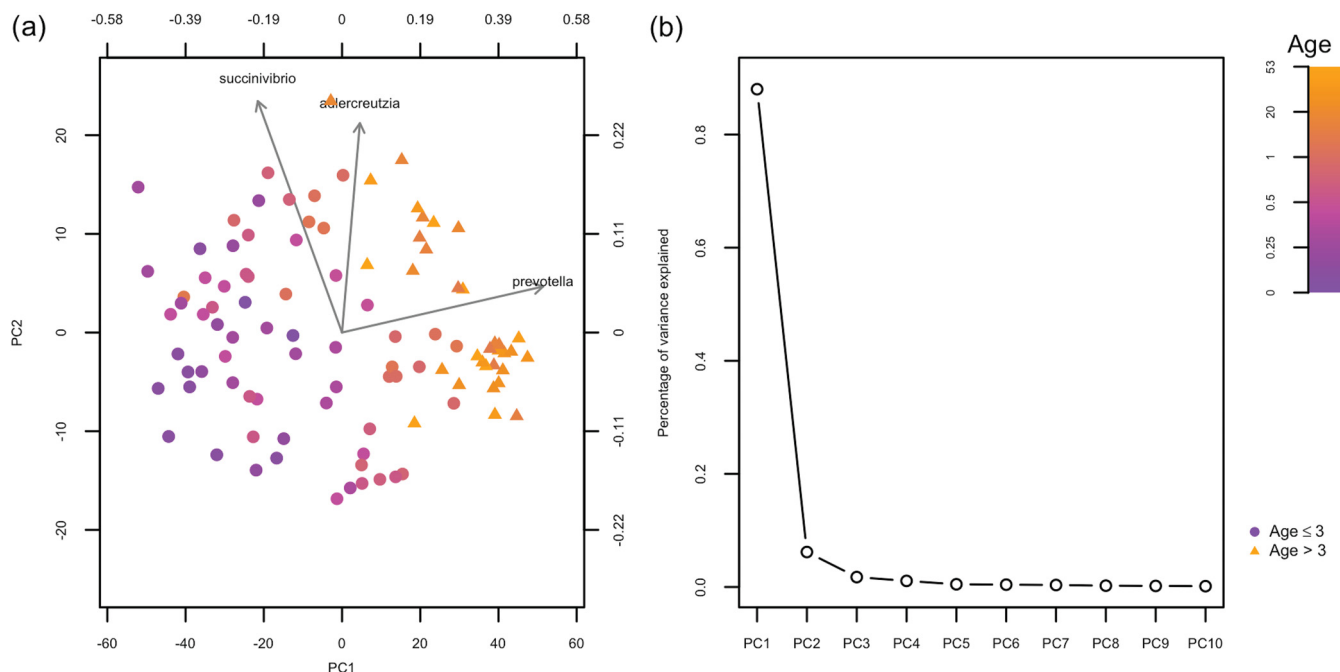


FIG 3 The biplot and scree plot for the analysis of the human gut microbiome data using both **H** and **R**. (a) GMD-biplot using both **H** and **R** based on the first and second GMD components. The top three taxa with the longest arrows are displayed. Symbols of sample points are based on the ages of the individuals when the samples were collected (age ≤ 3 years indicated by circles, and age > 3 years indicated by triangles). (b) GMD scree plot using both **H** and **R**. The contributions of top 10 components are displayed.

adults and children between 0.6 and 1 year of age in all studied populations, the GMD- and SVD-biplots may thus identify more biologically meaningful taxa than the AMD-biplot. It should, however, be noted that these bacterial are “identified” based on circumstantial, not statistical, evidence, and more work is needed to examine statistical associations in this context.

Incorporating a kernel for variables into the GMD-biplot. The GMD problem (see equation 3 in Materials and Methods) allows not only the similarity kernel for samples but also a kernel for the variables. Including both kernels may further improve the accuracy of sample classification as well as the identification of important variables. We illustrate this advantage by incorporating a kernel for variables in the analysis of the human gut microbiome data. More specifically, we first calculate a matrix $\Delta_R \in \mathbb{R}^{149 \times 149}$ of squared patristic distances between the tips of the phylogenetic tree for each pair of taxa and then derive a similarity matrix **R** as $\mathbf{R} = -\frac{1}{2} \mathcal{J} \Delta_R \mathcal{J}$. Figure 3a shows the GMD-biplot with the additional kernel **R** incorporated. The ROC analysis based on the leave-one-out cross validation for Fig. 3a yields an AUC of 0.984, which is higher than that of the AMD-biplot (Fig. 2b) but slightly lower than Fig. 2a and Fig. 2c. This may be because both **H** and **R** highly depend on the phylogenetic tree. Thus, incorporating **R** may be redundant and may reduce the accuracy of the sample clustering in this case. The top three taxa identified in Fig. 3a include *Prevotella* but not *Parabacteroides*, which may explain the lower clustering accuracy.

Including an additional kernel for variables in the GMD-biplot is related to the method of double-principal-coordinate analysis (DPCoA) (16). DPCoA, as shown in reference 17, is equivalent to a generalized PCoA which essentially incorporates an additional similarity kernel for variables into the analysis, as described in Proposition 1, but for $\mathbf{H} = \mathbf{I}_n$. As suggested in reference 18, DPCoA can provide a biplot representation of both samples and meaningful taxonomic categories. Hence, the GMD-biplot can also be viewed as an extension of DPCoA biplots because the GMD allows kernels for both samples and variables, while DPCoA allows a kernel only for variables.

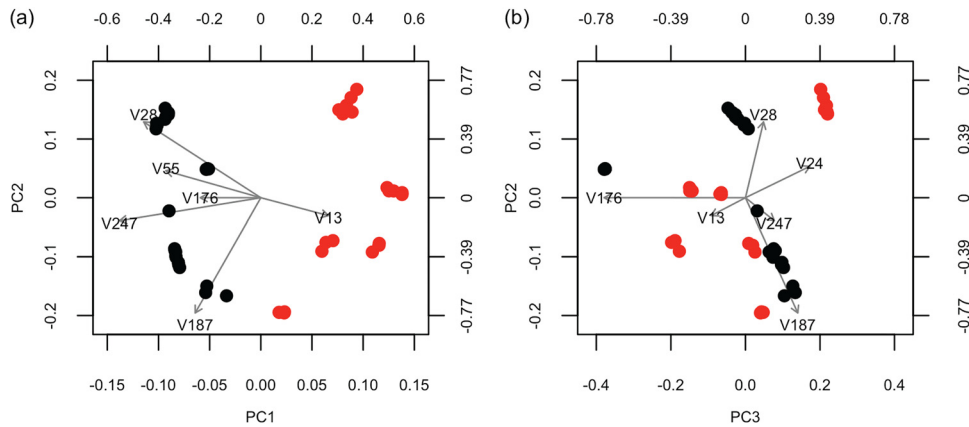


FIG 4 Biplots for the analysis of the simulated data. (a) GMD-biplot based on the first and second GMD components; (b) AMD-biplot based on the second and third components. Both biplots display the top six taxa with the longest arrows. The sample points are colored by the group index (1 [red] or 0 [black]).

Simulation. In this section, we conduct a simulation study based on the smokeless tobacco data to illustrate a scenario in which the AMD-biplot may fail to separate the samples, whereas the GMD-biplot performs well. Let \mathbf{H}^* and \mathbf{X} be the similarity kernel and data matrix from the smokeless tobacco data, respectively. We consider the eigen-decomposition of \mathbf{H}^* as $\mathbf{H}^* = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^T$; \mathbf{B} is a 45×27 matrix whose columns are eigenvectors of \mathbf{H}^* and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_{27})$ is a diagonal matrix whose elements are the eigenvalue of \mathbf{H}^* . Then, the AMD-biplot is based on the following approximated orthogonal decomposition of \mathbf{X} :

$$\mathbf{X} \approx \mathbf{B}\mathbf{D}\mathbf{V}^T, \tag{1}$$

where $\mathbf{D} = \text{diag}(d_1, \dots, d_{27})$ and \mathbf{V} is a 271×27 matrix with orthonormal columns. As shown in Fig. 2d, d_1, \dots, d_{27} may not be decreasing. For $j = 1, \dots, 27$, we define

$$d_{j,S} = \begin{cases} 0.6 & j = 1 \\ 0.8 & j = 2 \\ 1 & j = 3 \\ 0 & j > 3, \end{cases}$$

and construct the simulated data set \mathbf{X}_S as $\mathbf{X}_S = \mathbf{B}\mathbf{D}_S\mathbf{V}^T$, where $\mathbf{D}_S = \text{diag}(d_{1,S}, \dots, d_{27,S})$. For $i = 1, \dots, 45$, we define a binary outcome w_i that indicates the group index of the i th sample as:

$$w_i = \begin{cases} 1 & b_{i1} > 0 \\ 0 & b_{i1} \leq 0. \end{cases}$$

The GMD-biplot and the AMD-biplot of \mathbf{X}_S with similarity measure \mathbf{H}^* are presented in Fig. 4a and b, respectively. It can be seen that the two groups are completely mixed up in the AMD-biplot because the first column of \mathbf{B} is not selected for visualization. In contrast, the GMD-biplot successfully visualizes the sample groups by displaying the first and second GMD components.

To see why this occurs, we summarize the first three diagonal elements of $\mathbf{\Lambda}$, \mathbf{D}_S and $\mathbf{D}_S^2\mathbf{\Lambda}$ in Table 1 and notice that $d_{1,S} < d_{2,S} < d_{3,S}$. Consequently, the AMD-biplot displays

TABLE 1 The first three diagonal elements of $\mathbf{\Lambda}$, \mathbf{D}_S , and $\mathbf{D}_S^2\mathbf{\Lambda}$ in the simulation

Matrix	Value for the following diagonal elements:		
	1st	2nd	3rd
\mathbf{D}_S	0.6	0.8	1
$\mathbf{\Lambda}$	3.09	1.26	0.77
$\mathbf{D}_S^2\mathbf{\Lambda}$	1.11	0.81	0.77

the second and third columns of \mathbf{BD}_S , and hence, it completely fails to classify the samples because the group index w_i depends only on the first column of \mathbf{B} . In contrast, Proposition 1a (see Materials and Methods) shows that the GMD-biplot is based on the two largest eigenvalues and the corresponding eigenvectors of $\mathbf{X}_S\mathbf{X}_S^T\mathbf{H}^*$. It can be further seen that

$$\mathbf{X}_S\mathbf{X}_S^T\mathbf{H}^* = \mathbf{BD}_S\mathbf{V}^T\mathbf{VD}_S\mathbf{B}^T\mathbf{B}\mathbf{A}\mathbf{B}^T = \mathbf{BD}_S^2\mathbf{A}\mathbf{B}^T. \quad (2)$$

Equation 2 implies that the diagonal elements of $\mathbf{D}_S^2\mathbf{A}$ are the eigenvalues of $\mathbf{X}_S\mathbf{X}_S^T\mathbf{H}^*$ and columns of \mathbf{B} are the corresponding eigenvectors. Hence, it can be seen from Table 1 that $d_{1,S}^2\lambda_1 > d_{2,S}^2\lambda_2 > d_{3,S}^2\lambda_3$, even though $d_{1,S} < d_{2,S} < d_{3,S}$. Therefore, the GMD-biplot displays the first and second column of $\mathbf{BD}_S\mathbf{A}^{1/2}$ as sample points, which successfully captures sample classifications.

DISCUSSION

Biplots have gained popularity in the exploratory analysis of high-dimensional microbiome data. The traditional SVD-biplot is based on Euclidean distances between samples and cannot be directly applied when more general dissimilarities are used. Since Euclidean distances may not lead to an optimal low-dimensional representation of the samples, we have extended the concept of the SVD-biplot to allow for more general similarity kernels. The phylogenetically informed UniFrac distance, used in our examples, defines one such kernel. In settings where a general (possibly nonlinear) distance matrix is appropriate, our approach provides a mathematically rigorous and computationally efficient method, based on the GMD, that allows for plotting both the samples and variables with respect to the same coordinate system.

Our first data example with the smokeless tobacco data set from reference 11 demonstrates the merits of the proposed GMD-biplot. We found that the GMD-biplot successfully displays different types of products, while the AMD-biplot is not able to completely separate dry and moist snuff samples, and the SVD-biplot fails to separate moist and toombak samples. As shown in Table S1 in the supplemental material, the GMD-biplot is also able to identify biologically more meaningful taxa that are related to the different types of products, compared to the AMD-biplot and the SVD-biplot.

In our second example, the GMD-biplot also outperforms the AMD-biplot in terms of both the sample clustering and the identification of important taxa. However, there is not a clear advantage of the GMD-biplot over the SVD-biplot in this example. This difference between the two examples may be attributed to the relation between the Euclidean kernel and the non-Euclidean similarity measure. Denoting the Euclidean kernel and the non-Euclidean similarity measure by \mathbf{XX}^T and \mathbf{H} , respectively, it can be seen that the sample configuration in the AMD-biplot and the SVD-biplot depend solely on either \mathbf{H} or \mathbf{XX}^T , whereas the GMD-biplot uses the top two eigenvectors of $\mathbf{XX}^T\mathbf{H}$, the matrix product of the Euclidean kernel \mathbf{XX}^T and \mathbf{H} . Hence, if \mathbf{XX}^T contains substantially more information about sample clustering than \mathbf{H} , then taking \mathbf{H}^T into consideration may not further improve the accuracy of sample clustering. Indeed, this may be the case in our second example, where the clustering of samples using the Euclidean distance between samples of the CLR-transformed data is highly successful because the difference of the microbial profiles between infants and adults is obvious even without the help of the UniFrac distance. However, a possibly more common scenario is when both \mathbf{H} and \mathbf{XX}^T contain some, but different, information on sample clustering. In such cases, taking both \mathbf{XX}^T and \mathbf{H} into consideration may improve the sample clustering and provide better biological interpretation.

In practice, we typically do not know what the true configuration of samples looks like, so it is impossible to determine whether \mathbf{H} or \mathbf{XX}^T contains more information about sample clusters. Also, it is sensible to assume that \mathbf{XX}^T and \mathbf{H} are “coinformative” in the sense that they exhibit a shared eigenstructure; for instance, both may be informative for clustering samples. The coinformativeness can be quantified precisely using the Hilbert-Schmidt information criteria (HSIC) (19). For any two kernels \mathbf{K}_1 and \mathbf{K}_2 , the empirical HSIC is proportional to $\text{tr}(\mathbf{K}_1\mathbf{K}_2)$. Hence, by definition, the GMD

problem in equation 3 (see Materials and Methods) is equivalent to minimizing the HSIC between $(\mathbf{X} - \mathbf{USV}^T)(\mathbf{X} - \mathbf{USV}^T)^T$ and \mathbf{H} over \mathbf{U} , \mathbf{S} and \mathbf{V} . In other words, if we consider $\mathbf{X} - \mathbf{USV}^T$ as the residual matrix of \mathbf{X} , then the GMD solutions can be interpreted as the best approximation to \mathbf{X} in the sense that the HSIC between \mathbf{H} and the Euclidean kernel of the residual matrix is minimized. Thus, the GMD-biplot considers the coinformativeness of \mathbf{XX}^T and \mathbf{H} . Therefore, in many cases, it would be a more robust way to display the sample points compared to the AMD-biplot or the SVD-biplot. Another advantage of the GMD-biplot over the AMD-biplot is illustrated in our simulation study. Since AMD may not give decreasing singular values, the AMD-biplot may not be able to display the most informative eigenvectors of \mathbf{H} , and may thus fail to cluster the samples. In contrast, GMD assures that the resulting singular values are nonincreasing.

Our discussion in this paper has focused on the form biplot, which aims to visualize the relationship between variables and the sample clustering. In other scenarios, where the variation of the data matrix explained by each variable is of particular interest, the covariance biplot may be more appropriate. This biplot considers the GMD of \mathbf{X} with respect to \mathbf{H} ; i.e., $\mathbf{X} = \mathbf{USV}^T$, where $\mathbf{U}^T\mathbf{H}\mathbf{U} = \mathbf{I}_q$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_q$. Note that

$$\|\mathbf{X}\|_{\mathbf{H},\mathbf{I}}^2 = \text{tr}(\mathbf{XX}^T\mathbf{H}) = \text{tr}(\mathbf{X}^T\mathbf{H}\mathbf{X}) = \text{tr}(\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}) = \sum_{m=1}^q s_m^2,$$

where $\mathbf{S} = \text{diag}(s_1, \dots, s_q)$. Furthermore, since \mathbf{V} has orthogonal columns, it can be seen that $\sum_{m=1}^q s_m^2 = \sum_{j=1}^p (\sum_{m=1}^q v_{jm}^2 s_m^2)$. Thus, the value of $(\sum_{m=1}^q v_{jm}^2 s_m^2) / \|\mathbf{X}\|_{\mathbf{H},\mathbf{I}}^2$ gives the proportion of the variability in $\|\mathbf{X}\|_{\mathbf{H},\mathbf{I}}^2$ explained by the j th variable. Note that when $q = 2$, the length of the arrow of the j th variable in the covariance biplot is given by $\sqrt{(\sum_{m=1}^q v_{jm}^2 s_m^2)}$. Therefore, in a covariance biplot, the arrows shed light on how the total variance of the data is partitioned into parts explained by each variable.

MATERIALS AND METHODS

We denote the data matrix by $\mathbf{X} \in \mathbb{R}^{n \times p}$, where n is the number of samples and p is the number of variables (taxa). We assume that the columns of \mathbf{X} are centered to have mean 0 and $\text{rank}(\mathbf{X}) = K \leq \min(n, p)$. For any matrix \mathbf{M} , we denote its i th row (sample) and its (i, j) entry by \mathbf{m}_i and m_{ij} , respectively. We denote the transpose of \mathbf{M} by \mathbf{M}^T .

Biplot, distance measure, and AMD. A biplot is a graphical method to simultaneously represent, in two dimensions, both the rows (as points) and columns (as arrows) of the matrix \mathbf{X} on the same coordinate axes. Given a decomposition of \mathbf{X} as $\mathbf{X} = \mathbf{AB}^T$, a biplot displays two selected columns of \mathbf{A} and \mathbf{B} . The SVD-biplot is based on the singular value decomposition (SVD) of \mathbf{X} , i.e., $\mathbf{X} = \mathbf{USV}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_K$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_K$, and $\mathbf{S} = \text{diag}(\sigma_1, \dots, \sigma_K)$ with $\sigma_1, \dots, \sigma_K$ being a sequence of nonincreasing and positive singular values. Here \mathbf{I}_K is a rank K identity matrix. Based on the SVD, \mathbf{A} and \mathbf{B} can take various forms; examples include form and covariance biplots (7). Since our primary interest is to visualize the clustering of samples, we focus on the form biplot in this paper and comment on the covariance biplot in the Discussion.

The SVD-biplot displays the first two columns of \mathbf{US} and \mathbf{V} , which can explain $(\sigma_1^2 + \sigma_2^2) \text{tr}(\mathbf{XX}^T)$ of the total variance of \mathbf{X} . The SVD of \mathbf{X} is closely related to the eigen-decomposition of the similarity kernel \mathbf{XX}^T , as we can write $\mathbf{XX}^T = \mathbf{US}^2\mathbf{U}^T$. Thus, the eigen-decomposition of \mathbf{XX}^T provides a way to calculate \mathbf{U} and \mathbf{S} . Once \mathbf{U} and \mathbf{S} are calculated, one can calculate \mathbf{V} from the duality between \mathbf{U} and \mathbf{V} ; that is, $\mathbf{VS} = \mathbf{X}^T\mathbf{U}$. The similarity kernel \mathbf{XX}^T characterizes the Euclidean distance between samples. To see this, we define the Euclidean squared distance between the i th and j th sample as $d_{ij}^E = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbb{R}^p}^2$. Let $\mathcal{F} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ be the centering matrix where $\mathbf{1}_n$ is an $n \times 1$ vector of ones. It can then be seen that $\mathbf{XX}^T = -\frac{1}{2} \mathcal{F} \Delta_E \mathcal{F}$, where the (i, j) entry of Δ_E is d_{ij}^E .

Now, if Δ_E is replaced by a matrix \mathbf{D} of non-Euclidean squared dissimilarities, one can still define a similarity kernel by $\mathbf{H} = -\frac{1}{2} \mathcal{F} \mathbf{D} \mathcal{F}$. One such example is when \mathbf{D} arises from distances between sample vectors of microbial abundances (or presence/absence) which account for a phylogenetic tree, as in a weighted (respectively, unweighted) UniFrac distance matrix. In this case, one can construct a principal-coordinate analysis (PCoA) plot of the samples using $\mathbf{H} = \mathbf{U}_H \mathbf{S}_H^2 \mathbf{U}_H^T$. However, an SVD-biplot cannot be constructed, since there is no \mathbf{V} that corresponds to the variables. The approximate matrix duality (AMD) addresses this problem by fixing \mathbf{U}_H and then seeking a matrix \mathbf{V}_H with orthonormal columns and a diagonal matrix \mathbf{D}_H with nonnegative elements that minimize the objective function

$$(\mathbf{V}_H, \mathbf{D}_H) = \text{argmin}_{\mathbf{V}, \mathbf{D}} \|\mathbf{X} - \mathbf{U}_H \mathbf{D} \mathbf{V}^T\|_F^2.$$

Here, $\|\mathbf{M}\|_F^2 = \text{tr}(\mathbf{M}^T\mathbf{M})$ is the Frobenius norm of \mathbf{M} , and for any square matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, $\text{tr}(\mathbf{M}) = \sum_{j=1}^d m_{jj}$. The resulting AMD-biplot can be constructed by plotting the two columns of $\mathbf{U}_H \mathbf{D}_H$ as sample points and plotting \mathbf{V}_H as arrows for variables; the selected two columns/rows correspond to the two largest elements of \mathbf{D}_H .

GMD and the GMD-biplot. The concept of generalized matrix decomposition (GMD) was introduced by Escoufier (20) and further developed in reference 12. It is a generalization of the SVD with additional structural dependencies taken into consideration. We briefly review the key ideas behind the GMD. Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ and $\mathbf{R} \in \mathbb{R}^{p \times p}$ be two positive semidefinite matrices, which characterize the similarities between samples and between variables, respectively. The \mathbf{H}, \mathbf{R} -norm of \mathbf{X} is defined as $\|\mathbf{X}\|_{\mathbf{H}, \mathbf{R}} = \sqrt{\text{tr}(\mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{H})}$. For any $q \leq K$, the GMD solution $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}, \tilde{\mathbf{S}})$ finds the best rank- q approximation to \mathbf{X} with respect to the \mathbf{H}, \mathbf{R} -norm, that is,

$$(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}, \tilde{\mathbf{S}}) = \underset{\mathbf{U}, \mathbf{V}, \mathbf{S}}{\text{argmin}} \|\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T\|_{\mathbf{H}, \mathbf{R}}^2 \quad (3)$$

subject to $\mathbf{U}^T\mathbf{H}\mathbf{U} = \mathbf{I}_q$, $\mathbf{V}^T\mathbf{R}\mathbf{V} = \mathbf{I}_q$ and $\text{diag}(\mathbf{S}) \geq 0$. Here, $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are the left and right GMD vectors, respectively, and $\tilde{\mathbf{S}}$ is a diagonal matrix containing the GMD values. Note that $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are orthogonal with respect to \mathbf{H} and \mathbf{R} , respectively, but they may not be orthogonal with respect to the Euclidean norm unless $\mathbf{H} = \mathbf{I}_n$ and $\mathbf{R} = \mathbf{I}_p$. The following property of the GMD provides a way to calculate the GMD components; the proof can be found in reference 20.

Proposition 1: The GMD solutions $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}, \tilde{\mathbf{S}})$ satisfy:

$$(a) \mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{H}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}^2$$

$$(b) \tilde{\mathbf{V}}\tilde{\mathbf{S}} = \mathbf{X}^T\mathbf{H}\tilde{\mathbf{U}}$$

Proposition 1a suggests that the diagonal elements of $\tilde{\mathbf{S}}$ and corresponding columns of $\tilde{\mathbf{U}}$ are eigenvalues and corresponding eigenvectors of $\mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{H}$, respectively. Proposition 1b establishes the duality between $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$, meaning that $\tilde{\mathbf{V}}$ can be immediately obtained, given $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{S}}$. Alternatively, an efficient algorithm for finding the solution to equation 3 was proposed in reference 12, which is less computationally intensive compared to finding the eigenvalues and eigenvectors of $\mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{H}$. The algorithm also ensures that the diagonal elements of $\tilde{\mathbf{S}}$ are nonincreasing.

Note that the GMD can handle the non-Euclidean similarity kernel \mathbf{H} just by taking $\mathbf{R} = \mathbf{I}_p$. Based on the GMD of \mathbf{X} with respect to \mathbf{H} , the GMD-biplot can be constructed with respect to the coordinate system provided by the first two columns of \mathbf{V} . More specifically, letting \mathbf{v}_j be the j th column of \mathbf{V} , the i th sample point can be configured by the coordinates of \mathbf{x}_i , given by $(\mathbf{x}_i^T\mathbf{v}_1, \mathbf{x}_i^T\mathbf{v}_2)$. To plot the arrow for the j th variable, we consider the vector $\mathbf{e}_j \in \mathbb{R}^p$, which has a 1 in the j th element and 0's elsewhere. Then, the arrow for the j th variable can be configured by the coordinates of \mathbf{e}_j , given by $(\mathbf{e}_j^T\mathbf{v}_1, \mathbf{e}_j^T\mathbf{v}_2)$. This coordinate system also allows the configuration of future samples. Letting $\mathbf{x}_* \in \mathbb{R}^p$ be a future sample, we can add \mathbf{x}_* into the GMD-biplot as a point located at $(\mathbf{x}_*^T\mathbf{v}_1, \mathbf{x}_*^T\mathbf{v}_2)$. Similar to the SVD-biplot, the GMD-biplot can explain $(\bar{\sigma}_1^2 + \bar{\sigma}_2^2)/\text{tr}(\mathbf{X}\mathbf{R}\mathbf{X}^T\mathbf{H})$ of the total variance of \mathbf{X} with respect to the \mathbf{H}, \mathbf{I}_p norm, where $\bar{\sigma}_k$ is the k th diagonal element of $\tilde{\mathbf{S}}$ for $k = 1, 2$.

Since the GMD values are nonincreasing, for the purpose of constructing the GMD-biplot, we can choose $q = 2$ in the GMD problem (equation 3), which may save considerable computational time. In contrast, since the AMD may produce nondecreasing "singular values," we have to find the full decomposition of \mathbf{X} by the AMD before deciding which singular vectors to plot in the AMD-biplot; this may become computationally intensive for large n and p .

Data availability. All data used are publicly available in references 11 and 14. All computations are conducted in the R programming language, and the proposed biplot is implemented in our R package "GMDcomp," available at <https://github.com/taryue/GMDcomp>.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00504-19>.

FIG S1, TIF file, 1.2 MB.

TABLE S1, PDF file, 0.02 MB.

ACKNOWLEDGMENTS

This work was partially supported by grant R01 GM114029 from the NIH. A.S. also acknowledges support from the NSF through grant DMS-1561814. J.M. and T.W.R. acknowledge support by grant R01 GM129512 from the NIH. T.W.R. also acknowledges support from NIH grants R01 CA192222 and P01 CA168530.

The content of this article is solely our responsibility and does not necessarily represent the official views of the NIH or any other funding agency.

We thank Parker Knight for his assistance with the software development.

REFERENCES

- Gabriel KR. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58:453–467. <https://doi.org/10.1093/biomet/58.3.453>.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of

- high-throughput community sequencing data. *Nat Methods* 7:335–336. <https://doi.org/10.1038/nmeth.f.303>.
3. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541. <https://doi.org/10.1128/AEM.01541-09>.
 4. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>.
 5. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73:1576–1585. <https://doi.org/10.1128/AEM.01996-06>.
 6. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, Huttenhower C, Ley RE. 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol* 9:e1002863. <https://doi.org/10.1371/journal.pcbi.1002863>.
 7. Daunis-i-Estadella J, Thió-Henestrosa S, Mateu-Figueras G. 2011. Including supplementary elements in a compositional biplot. *Comput Geosci* 37:696–701. <https://doi.org/10.1016/j.cageo.2010.11.003>.
 8. McCune B, Grace JB, Urban DL. 2002. Analysis of ecological communities, vol 28. MjM Software Design, Gleneden Beach, OR.
 9. Dray S, Dufour A. 2007. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 22:1–20.
 10. Greenacre M. 2017. Ordination with any dissimilarity measure: a weighted Euclidean solution. *Ecology* 98:2293–2300. <https://doi.org/10.1002/ecy.1937>.
 11. Satten GA, Tyx RE, Rivera AJ, Stanfill S. 2017. Restoring the duality between principal components of a distance matrix and linear combinations of predictors, with application to studies of the microbiome. *PLoS One* 12:e0168131. <https://doi.org/10.1371/journal.pone.0168131>.
 12. Allen GI, Grosenick L, Taylor J. 2014. A generalized least-square matrix decomposition. *J Am Stat Assoc* 109:145–159. <https://doi.org/10.1080/01621459.2013.852978>.
 13. Gloor GB, Reid G. 2016. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* 62:692–703. <https://doi.org/10.1139/cjm-2015-0821>.
 14. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. *Nature* 486:222–227. <https://doi.org/10.1038/nature11053>.
 15. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. <https://doi.org/10.1371/journal.pone.0061217>.
 16. Pavoine S, Dufour AB, Chessel D. 2004. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J Theor Biol* 228:523–537. <https://doi.org/10.1016/j.jtbi.2004.02.014>.
 17. Purdom E. 2011. Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann Appl Stat* 5:2326–2358. <https://doi.org/10.1214/10-AOAS402>.
 18. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes S. 2016. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Res* 5:1492. <https://doi.org/10.12688/f1000research.8986.2>.
 19. Gretton A, Bousquet O, Smola A, Schölkopf B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms, p 63–77. *In Proceedings of the 16th International Conference on Algorithmic Learning Theory ALT'05*. Springer-Verlag, Berlin, Germany.
 20. Escoufier Y. 2006. Operator related to a data matrix: a survey, p 285–297. *In Rizzi A, Vichi M (ed), Compstat 2006 - Proceedings in Computational Statistics*. Springer-Verlag, Berlin, Germany.