genes MDPI

Article

Principal Amalgamation Analysis for Microbiome Data

Yan Li [1], Gen Li [2] and Kun Chen [1,*]

[1] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA; yanlisph@umich.edu
[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; ligen@umich.edu
* Correspondence: kun.chen@uconn.edu

Abstract: In recent years microbiome studies have become increasingly prevalent and large-scale. Through high-throughput sequencing technologies and well-established analytical pipelines, relative abundance data of operational taxonomic units and their associated taxonomic structures are routinely produced. Since such data can be extremely sparse and high dimensional, there is often a genuine need for dimension reduction to facilitate data visualization and downstream statistical analysis. We propose *Principal Amalgamation Analysis* (PAA), a novel amalgamation-based and taxonomy-guided dimension reduction paradigm for microbiome data. Our approach aims to aggregate the compositions into a smaller number of *principal compositions*, guided by the available taxonomic structure, by minimizing a properly measured loss of information. The choice of the loss function is flexible and can be based on familiar diversity indices for preserving either within-sample or between-sample diversity in the data. To enable scalable computation, we develop a hierarchical PAA algorithm to trace the entire trajectory of successive simple amalgamations. Visualization tools including dendrogram, scree plot, and ordination plot are developed. The effectiveness of PAA is demonstrated using gut microbiome data from a preterm infant study and an HIV infection study.

Keywords: data aggregation; dimension reduction; microbiome data; taxonomic hierarchy

check for updates

Citation: Li, Y.; Li, G.; Chen, K. Principal Amalgamation Analysis for Microbiome Data. *Genes* **2022**, *13*, 1139. https://doi.org/10.3390/genes13071139

Academic Editor: Yi-Juan Hu

Received: 3 May 2022
Accepted: 21 June 2022
Published: 24 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Microbiome, defined as the set of microorganisms inhabiting a specific biological niche, plays a critical role in the development, nutrition, immunity, and health of their host organisms when the microorganisms are host associated [1]. The human gut microbiome, for instance, is known to not only control digestion but also affect the immune and nervous systems of human [2–4]. Whether host associated or free living, microbial communities are indispensable components of their ecosystems, and a deep understanding of their community structure and their interactions with the environment could lead to important biological and ecological insights. Indeed, in recent years, microbiome studies have become increasingly prevalent and large-scale, in part due to the rapid advances in high-throughput sequencing technologies. Raw sequencing reads can now be processed by well-established analytical tools such as quantitative insights into microbial ecology (QIIME) and mothur, to produce abundance tables of operational taxonomic units (OTU) [5–7]. Since the number of sequencing reads per sample (i.e., library size) may vary dramatically, proper sampling and normalization procedures are further adopted to produce relative abundance data of the OTUs [8,9], from which downstream analysis is performed.

Microbiome data is complex in nature, subject to constraints such as compositionality, high dimensionality, zero inflation, overdispersion, and taxonomic hierarchy. Specifically, microbiome data, as presented in relative abundances or proportions, are compositional; each compositional vector resides in a simplex that does not admit the standard Euclidean geometry. Second, the data are often very sparse with a large portion of zeros, arising from either under-sampling or true absence of the corresponding taxa or outlier mechanism defined by Kaul et al. [10]. Third, the number of OTUs or taxa is often much larger than the number of samples, making the data analysis prone to many curses of dimensionality. In addition, a unique feature of microbiome data is the presence of the evolutionary history of the taxa charted through a taxonomic tree. This hierarchical structure provides crucial information about the relationship between different microbes and is proven useful in various studies [11–14]. These inherent characteristics of microbiome data impose various statistical challenges and stress the need for developing novel methods to better harness the power of such data.

As the microbiome data is often extremely sparse and high dimensional in many studies, there is a genuine need to properly reduce its dimension to facilitate data exploration, visualization, and downstream analysis. Besides some "naive" data reduction methods such as directly using certain diversity measures as coarse summaries or keeping only the most prevalent taxon [15], ordination methods such as principal component analysis (PCA) and principal coordinate analysis (PCoA) are among the most commonly-adopted approaches in practice. PCA generally relies on transformations that neglect the unique features of microbiome data [16,17], such as zero inflation and taxonomic tree structure; PCoA, on the other hand, is based on a proximity matrix that may accommodate the data features but fails to pinpoint relevant microbes that drive the data reduction [18]. Moreover, for microbiome data with taxonomic information, there is a trade-off between data resolution and accuracy: the lower the taxonomic rank, the higher the data resolution (with more taxa and thus more information), but the sparser and less accurate the data (there are more zeros and each composition is converted from a smaller count). Most existing methods only apply to a prefixed taxonomic rank and/or rely on transformations (with ad-hoc replacement of zeros) that may inflate inaccuracy [19]. These often lead to unstable and biased results [20,21].

We concern a fundamental question: *what constitutes an interpretable and effective dimension reduction of microbiome data?* It is apparent that the answers from the aforementioned approaches are with flaws. Our answer is radically different and yet strikingly intuitive: we argue that for microbiome (relative abundance) data, an effective and interpretable operation for dimension reduction is through aggregating the compositional components, i.e., through the so-called *amalgamation*, a fundamental operation on compositional data [22]. More precisely, if the components of a length-p compositional vector are separated into $k < p$ mutually exclusive and exhaustive subsets and the components of each subset are added together, the resulting length-k compositional vector is termed an amalgamation. For instance, $(x_1 + x_2, x_3, x_4 + x_5)$ is an amalgamation of $(x_1, x_2, x_3, x_4, x_5)$. Given its simplicity, it is not surprising that amalgamation has been widely used in practice, but mostly in a rather ad-hoc way, e.g., combining a number of compositional components with the lowest prevalence. Not until recently, a few studies on amalgamation-based dimension reduction emerged [23,24]. A more detailed review on existing dimension reduction methods for microbiome data is provide in Section 2.

We propose *Principal Amalgamation Analysis* (PAA), an amalgamation-based and taxonomy-guided dimension reduction paradigm for microbiome data. Our PAA approach directly handles the compositional data without the need for transformation and reduces its dimension by clustering and aggregating the compositions based on minimizing certain information loss, subject to confinement to the taxonomic hierarchy. The choice of the loss function can be flexible and problem specific; for example, it can be based on diversity measures such as α diversity and β diversity to examine and preserve either within-sample (between-species) or between-sample (between-habitat) diversity of the data. To enable scalable computation, we develop and implement an efficient agglomerative clustering algorithm to identify the entire trajectory of the successive simple amalgamation steps. This allows us to start from the raw OTUs at their lowest taxonomic ranks and gradually amalgamate them until a desired balance between information loss and dimension reduction is reached. As such, PAA alleviates the bias and instability introduced by zero replacement and data transformations, maintains the compositional and taxonomic structures of the reduced data, and offers superior interpretation and visualization through the resulting "principal compositions".

2. Existing Work on Dimension Reduction of Microbiome Data

Microbiome data are often normalized as compositions [8,9], which reside in a simplex that does not admit the standard Euclidean geometry. It creates significant challenges for statistical analysis, as many standard methods do not directly apply. There have been developments on compositional data analysis based on transformations and the so-called Aitchison geometry [16,22,25–27]. However, these transformations could be inadequate to accommodate the unique features of microbiome data such as zero inflation, over dispersion and the presence of taxonomic tree structure among microbes.

The existing data reduction approaches for microbiome compositional data can be summarized to four categories, namely, the indexing approach, the selection approach, the transformation-based approach, and the amalgamation-based approach. The indexing approach, which represents data by some diversity or complexity indices, typically disables taxon-level analysis and results in oversimplification of data [28–30]. The selection approach [31,32], which only keeps a subset of "dominant" compositions, often ignores intrinsic relations due to compositionality and can be vulnerable to the extremely low abundances of many OTUs. In practice, such selection could trivially end up with a few most prevalent ones [33]. The transformation-based approach conveniently utilizes existing reduction methods such as PCA after transforming data to the Euclidean space [25,34–37]. The required transformations usually involve logarithm operations and cannot be directly applied on the excessive zeros in the microbiome data. An alternative is to use power transformations such as square-root transformation [36,38], which avoids zero replacement and puts the data onto the unit sphere to enable manifold-based PCA. However, these PCA approaches may compromise interpretation of the data in terms of individual taxon and impede incorporation of the extrinsic taxonomic tree structure. Other proximity-based methods such as PCoA [39,40] could accommodate several special features of the data but fail to pinpoint specific taxon that drive data reduction.

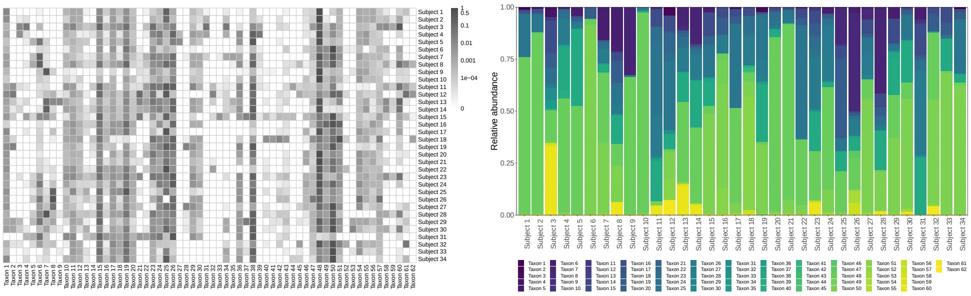
While Aitchison's formal terminology of "amalgamation" may not be as widely spread as it should be, the operation itself has nevertheless been widely used in microbiome data analysis, although often as a pragmatic and rather ad-hoc way of dealing with the most rare compositional components in the data. For example, rare taxon with excessive number of zeros or low abundances at lower taxonomic ranks are aggregated to a higher rank for analysis [11,41]. It is also common in the microbiome analysis that rare taxon are simply removed by some ad-hoc filtering process [42]. These naive approaches may lead to unwanted information loss and potential conflicts between analyses performed at different ranks or with different filter rules.

Not until recently, a few studies on amalgamation-based dimension reduction emerged. Greenacre [43] and Greenacre et al. [23] argued that amalgamation provides an interpretable way to reduce the dimensionality of compositions and could make substantive sense in practical applications, despite the non-linearity in the Aitchison geometry of the simplex and its possibility to distort between-sample distances. They advocated for expert-driven amalgamation, i.e., the use of domain-knowledge to perform amalgamation, and proposed amalgamation-based hierarchical clustering with log-ratio transformed data. Quinn and Erb [24] further discussed the usage of amalgamation as an alternative to the commonly-adopted dimension reduction methods and proposed an optimization approach to preserve a suitable between-sample distance measure with centered log-ratio transformation. However, the method does not work without zero-replacement, and its genetic algorithm can be extremely computational intensive and hinders the incorporation of extrinsic information such as the taxonomic tree structure.

3. Setup with an Illustrative Example

To illustrate the proposed taxonomy-guided dimension reduction, we consider a preterm infant study conducted at a Neonatal Intensive Care Unit (NICU) in the northeast region of the U.S. Fecal samples of preterm infants were collected daily when available during the infant's first month of postnatal age. Bacterial DNA was isolated and extracted from each sample [44,45]; V4 regions of the 16S rRNA gene were sequenced using the Illumina platform and clustered and analyzed using QIIME [6] to produce microbiome abundance data. When infant reached 36–38 weeks of postmenstrual age, neurobehavioral outcomes were measured using the NICU Network Neurobehavioral Scale (NNNS) [45]. The main interest was examining the gut-brain axis, i.e., whether and how gut microbiome compositions during early postnatal stage impact later neurobehavioral outcomes.

The raw microbiome data is longitudinal and has more than one thousand operational taxonomic units (OTU); these OTUs were classified up to the genus level using the Ribosomal Database Project (RDP) Classifier [46]. For the purpose of illustration, we consider the average compositions over the postnatal period at the genus level, which results in a single dataset with $n = 34$ subjects and $p = 62$ taxa. Figure 1a displays a heatmap of the data and Figure 1b shows the relative abundance barplot of the data. Figure 2 displays the taxonomic tree of the 62 taxa up to the genus level.



(**a**) Heatmap

(**b**) Barplot of relative abundance

Figure 1. The NICU data: Heatmap and barplot of the relative abundance data.

To set up, suppose we observe n independent compositional samples on p taxa; let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_p) = [x_{ij}]_{n\times p}$ be the observed $n \times p$ compositional data matrix, where each row $\boldsymbol{x}_i$, $i = 1, \ldots, n$, lies in $\mathbb{S}^{p-1}$, with $\mathbb{S}^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : \sum_{j=1}^{p} x_j = 1, x_j \geq 0, j = 1, \ldots, p\}$ representing a $(p-1)$-simplex in $\mathbb{R}^p$. As seen from the heatmap in Figure 1, microbiome data is often very sparse with the presence of many rare taxa; even after being aggregated to the genus level, the percentage of zero entries in the data is still close to 40%.

In addition, we assume the availability of a taxonomic tree structure of the p taxa. Some general terminologies of a tree structure are defined as follows. Let T represent a p-leafed taxonomic tree, $I(T)$ the set of internal nodes, and $|T|$ the total number of nodes in a tree. Each leaf node of the tree corresponds to a taxon, and each internal node corresponds to a group of taxa. We follow the commonly used notions of child, parent, sibling, descendant, and ancestor to describe relations between nodes. Let the depth of a node E, denoted as $\mathcal{D}(E)$, be the number of ancestors from the node to the root, and let the depth of a tree, denoted as $\mathcal{D}^*(T)$, be the maximum depth of its leaf nodes. For a leaf node E, we use $A^*(E)$ to denote its lowest multi-child ancestor that has more than one child. For example, in Figure 2, the depth of Taxon 1 is 1, while that of Taxon 2 is 5. The lowest multi-child ancestor of Taxon 12 is its parent, while that of Taxon 13 is its grandparent as its parent has only one child; that is, they share the same lowest multi-child ancestor. Taxon 26 and Taxon 27, on the other hand, do not share the same parent nor the lowest multi-child ancestor. We remark that we do not require the tree to be "complete", which means that different taxa could be classified at different taxonomic ranks. For instance, Taxon 1 is classified at the Phylum level, and Taxa 10 is only identified at the Class level.

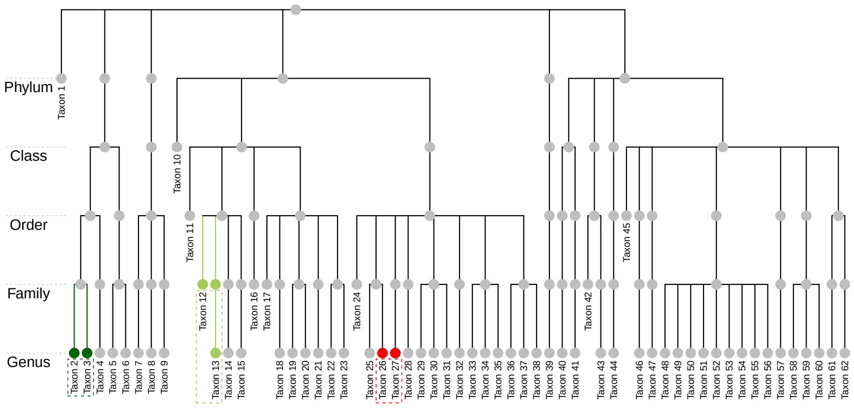


Figure 2. The NICU data: Taxonomic structure of the 62 taxa at the genus level. Taxon 2 and Taxon 3 (dark green nodes) share the same parent and are at the same taxonomic rank. Taxon 12 and Taxon 13 (light green nodes) are at different ranks but they share the same lowest multi-child ancestor. Taxon 26 and Taxon 27 (red nodes) do not share the same lowest multi-child ancestor.

4. Taxonomy-Guided Principal Amalgamation Analysis

Analogous to PCA, which finds a number of principal components to best preserve the total variation in the data, PAA aims to aggregate the compositions to a smaller number of *principal compositions*, guided by the available taxonomic structure, to best preserve a proper measure of information in the data.

In this section, we start with a general framework of PAA as an information-preserving optimization procedure in Section 4.1. To achieve scalable computation and conveniently utilize the taxonomic structure, a hierarchical agglomerative PAA (HPAA) algorithm is developed in Section 4.2. The computational details with various loss functions of interest are provided in Section 4.3, and the graphical tools for visualizations of PAA are illustrated in Section 4.4.

4.1. Framework

The *amalgamation* is a fundamental operation for compositions; it is formally defined as follows.

Definition 1 (Amalgamation)**.** *Let $\boldsymbol{x} \in \mathbb{S}^{p-1}$. For any $1 < k < p$, define*

$$\mathcal{M}_0(k,p) = \Big\{\boldsymbol{R} = [r_{ij}]_{k\times p}; r_{ij} \in \{0,1\}, \textstyle\sum_{i=1}^{k} r_{ij} = 1 \mathit{\ for\ } j = 1,\ldots,p\Big\}.$$

Then, $\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x}$ gives an amalgamation of $\boldsymbol{x}$ when $\boldsymbol{R} \in \mathcal{M}_0(k,p)$.

The matrix $\boldsymbol{R} = (\boldsymbol{r}_1,\ldots,\boldsymbol{r}_k)^T \in \mathbb{R}^{k\times p}$ is called an amalgamation matrix. It is clear that the k many $\boldsymbol{r}_j$ vectors represent k mutually exclusive and exhaustive subsets of the p components in $\boldsymbol{x}$, and each $\boldsymbol{r}_j^T\boldsymbol{x}$ then computes a sum of the components of $\boldsymbol{x}$ in the jth subset. The operation of amalgamation reduces the original compositional vector in $\mathbb{S}^{p-1}$ to a simplex with dimension at most k, as it is possible that $\sum_{j=1}^{p} r_{ij} = 0$ for some $i = 1,\ldots,k$. Apparently, this operation may result in a loss of information whenever $k < p$.

Naturally, given the observed data $\boldsymbol{X}$, PAA can be formulated as a set of optimization problems: for $k = 2,\ldots,p-1$,

$$\widehat{\boldsymbol{R}}_k \in \arg \min_{\boldsymbol{R}\in\mathcal{M}_0(k,p)} L(\boldsymbol{R};\boldsymbol{X}), \tag{1}$$

where $L(\cdot)$ is a properly specified loss function that measures certain information reduction from $\boldsymbol{X}$ to $\boldsymbol{X}\boldsymbol{R}^T$. Borrowing the terminology from PCA, we call the resulting $\widehat{\boldsymbol{R}}_k$ matrix as the *loading matrix* or the *principal amalgamation matrix* and the amalgamated data $\boldsymbol{X}\widehat{\boldsymbol{R}}_k^T$

as the *score matrix*. Subsequently, the k amalgamated vectors, $\boldsymbol{X}\widehat{\boldsymbol{r}}_1, \ldots, \boldsymbol{X}\widehat{\boldsymbol{r}}_k$ are called *principal compositions*.

The construction of the loss function is flexible, and it is tied to the choice of how to measure the information in the data. For microbiome data or compositional data in general, of particular interest in practice is to measure the information loss by the reduction in some diversity index, for preserving a specific aspect of diversity in the original data as much as possible. Popular choices include the family of α diversity such as Simpson's diversity index (SDI) and Shannon–Wiener index (SWI), which measures within-sample diversity, and the family of β diversity such as Bray-Curtis dissimilarity (BC) and Weighted UniFrac (WUF), which measures between-sample diversity [29,47–49]. There are also entropy-based or model-based measures that incorporate different aspects of the above indices [50–54]. Other methods for constructing the loss function include the likelihood approach, i.e., through making distributional assumption of the data (e.g., Dirichlet), and the transformation approach, i.e., through transforming the data to the Euclidean space such that familiar statistics such as sample variance can be used. To focus on the main idea, we defer the detailed discussions on these choices in Section 4.3.

However, due to the complex structure of $\mathcal{M}_0(k,p)$, conducting PAA through exactly solving the optimization problems in Equation (1) would be computationally prohibitive when p is large, not even to mention that utilizing the taxonomic structure may introduce further complications. Intriguingly, we realize that PAA can be pursued from a cluster analysis perspective. With any specified number of principal compositions, the objective of PAA is essentially to search for a cluster pattern of the compositions, such that each cluster of compositions aggregates into a new principal composition.

4.2. Hierarchical PAA Algorithm

Motivated by the connection between PAA and clustering, we develop an agglomerative hierarchical PAA (HPAA) algorithm to utilize taxonomic structure and enable scalable computation. Our approach starts from the original compositions and gradually amalgamates them through a sequence of simple amalgamations, i.e., at each step only a single pair of compositions is being aggregated. As such, HPAA generates the entire path of the simple amalgamations for reducing the data from its most informative original form to an utterly non-informative vector of ones. Along this process, a dendrogram of the successive amalgamations and the associated information losses is naturally generated.

We first describe the HPAA algorithm and the growth of the associated dendrogram in its basic form, without consideration of taxonomic guidance. To initialize, let $t = 0$ and denote $\boldsymbol{X}_0 = \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_p)$ as the original $n \times p$ compositional data matrix. We start from the p taxa in the original data $\boldsymbol{X}_0$, each of which forms its own cluster. Let $\mathcal{S}_0 = \{\{1\}, \ldots, \{p\}\}$ be the initial partition of the p taxa, which correspondingly forms the initial leaf nodes of the dendrogram, denoted as $E_{0,1}, \ldots, E_{0,p}$.

At the tth step, for $t = 1, \ldots, p-1$, let $\mathcal{S}_{t-1}$ denote the set of $|\mathcal{S}_{t-1}|$ (i.e., $p - t + 1$) nodes and $\boldsymbol{X}_{t-1}$ denote the $n \times |\mathcal{S}_{t-1}|$ current amalgamated data from the last step. With these inputs, the core problem is to search for a pair of the current nodes, $(E_{t-1,\widehat{j}}, E_{t-1,\widehat{j'}})$, to be aggregated into a new node such that the information loss of the amalgamated data is minimized. That is,

$$(\widehat{j}, \widehat{j'}) = \arg \min_{(j,j') \in \mathcal{P}_{t-1}} L(\boldsymbol{R}(j,j'); \boldsymbol{X}_{t-1}), \tag{2}$$

where $\boldsymbol{R}(j,j')$ is a simple amalgamation matrix in $\mathcal{M}_0(|\mathcal{S}_t|, |\mathcal{S}_{t-1}|)$ that aggregates the jth and j'th columns of $\boldsymbol{X}_{t-1}$, and $\mathcal{P}_{t-1}$ is the active set of "legitimate" pairs of nodes that can be amalgamated. For instance, if no restriction is imposed, we set $\mathcal{P}_{t-1} = \{(j,j'); 1 \leq j < j' \leq |\mathcal{S}_{t-1}|\}$, consisting of all possible pairs of the current leaf nodes.

With the solution from Equation (2), we then update

$$\boldsymbol{X}_t \leftarrow \boldsymbol{X}_{t-1}\boldsymbol{R}(\widehat{j}, \widehat{j'})^T$$

and denote the reduced set of nodes as $E_{t,1}, \ldots, E_{t,p-t}$. For example, if at the first step ($t = 1$), $E_{0,1}$ and $E_{0,2}$ are chosen to be combined, we have $\mathcal{S}_1 = \{\{1,2\}, \{3\}, \ldots, \{p\}\}$, $\boldsymbol{X}_1 = (\widetilde{\boldsymbol{x}}_1 + \widetilde{\boldsymbol{x}}_2, \widetilde{\boldsymbol{x}}_3, \ldots, \widetilde{\boldsymbol{x}}_p)$, and the new reduced set of nodes are denoted as $E_{1,1}, \ldots, E_{1,p-1}$.

The above procedure is repeated until only two nodes are left; they are then bound to be combined as a vector of ones. The proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Hierarchical principal amalgamation analysis (HPAA) via agglomerative clustering

```
**Parameters:** Compositional data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, and a user-specified loss function $L(\boldsymbol{R}; \boldsymbol{X})$.
**Initialization**: Set $\boldsymbol{X}_0 = \boldsymbol{X}$. Set the initial partition as $\mathcal{S}_0 = \{E_{0,j} = \{j\}, j = 1, \ldots, p\}$,
    where $E_{0,j} = \{j\}$ means the node/cluster $E_{0,j}$ is formed by the $j$th taxon only.
**For** $t = 1, 2, \ldots, p-1$,
    • Search for a pair of current nodes $E_{t-1,\widehat{j}}$ and $E_{t-1,\widehat{j'}}$ to perform amalgamation, by
      solving Equation (2).
    • Combine $E_{t-1,\widehat{j}}$ and $E_{t-1,\widehat{j'}}$ to be a new node, and accordingly update $\mathcal{S}_t$, $\boldsymbol{X}_t$ and
      $E_{t,j}$ $(j = 1, \ldots, p-t)$.
**End For.**
```

We propose three levels of taxonomy guidance: unconstrained, weak taxonomic hierarchy, and strong taxonomic hierarchy, which, as the names suggest, produce amalgamation patterns with different degrees of conformity with the taxonomic tree. It all boils down to properly set the active set of the paired nodes $\mathcal{P}_{t-1}$ in solving Equation (2). Moreover, when either weak or strong taxonomic hierarchy is enforced, the successive growth of the dendrogram through guided amalgamations is always coupled with the successive reduction of the taxonomic tree.

- *Unconstrained.* In each step, we search over all possible pairs of nodes in solving Equation (2),
$$\mathcal{P}_{t-1} = \{(j, j'); 1 \leq j < j' \leq |\mathcal{S}_{t-1}|\}.$$
- *Weak taxonomic hierarchy*. In each step, we only search over pairs of nodes that share the same lowest multi-child ancestor in the reduced taxonomic tree. That is, Equation (2) is solved over
$$\mathcal{P}_{t-1} = \{(j, j'); 1 \leq j < j' \leq |\mathcal{S}_{t-1}|, A^*(E_{t-1,j}) = A^*(E_{t-1,j'})\}.$$
For example, consider the first step of HPAA ($t = 1$) with the $p = 62$ leaf nodes in Figure 2. Both (Taxon 2, Taxon 3) and (Taxon 12, Taxon 13) are in $\mathcal{P}_0$, while (Taxon 26, Taxon 27) is not.
- *Strong taxonomic hierarchy*. In each step, we further restrict the search to be among pairs of nodes that have the largest depth in the taxonomic tree. As a result, taxa at the lowest taxonomic rank will always be aggregated first. That is, Equation (2) is solved over
$$\begin{aligned}\mathcal{P}_{t-1} = \{(j, j'); & 1 \leq j < j' \leq |\mathcal{S}_{t-1}|,\\ & A^*(E_{t-1,j}) = A^*(E_{t-1,j'}),\\ & \mathcal{D}(E_{t-1,j}) = \mathcal{D}(E_{t-1,j'}) = \mathcal{D}^*_{t-1}\}.\end{aligned}$$
For example, in Figure 2, the pair (Taxon 2, Taxon 3) remains in $\mathcal{P}_0$, while (Taxon 12, Taxon 13) is no longer in $\mathcal{P}_0$ as they are not of the lowest taxonomic rank.

4.3. Construction of Loss Function with Common Diversity Measures

We illustrate the implementation of HPAA using loss functions constructed from several commonly-used α diversity and β diversity measures.

The α diversity measures the richness (number of different entities) and evenness (the homogeneity in abundance of the entities) within each compositional sample. It can be calculated for each sample in the data, i.e., $\alpha(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$. In general, larger $\alpha(\boldsymbol{x}_i)$ indicates larger within-sample diversity among species, and the index is non-increasing along successive amalgamations. Therefore, a general loss function based on α diversity can be constructed as

$$L_\alpha(\boldsymbol{R}; \boldsymbol{X}) = -\sum_{i=1}^{n} \alpha(\boldsymbol{R}\boldsymbol{x}_i). \tag{3}$$

4.3.1. Simpson's Diversity Index (SDI)

Consider first the Simpson's diversity index (SDI), defined as

$$\mathrm{SDI}(\boldsymbol{x}_i) = 1 - \textstyle\sum_{j=1}^{p} x_{ij}^2 = 1 - \boldsymbol{x}_i^T \boldsymbol{x}_i,$$

where x_{ij} represents the abundance of the jth components in the ith sample with $\boldsymbol{x}_i \in \mathbb{S}^{p-1}$. The SDI can be understood as the probability that two individuals randomly selected from a sample will belong to different species. A small SDI indicates that a few components dominate, while a large SDI indicates a diverse and balanced distribution among components. It is seen that SDI is non-increasing along successive amalgamations, as $x_{ij}^2 + x_{ij'}^2 \leq (x_{ij} + x_{ij'})^2$ for $x_{ij}, x_{ij'} \geq 0$. Therefore, with the form of the loss function in Equation (3), the general PAA criterion in Equation (1) becomes

$$\min_{\boldsymbol{R} \in \mathcal{M}_0(k,p)} \mathrm{tr}(\boldsymbol{R}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{R}^T),$$

and the tth step simple amalgamation problem in Equation (2) becomes

$$\min_{(j,j') \in \mathcal{P}_{t-1}} \widetilde{\boldsymbol{x}}_{j,t-1}^T \widetilde{\boldsymbol{x}}_{j',t-1},$$

where $\widetilde{\boldsymbol{x}}_{j,t=1}$ denotes the jth column of $\boldsymbol{X}_{t-1}$, which is equivalent to find the minimal off-diagonal element of $\boldsymbol{X}_{t-1}^T\boldsymbol{X}_{t-1}$ within the active set specified by $\mathcal{P}_{t-1}$. We remark that in each step only two columns of the amalgamated data are affected; this is utilized to simplify the computation.

4.3.2. Shannon–Wiener Index (SWI)

Unlike the SDI which weights more on dominant components, the Shannon–Wiener index (SWI) is equally sensitive to rare and dominant components, defined as

$$\mathrm{SWI}(\boldsymbol{x}_i) = -\textstyle\sum_{j=1}^{p} x_{ij} \log x_{ij} = -\boldsymbol{x}_i^T \log(\boldsymbol{x}_i).$$

The logarithmic transformation is applied entrywisely on the enclosed vector or matrix. As such, to compute SWI, we do need to first replace zeros in the data. The SWI is non-increasing along successive amalgamations since $x_{ij} \log x_{ij} + x_{ij'} \log x_{ij'} \leq (x_{ij} + x_{ij'}) \log(x_{ij} + x_{ij'})$ for $x_{ij}, x_{ij'} > 0$. Therefore, with the loss function form in Equation (3), the general PAA criterion in Equation (1) becomes

$$\min_{\boldsymbol{R} \in \mathcal{M}_0(k,p)} \mathrm{tr}\{\boldsymbol{R}\boldsymbol{X}^T \log(\boldsymbol{X}\boldsymbol{R}^T)\},$$

and the tth step simple amalgamation problem in Equation (2) becomes

$$\begin{aligned}\min_{(j,j') \in \mathcal{P}_{t-1}} &\{(\widetilde{\boldsymbol{x}}_{j,t-1} + \widetilde{\boldsymbol{x}}_{j',t-1})^T \log(\widetilde{\boldsymbol{x}}_{j,t-1} + \widetilde{\boldsymbol{x}}_{j',t-1}) \\ &- \widetilde{\boldsymbol{x}}_{j,t-1}^T \log(\widetilde{\boldsymbol{x}}_{j,t-1}) - \widetilde{\boldsymbol{x}}_{j',t-1}^T \log(\widetilde{\boldsymbol{x}}_{j',t-1})\}.\end{aligned}$$

While the α diversity focuses on within-sample diversity, the β diversity reflects the between-sample differences. It can be calculated for each pair of samples in the data, i.e., $\beta(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ for $i, i' = 1, \ldots, n$, resulting in a between-sample distance or dissimilarity matrix $\boldsymbol{D}(\boldsymbol{X}) = [\beta(\boldsymbol{x}_i, \boldsymbol{x}_{i'})]_{n \times n}$. As such, PAA aims to best preserve the dissimilarity pattern in the amalgamated data. We thus construct the loss function based on β diversity as the sum of squared differences between the original distance matrix and that of the amalgamated data,

$$L_\beta(\boldsymbol{R}; \boldsymbol{X}) = \sum_{i<i'} \{\beta(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) - \beta(\boldsymbol{R}\boldsymbol{x}_i, \boldsymbol{R}\boldsymbol{x}_{i'})\}^2. \tag{4}$$

4.3.3. Bray-Curtis Dissimilarity Index (BC)

Consider the Bray-Curtis dissimilarity index (BC) defined as

$$\begin{aligned} \mathrm{BC}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) &= \frac{\sum_{j=1}^p |x_{ij} - x_{i'j}|}{\sum_{j=1}^p (x_{ij} + x_{i'j})} = \frac{\sum_{j=1}^p |x_{ij} - x_{i'j}|}{2} \\ &= \frac{\sum_{j=1}^p [(x_{ij} + x_{i'j}) - 2\min(x_{ij}, x_{i'j})]}{2} \\ &= 1 - \sum_{j=1}^p \min(x_{ij}, x_{i'j}), \end{aligned}$$

for any pair of samples $\boldsymbol{x}_i, \boldsymbol{x}_{i'} \in \mathbb{S}^{p-1}$. Note in the context of compositional data, the Bray-Curtis dissimilarity is simplified to the Manhattan (City-Block) distance. It is non-increasing along successive amalgamations as $\min(x_{ij}, x_{i'j}) + \min(x_{ij'}, x_{i'j'}) \leq \min(x_{ij} + x_{ij'}, x_{i'j} + x_{i'j'})$. The tth step simple amalgamation problem in Equation (2) can be expressed as

$$\begin{aligned} \min_{(j,j') \in \mathcal{P}_{t-1}} \{ \textstyle\sum_{i<i'} [& \min(x_{ij,t-1}, x_{i'j,t-1}) + \min(x_{ij',t-1}, x_{i'j',t-1}) \\ & - \min(x_{ij,t-1} + x_{ij',t-1}, x_{i'j,t-1} + x_{i'j',t-1})]^2 \}. \end{aligned}$$

4.3.4. Weighted UniFrac Distance (WUF)

The weighted UniFrac distance (WUF) further incorporates information from the phylogenetic or taxonomic tree when computing the between-sample distance. It can also be viewed as a plug-in estimate of the Wasserstain distance between two probability distributions defined on the taxonomic tree [55]. It is commonly used in exploratory microbiome data analysis and a number of variants were developed. To mention a few, double principal coordinate analysis (DPCoA) proposed by Pavoine et al. [56] generalized PCA by incorporating the relationship among variables from the phylogenetic structure that can be described using dissimilarity measures like UniFrac or weighted UniFrac. Chen et al. [57] compared the power of statistical tests using a number of variants of UniFrac including unweighted/weighted UniFrac and generalized UniFrac. Randolph et al. [11] proposed a kernel-based regression framework that incorporates the unweighted/weighted UniFrac dissimilarity matrix from the phylogenetic structure.

Here we briefly illustrate HPAA with weighted UniFrac distance. For any pair of samples $\boldsymbol{x}_i, \boldsymbol{x}_{i'} \in \mathbb{S}^{p-1}$, WUF is defined as

$$\mathrm{WU}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{\sum_{j=1}^p l_j |x_{ij} - x_{i'j}|}{\sum_{j=1}^p L_j (x_{ij} + x_{i'j})},$$

where l_j for $j = 1, \ldots, p$ denotes the length of the j branch, i.e., the length between the node for jth entity and its parent, and L_j denotes the distance of jth entity from the root node of the phylogenetic tree. Here the length of branches may change with compositions at lower levels of taxonomic tree amalgamated to higher level.

At the tth step, the simple amalgamation problem in Equation (2) can be expressed as

$$\min_{(j,j')\in\mathcal{P}_{t-1}} \{\sum_{i<i'} [\frac{\sum_{j=1}^{p} l_j|x_{ij} - x_{i'j}|}{\sum_{j=1}^{p} L_j(x_{ij} + x_{i'j})} - \frac{\sum_{k=1,k\neq j,j'}^{p} l_k|x_{ik} - x_{i'k}| + l_{j,j'}|(x_{ij} + x_{ij'}) - (x_{i'j} + x_{i'j'})|}{\sum_{k=1,k\neq j.j'}^{p} L_k(x_{ik} + x_{i'k}) + L_{j,j'}(x_{ij} + x_{ij'} + x_{i'j} + x_{i'j'})}]^2\},$$

where $l_{j,j'}$ denotes the the length between the newly formed entity from jth and j'th entities and its parent, and $L_{j,j'}$ denotes the distance of new entity from the root node of the phylogenetic tree. During the successive amalgamations, the lengths of branches in computing WUF are also getting updated.

4.4. Visualization with Examples

We use the NICU data to illustrate the graphical tools developed for visualizing the PAA results. These tools can be extremely useful for visualizing and understanding compositional data, as well as helping to determine the desired number of principal compositions in practice.

4.4.1. Dendrogram

We construct a HPAA dendrogram to simultaneously visualize both the tree diagram of the successive amalgamations and the taxonomic structure of the taxon. To illustrate, Figure 3 shows the HPAA dendrogram from performing HPAA with SDI loss and strong taxonomic hierarchy on the NICU data. The top part of figure shows the dendrogram of amalgamations, where the y-axis shows the percentage decrease in total diversity as measured by SDI (on the log-scale) along the successive amalgamations, from the bottom to the top. As such, any horizontal cut of the dendrogram at a desired level of diversity loss/preservation shows the corresponding amalgamated data. In particular, each red dashed horizontal line indicates the steps at which the original data are aggregated to a higher taxonomic rank. It shows that, for example, aggregating data to the order level (22 taxa or principal compositions left) through HPAA leads to 22.3% loss in total SDI. At the bottom part, we use color bars to show taxonomic structure of the taxa (as shown in Figure 1), where in each horizontal bar taxa of the same color belong to the same category of that rank.

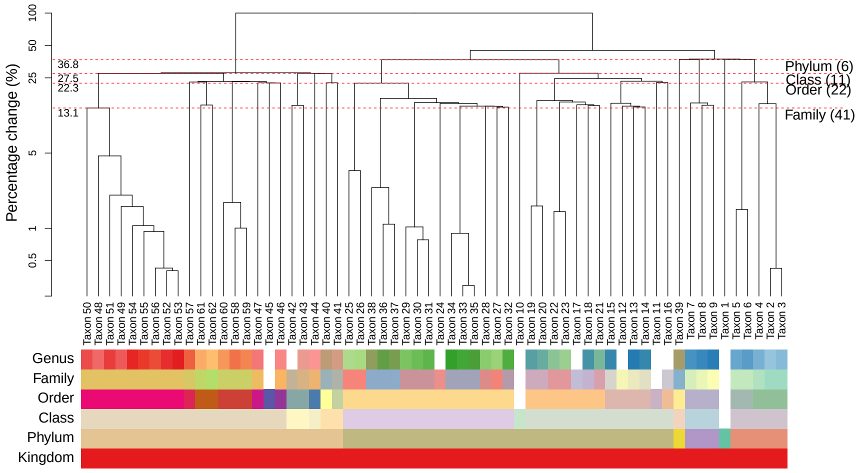


Figure 3. The NICU data: Dendrogram of HPAA with SDI and strong taxonomic hierarchy.

Figures S1 and S2 in Supplementary Information A show HPAA dendrograms with SDI loss and BC loss, respectively, under all three levels of taxonomy guidance. Not surprisingly, the patterns of amalgamations vary under different settings. Without taxonomic constraint, the change in diversity appears to be very smooth along the amalgamations, but the resulting principal compositions may not be easily interpretable, as indicated by the mixed color patterns in the color bars of the taxonomic rank. On the other hand, for the setting of strong taxonomic hierarchy, while the principal compositions are forced to closely follow the taxonomic structure, the percentage change in diversity tends to exhibit dramatic jumps, especially at the steps that the last remaining taxon at a lower taxonomic rank is forced to be aggregated to a higher rank. As a compromise, for the setting of weak taxonomic hierarchy, the resulting principal compositions remain interpretable, and the percentage change in diversity remains smooth and can be quite close to that of the unconstrained setting in the early stage of amalgamations.

The HPAA dendrogram also reveal several interesting insights on the microbiome of preterm infants. As shown in Figure 3, while Taxa 49–56 are all genus of the Enterobacteriaceae family, the pattern of amalgamation suggests that Taxon 50, which is Klebsiella, is distinctive with the rest. It turns out that Klebsiella is a genus of Enterobacteriaceae that has emerged as a significant nosocomial pathogen in neonates [58,59], and its species have been implicated as a cause of various neonatal infections [60,61] and neonatal sepsis [62,63].

4.4.2. Scree Plot

The scree plot shows the percentage change in the diversity loss as a function of the number of principal compositions. Figure 4 shows the scree plots from performing HPAA on the NICU data under different settings. The difference among the three levels of taxonomic guidance is very revealing, which confirms the previous observation from the dendrograms that the setting of weak taxonomic hierarchy reaches a good balance between preserving information and interpretability.

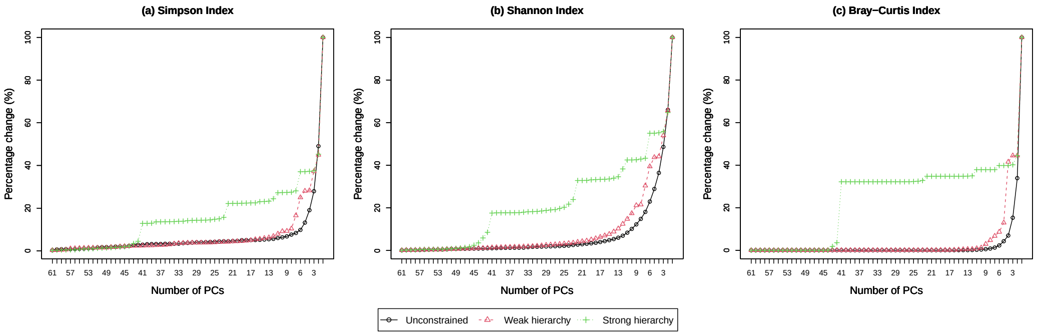


Figure 4. The NICU data: Scree plots for HPAA with (**a**) Simpson's index, (**b**) Shannon's index and (**c**) Bray-Curtis dissimilarity under different hierarchy settings (Percentage change in diversity vs. number of principal compositions).

4.4.3. Ordination Plot

We construct an ordination plot to visualize the changes in the between-sample distance patterns before and after HPAA with any given number of principal compositions. Specifically, we perform the non-metric multidimensional scaling (NMDS) analysis with Bray–Curtis dissimilarity on the combined original data and the principal compositions from HPAA, which produces a low-dimensional ordination plot of all samples before and after amalgamation. For each sample, it is represented by a pair of points from either the original data or the principal compositions; the smallest circle that covers the pair is drawn, whose radius then indicates the level of distortion due to HPAA data reduction.

Figure 5 shows the ordination plots from performing HPAA on the NICU data with three different loss functions and weak taxonomic hierarchy, in which 20 principal compositions are kept. All three settings preserve the between-sample diversity reasonably well, as indicated by the fact that the circles generally have a small radius; as expected, HPAA with the BC loss performs the best as it directly targets on preserving between-sample diversity.

Figure S3 in Supplementary Information A shows the ordination plots from performing HPAA on the NICU data with the BC loss and weak taxonomic hierarchy, with different numbers of principal compositions. As expected, the larger the number of principal compositions, the better the preservation of the between-sample diversity and the less reduction of the size of the data. In practice, such plots, together with the associated statistics, could be very useful in determining the appropriate number of principal compositions.

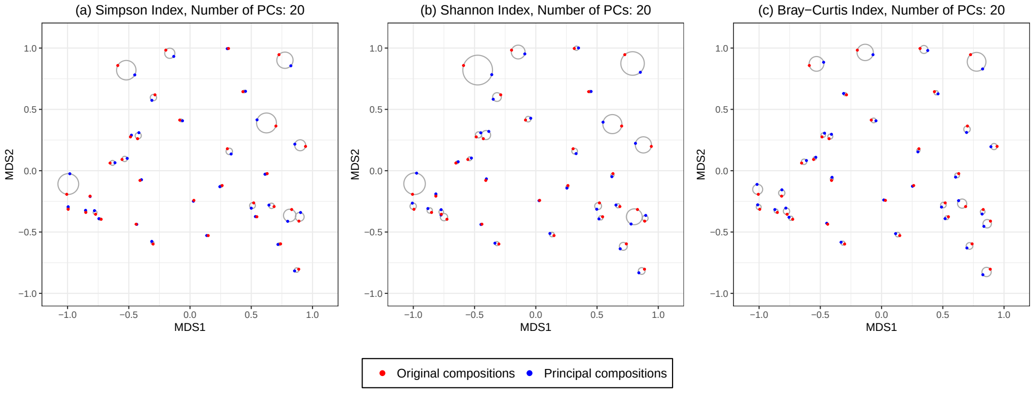


Figure 5. The NICU data: 2D NMDS ordination plots for comparing original and principal compositions from HPAA with (**a**) Simpson's index, (**b**) Shannon's index and (**c**) Bray-Curtis dissimilarity under weak taxonomic hierarchy.

5. Comparison with Competing Methods

To the best of our knowledge, a highly relevant competitor of PAA is the approach proposed by Quinn and Erb [24]; it is implemented in an R package `amalgam` that is available from GitHub repository amalgam https://github.com/tpq/amalgam (accessed on 13 June 2022). For any prespecified dimension of the amalgamated data, the method, referred to herein as `amalgam`, maximizes the correlation between the two Euclidean distance matrices computed from centered log-ratio transformed data before and after amalgamation. Due to the combinatory and nonconvex nature of the problem, a genetic algorithm was proposed to conduct local optimization. Besides `amalgam`, a similar amalgamation method by Greenacre [43] is implemented in an R package `easyCODA`, which is based on preserving the variance of log-ratio transformed data. The method, referred to as ACLUST, performs hierarchical clustering, in which two clusters that give the least loss of variation in the log-ratio transformed data [64] are amalgamated at each step. ACLUST can be regarded as an unconstrained HPAA with a transformation-based loss function that requires zero replacement. Both `amalgam` and ACLUST algorithms can be extremely computational intensive and hinder the incorporation of the taxonomic tree structure. We also consider naive prevalence-based filtering method that simply discard taxa with low abundance.

5.1. Simulation

We first compare the computation efficiency of HPAA, `amalgam`, and ACLUST. The results show that HPAA is very computationally efficient and scales well with the increase of the dimension or the sample size. In contrast, `amalgam` and ACLUST are very computationally intensive even for moderately large p or n, making it unsuitable for large-scale microbiome studies.

We also compare different dimension reduction methods on how well they preserve the between-sample distance pattern, which is very importance in many biological applications. The results show that HPAA methods with different loss functions outperform the baseline, the `amalgam`, and the ACLUST methods. PAA with the Bray-Curtis loss performs the best, as it directly aims at preserving the Bray-Curtis dissimilarity. To our surprise, the `amalgam` method performs worse than the baseline, which may be due to its requirement of zero-replacement and log-ratio transformation and the slow convergence of its genetic algorithm. Moreover, the ACLUST method performs even worse than the `amalgam`. See details in Supplementary Information B.

5.2. Application: Microbiome and HIV Infection

Understanding the association between microbiome richness and HIV-1 risk may help to design novel interventions to improve HIV-1-associated immune dysfunction. Here we considered a cross-sectional HIV microbiome study conducted in Barcelona, Spain, that included both HIV-infected subjects and HIV-negative controls [65]. Gut microbiome data were obtained from MiSeq 16S rRNA sequence data on fecal microbiomes and bioinformatically processed with mothur. The main goal of the study is to find the association between HIV transmission group (MSM: men who have sex with men vs non-MSM), HIV infection status and relative abundance of microbiome composition. As reported by Noguera-Julian et al. [65], risk factors related with sexual preference such as MSM and non-MSM might greatly affect the gut microbiome composition, and thus the relative abundance of taxa might be able to identify the risk clusters of subjects. Following Quinn and Erb [24], the microbiome abundance data were preprocessed to produce a genus-level relative abundance data matrix for $p = 60$ taxa and $n = 128$ HIV-infected subjects, including 60 MSM and 55 non-MSM subjects. The percentage of zeros is 36.6%. The taxonomic tree structure of the $p = 60$ taxa was also available as extrinsic information.

With this dataset, we compared different dimension reduction methods in terms of their performance on preserving the between-sample distance and on the classification accuracy of the MSM factor of subjects with the reduced data. For `amalgam`, the number of amalgamations is fixed at $k = 20$. We omitted ACLUST as its implementation in the R package `easyCODA` becomes computationally infeasible for the dimension $p = 60$. To be comparable, we use HPAA with BC loss and weak taxonomic hierarchy to produce $k = 20$ principal compositions. We also include a simple prevalence-based filtering method that only keeps the $k = 20$ taxa with the highest prevalence.

Figure 6 shows the ordination plots of the three methods. The average Euclidean distance (with standard deviation in brackets) between the points representing the original compositions and principal compositions are 0.05 (0.04), 0.09 (0.07), 0.11 (0.09) for HPAA, the naive method, and the amalgam method, respectively. It is revealing that HPAA performs the best in preserving the between-sample distances of the data, which is partly owing to the proper use of the taxonomic structure.

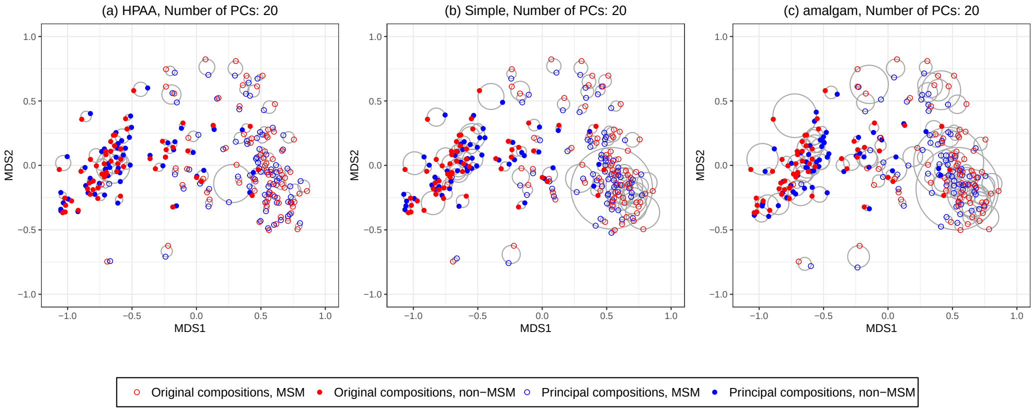


Figure 6. The HIV data: NMDS ordination plots for comparing original data and different principal compositions from different dimension reduction methods including (**a**) HPAA with Bray-Curtis and weak taxonomic hierarchy (HPAA), (**b**) simple approach (Simple), and (**c**) the `amalgam` method by Quinn and Erb [24].

To evaluate the predictive performance of the dimension reduction methods in downstream classification analysis, we conduct an out-of-sample random splitting procedure. For completeness, we also included transformation-based methods, namely, PCA and

PCoA with log-ratio transformed data and $k = 20$. In each run, we random split the original data into a training set of 80% samples and a testing set of 20% samples. Each of the three methods is applied on the training data to produce $k = 20$ features, which are then used as predictors to train a logistic regression model of MSM status. The trained feature construction approach and the logistic regression model are then applied to the testing data, for which the classification performance is measured by the value of the area under the receiver operating characteristic curve (AUC). The whole procedure is repeated 100 times. The average AUC values are 0.84 (0.07), 0.83 (0.08), 0.82 (0.07), 0.82 (0.07), 0.83 (0.08) for HPAA, the naive method, the `amalgam` method, PCA, and PCoA, respectively. We see that the principal compositions from HPAA leads to slightly improved classification. The results showcase the potential of HPAA for improving downstream statistical analysis on both interpretability and prediction.

6. Discussion

We have developed a new approach, principal amalgamation analysis, to perform dimension reduction of microbiome compositional data. The proposed method aggregates the compositions to a smaller number of principal compositions, by minimizing a user-specified loss function subject to conformity to the taxonomic structure. We hope to advocate using it as a preprocessing tool to reduce the dimension of highly-sparse OTU-level relative abundance data.

It is also of particular interest to generalize the current framework to other microbiome data, e.g., metagenomes, metaproteomes, among others. Although we mainly focus on the microbiome compositional data in the proposed principal amalgamation analysis, the proposed framework is applicable as long as the amalgamation operation is meaningful for a particular data type and an appropriate loss function can be used to measure the information loss before and after the amalgamation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13071139/s1, Section A: Visualization Example; Section B: Simulation; Figure S1: HPAA dendrograms with SDI and different constraints on taxonomic hierarchy for the NICU data; Figure S2: HPAA dendrograms with Bray-Curtis and different constraints on taxonomic hierarchy for the NICU data; Figure S3: 2D NMDS ordination plots for comparing original data and different numbers of principal compositions from HPAA with Bray-Curtis and weak taxonomic hierarchy for the NICU data; Figure S4: Simulation on average running time (in second) of HPAA methods with Simpson's index (SDI), Shannon's index (SWI) and Bray-Curtis dissimilarity (BC), and the `amalgam` method by Quinn and Erb [24]. Figure S5: Simulation on accuracy in preserving between-sample Bray-Curtis dissimilarity after dimension reduction.

Author Contributions: Conceptualization and methodology, K.C. and G.L.; analysis design, K.C.; method implementation and numerical analysis, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Source code and the HIV data for numerical studies are available in a public GitHub repository https://github.com/LiYanStat/paaPack(accessed on 13 June 2022). The NICU data from preterm infant study that support the findings in this paper are provided by Dr. Xiaomei Cong. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://figshare.com/s/8f0d7f9a5078c2030c2a (accessed on 13 June 2022). with the permission of Dr. Xiaomei Cong.

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

References

1. Turnbaugh, P.J.; Ley, R.E.; Hamady, M.; Fraser-Liggett, C.M.; Knight, R.; Gordon, J.I. The human microbiome project. *Nature* **2007**, *449*, 804–810. [CrossRef] [PubMed]
2. Turnbaugh, P.J.; Ley, R.E.; Mahowald, M.A.; Magrini, V.; Mardis, E.R.; Gordon, J.I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **2006**, *444*, 1027–1031. [CrossRef]
3. Tremlett, H.; Bauer, K.C.; Appel-Cresswell, S.; Finlay, B.B.; Waubant, E. The gut microbiome in human neurological disease: A review. *Ann. Neurol.* **2017**, *81*, 369–382. [CrossRef] [PubMed]
4. Kau, A.L.; Ahern, P.P.; Griffin, N.W.; Goodman, A.L.; Gordon, J.I. Human nutrition, the gut microbiome and the immune system. *Nature* **2011**, *474*, 327–336. [CrossRef] [PubMed]
5. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef] [PubMed]
6. Caporaso, J.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.; Costello, E.; Fierer, N.; Pea, A.; Goodrich, J.; Gordon, J.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335–336. [CrossRef]
7. Chong, J.; Liu, P.; Zhou, G.; Xia, J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* **2020**, *15*, 799–821. [CrossRef]
8. Gloor, G.B.; Wu, J.R.; Pawlowsky-Glahn, V.; Egozcue, J.J. It's all relative: Analyzing microbiome data as compositions. *Ann. Epidemiol.* **2016**, *26*, 322–329. [CrossRef]
9. Tsilimigras, M.C.; Fodor, A.A. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Ann. Epidemiol.* **2016**, *26*, 330–335. [CrossRef]
10. Kaul, A.; Mandal, S.; Davidov, O.; Peddada, S.D. Analysis of Microbiome Data in the Presence of Excess Zeros. *Front. Microbiol.* **2017**, *8*, 2114. [CrossRef]
11. Randolph, T.W.; Zhao, S.; Copeland, W.; Hullar, M.; Shojaie, A. Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* **2018**, *12*, 540–566. [CrossRef] [PubMed]
12. Xiao, J.; Chen, L.; Yu, Y.; Zhang, X.; Chen, J. A phylogeny-regularized sparse regression model for predictive modeling of microbial community data. *Front. Microbiol.* **2018**, *9*, 3112. [CrossRef] [PubMed]
13. Tanaseichuk, O.; Borneman, J.; Jiang, T. Phylogeny-based classification of microbial communities. *Bioinformatics* **2013**, *30*, 449–456. [CrossRef] [PubMed]
14. Garcia, T.P.; Müller, S.; Carroll, R.J.; Walzem, R.L. Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics* **2013**, *30*, 831–837. [CrossRef] [PubMed]
15. Greenacre, M. Comments on: Compositional data: The sample space and its structure. *TEST* **2019**, *28*, 644–652. [CrossRef]
16. Aitchison, J. Principal component analysis of compositional data. *Biometrika* **1983**, *70*, 57–65. [CrossRef]
17. Aitchison, J.; Greenacre, M. Biplots of compositional data. *J. R. Stat. Soc. Ser.* **2002**, *51*, 375–392. [CrossRef]
18. Lozupone, C.; Lladser, M.E.; Knights, D.; Stombaugh, J.; Knight, R. U niFrac: An effective distance metric for microbial community comparison. *ISME J.* **2011**, *5*, 169–172. [CrossRef]
19. Weiss, S.; Xu, Z.Z.; Peddada, S.; Amir, A.; Bittinger, K.; Gonzalez, A.; Lozupone, C.; Zaneveld, J.R.; Vázquez-Baeza, Y.; Birmingham, A.; et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **2017**, *5*, 27. [CrossRef]
20. Palarea-Albaladejo, J.; Martin-Fernandez, J. Values below detection limit in compositional chemical data. *Anal. Chim. Acta* **2013**, *764*, 32–43. [CrossRef]
21. McMurdie, P.J.; Holmes, S. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **2014**, *10*, e1003531. [CrossRef] [PubMed]
22. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser.* **1982**, *44*, 139–160. [CrossRef]
23. Greenacre, M.; Grunsky, E.; Bacon-Shone, J. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Comput. Geosci.* **2021**, *148*, 104621. [CrossRef]
24. Quinn, T.P.; Erb, I. Amalgams: Data-driven amalgamation for the dimensionality reduction of compositional data. *NAR Genom. Bioinform.* **2020**, *2*, lqaa076. [CrossRef]
25. Aitchison, J.; Egozcue, J.J. Compositional data analysis: Where are we and where should we be heading? *Math. Geol.* **2005**, *37*, 829–850. [CrossRef]
26. Aitchison, J.; Bacon-shone, J. Log contrast models for experiments with mixtures. *Biometrika* **1984**, *71*, 323–330. [CrossRef]
27. Bacon-Shone, J. A Short History of Compositional Data Analysis. In *Compositional Data Analysis: Theory and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2011; pp. 1–11. [CrossRef]
28. Johnson, K.V.; Burnet, P.W. Microbiome: Should we diversify from diversity? *Gut Microbes* **2016**, *7*, 455–458. [CrossRef]
29. Wagner, B.D.; Grunwald, G.K.; Zerbe, G.O.; Mikulich-Gilbertson, S.K.; Robertson, C.E.; Zemanick, E.T.; Harris, J.K. On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities. *Front. Microbiol.* **2018**, *9*, 1037. [CrossRef]
30. Willis, A.D. Rarefaction, Alpha Diversity, and Statistics. *Front. Microbiol.* **2019**, *10*, 2407. [CrossRef]
31. Chen, J.; Li, H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **2013**, *7*, 418–442. [CrossRef]

32. Susin, A.; Wang, Y.; Le Cao, K.A.; Calle, M.L. Variable selection in microbiome compositional data analysis. *NAR Genom. Bioinform.* **2020**, *2*, lqaa029. [CrossRef] [PubMed]
33. Aitchison, J. Reducing the dimensionality of compositional data sets. *J. Int. Assoc. Math. Geol.* **1984**, *16*, 617–635. [CrossRef]
34. Zou, H.; Hastie, T.; Tibshirani, R. Sparse Principal Component Analysis. *J. Comput. Graph. Stat.* **2006**, *15*, 265–286. [CrossRef]
35. Filzmoser, P.; Hron, K.; Reimann, C. Principal component analysis for compositional data with outliers. *Environ. Off. J. Int. Environ. Soc.* **2009**, *20*, 621–632. [CrossRef]
36. Scealy, J.; De Caritat, P.; Grunsky, E.C.; Tsagris, M.T.; Welsh, A. Robust principal component analysis for power transformed compositional data. *J. Am. Stat. Assoc.* **2015**, *110*, 136–148. [CrossRef]
37. Wang, H.; Shangguan, L.; Guan, R.; Billard, L. Principal component analysis for compositional data vectors. *Comput. Stat.* **2015**, *30*, 1079–1096. [CrossRef]
38. Dai, X.; Müller, H.G. Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Stat.* **2018**, *46*, 3334–3361. [CrossRef]
39. Anderson, M.J.; Willis, T.J. Canonical Analysis of Principal Coordinates: A Useful Method Of Constrained Ordination for Ecology. *Ecology* **2003**, *84*, 511–525. [CrossRef]
40. Verma, S.P. Multidimensional Techniques for Compositional Data Analysis. In *Road from Geochemistry to Geochemometrics*; Springer: Singapore, 2020; pp. 441–479. [CrossRef]
41. Lin, W.; Shi, P.; Feng, R.; Li, H. Variable selection in regression with compositional covariates. *Biometrika* **2014**, *101*, 785–797. [CrossRef]
42. Cao, Q.; Sun, X.; Rajesh, K.; Chalasani, N.; Gelow, K.; Katz, B.; Shah, V.H.; Sanyal, A.J.; Smirnova, E. Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. *Front. Microbiol.* **2021**, *11*, 3203. [CrossRef]
43. Greenacre, M. Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Appl. Comput. Geosci.* **2020**, *5*, 100017. [CrossRef]
44. Bomar, L.; Maltz, M.; Colston, S.; Graf, J. Directed Culturing of Microorganisms Using Metatranscriptomics. *mBio* **2011**, *2*, e00012-11. [CrossRef] [PubMed]
45. Cong, X.; Judge, M.; Xu, W.; Diallo, A.; Janton, S.; Brownell, E.A.; Maas, K.; Graf, J. Influence of Infant Feeding Type on Gut Microbiome Development in Hospitalized Preterm Infants. *Nurs. Res.* **2017**, *66*, 123–133. [CrossRef] [PubMed]
46. Cole, J.R.; Wang, Q.; Fish, J.A.; Chai, B.; McGarrell, D.M.; Sun, Y.; Brown, C.T.; Porras-Alfaro, A.; Kuske, C.R.; Tiedje, J.M. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **2014**, *42*, D633–D642. [CrossRef]
47. Whittaker, R.H. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* **1960**, *30*, 279–338. [CrossRef]
48. Whittaker, R.H. Evolution And Measurement of Species Diversity. *Taxon* **1972**, *21*, 213–251. [CrossRef]
49. Goodrich, J.K.; Di Rienzi, S.C.; Poole, A.C.; Koren, O.; Walters, W.A.; Caporas, J.G.; Knight, R.; Ley, R.E. Conducting a microbiome study. *Cell* **2014**, *158*, 250–262. [CrossRef]
50. Renyi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
51. Hill, M.O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **1973**, *54*, 427–432. [CrossRef]
52. Jost, L. Entropy and diversity. *Oikos* **2006**, *113*, 363–375. [CrossRef]
53. Gotelli, N.; Chao, A. Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data. In *Encyclopedia of Biodiversity*; Academic Press: Cambridge, MA, USA, 2013; pp. 195–211. [CrossRef]
54. Rajaram, R.; Castellani, B. An entropy based measure for comparing distributions of complexity. *Phys. A Stat. Mech. Its Appl.* **2016**, *453*, 35–43. [CrossRef]
55. Evans, S.N.; Matsen, F.A. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser.* **2012**, *74*, 569–592. [CrossRef] [PubMed]
56. Pavoine, S.; Dufour, A.B.; Chessel, D. From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis. *J. Theor. Biol.* **2004**, *228*, 523–537. [CrossRef] [PubMed]
57. Chen, J.; Bittinger, K.; Charlson, E.S.; Hoffmann, C.; Lewis, J.; Wu, G.D.; Collman, R.G.; Bushman, F.D.; Li, H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **2012**, *28*, 2106–2113. [CrossRef]
58. Hervas, J.A.; Ballesteros, F.; Alomar, A.; Gil, J.; Benedi, V.J.; Alberti, S. Increase of Enterobacter in neonatal sepsis: A twenty-two-year study. *Pediatr. Infect. Dis. J.* **2001**, *20*, 134–140. [CrossRef] [PubMed]
59. Gupta, A. Hospital-acquired infections in the neonatal intensive care unit-Klebsiella pneumoniae. *Semin. Perinatol.* **2002**, *26*, 340–345. [CrossRef]
60. Sood, S.K.; Mulvihill, D.; Daum, R.S. Intrarenal abscess caused by Klebsiella pneumoniae in a neonate: Modern management and diagnosis. *Am. J. Perinatol.* **1989**, *6*, 367–370. [CrossRef] [PubMed]
61. Basu, S.; Mukherjee, K.K.; Poddar, B.; Goraya, J.S.; Chawla, K.; Parmar, V.R. An Unusual Case of Neonatal Brain Abscess Following Klebsiella pneumoniae Septicemia. *Infection* **2001**, *29*, 283–285. [CrossRef]
62. Podschun, R.; Acktun, H.; Okpara, J.; Linderkamp, O.; Ullmann, U.; Borneff-Lipp, M. Isolation of Klebsiella planticola from newborns in a neonatal ward. *J. Clin. Microbiol.* **1998**, *36*, 2331–2332. [CrossRef]
63. Westbrook, G.L.; O'Hara, C.M.; Roman, S.B.; Miller, J.M. Incidence and identification of Klebsiella planticola in clinical isolates with emphasis on newborns. *J. Clin. Microbiol.* **2000**, *38*, 1495–1497. [CrossRef]

64. Greenacre, M. Variable selection in compositional data analysis using pairwise logratios. *Math. Geosci.* **2019**, *51*, 649–682. [CrossRef]
65. Noguera-Julian, M.; Rocafort, M.; Guillén, Y.; Rivera, J.; Casadellà, M.; Nowak, P.; Hildebrand, F.; Zeller, G.; Parera, M.; Bellido, R.; et al. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **2016**, *5*, 135–146. [CrossRef] [PubMed]