Submitted to ISMB 2024, , pp. 1–8

doi:

Multidimensional scaling informed by F-ratio: Visualizing microbiome for inference

Hyungseok Kim,[1,2,*,†] Soobin Kim,[3,*] Megan M. Morris,[4] Jeffrey A. Kimbrel,[4] Xavier Mayali[4] and Cullen R. Buie[1,‡]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, [2]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA, [3]Department of Statistics, University of California, Davis, Davis, CA, USA and [4]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA

[*]Equal contribution.

[†]Present address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[‡]Corresponding author. crb@mit.edu

FOR PUBLISHER ONLY Received on 25 January 2024

Abstract

Multidimensional scaling (MDS) is a dimension reduction technique that preserves pairwise distances between observations and has commonly been used for displaying multivariate data including microbiome. Supervised learning-based MDS alterations have recently been proposed by incorporating labels for training, but their configurations depend on the choice of a hyperparameter, making them susceptible to false positives. In this study, we present a weakly supervised MDS approach informed by F-ratio, the dispersion of observations that share a common binary label. We compare our method with existing dimension reduction approaches and show that our visualization accurately represents the F-ratio while consistently preserving the global structure. Using an algal-associated and human gut microbiome datasets, we show that our new method better illustrates the community's response to the host presence or diseases, suggesting its potential impact on microbiology and ecology data analysis.

Key words: Multidimensional scaling, Microbiome, F-statistic, Dimension reduction

Introduction

Recent advances in microbiome research have empowered multivariate analytics, keeping pace with the growing size and dimensionality of biological data. Among diverse multivariate datasets, microbiome data are distinguished by their sparse and compositional structure [16]. Moreover, the analysis of microbiome data involves incorporating a context derived from the taxonomy of each microbial component based on its gene information [22, 25, 3].

There are two representative statistical approaches in microbiome analysis. First, the microbiome data is visualized by reducing its structure to a lower dimensional space, a process known as dimensionality reduction. The most conventional approach is multidimensional scaling (MDS), where a configuration is sought in a way that preserves every pairwise distance between the community samples. Because the microbiome samples are structurally distinguished among multidimensional data (e.g., highly skewed, zero-inflated [15]), the simplicity of MDS has been preferred over the other methods in dimensionality reduction methods since its inception [12, 3]. On the other hand, recent methods such as UMAP [26], have been used to visualize a high dimensional microbiome dataset by effectively clustering its features associated with their sampling site within the host [2].

Another approach in the microbiome analysis is to infer its significance using hypothesis testing or statistical regression. Commonly used inference procedures in the field compare multidimensional structures by incorporating data labels, treatments, or explanatory variables. For example, one way to test for the significance of treatments is by comparing structural dispersion calculated across and within sample groups, such as the F-ratio. In cases where the distribution of the structure is unknown, empirical hypothesis testing is performed using nonparametric statistics [17, 15, 9, 1]. Specifically, if the data can be assumed to be independent, the statistic can be obtained by permuting the data labels [17].

One of the distinctions between dimension reduction and statistical inference in microbiome analysis lies in the utilization of label information. Most dimension reduction methods, such as principal component analysis (PCA), MDS, UMAP [26], t-SNE [35], Isomap [34], do not necessarily require the data labels in principle. While recent unsupervised visualization approaches have differentiated microbiome data groups by clustering [2], there still remains a precaution in choosing an appropriate visualization tool. This is because an inappropriate dimension reduction can lead to biased results or false positives [24], considering that designing a microbiome study is challenging due to a stochastic and founder effect [27]. For example, if sample groups possess a significantly different structure which is not captured through a conventional unsupervised dimension reduction, e.g., samples with a large dispersion, then visualization results may fail to provide an insightful interpretation compared to performing the statistical test.

Indeed, employing labels for visualizing multidimensional data has been accepted in other domain applications. The motivation underlies by revising the classical approach, such as classical MDS, by including external information that is conferred by group labels [14]. Broadly termed as the confirmatory MDS [4], the approach includes applying an external constraint to the MDS, thereby providing a more contextual illustration of the data structure. These altered MDS configuration are generally accepted up to a point where they do not deviate heavily from the original configuration, as classical MDS can also produce different configurations [5]. In the confirmatory MDS methods, an objective function is constructed by adding a label-informed confirmatory term to the MDS objective with a hyperparameter. These approaches have been useful in differentiating multivariate data by groups in lower dimensions [10, 37]. However, the configurations can highly be sensitive to the choice of the hyperparameter. For example, setting an excessive weight on label information may lead to an unbalanced configuration, a phenomenon commonly observed in existing dimension reduction methods. To prevent such potentially misleading outcomes, such as false positives, the hyperparameter is empirically determined through iteration, while this iterative process can be time-limiting, restricting its broader application in biological analysis.

In this work, we introduce a weakly supervised version of MDS that is informed by data labels through F-statistic, which we refer to as FMDS. In other words, FMDS is a visualization method that configures multivariate data structures through their statistical significance. Under a binary class setting, the purpose of FMDS is to explain a difference between groups (even if it is small), as an adjustment of the MDS configuration. The motivation also sets itself apart from other confirmatory approaches in that FMDS does not directly aim to discriminate between groups. Instead, it focuses on creating configurations that separate groups based on the extent of their true differences. By characterizing the behavior of the proposed framework we show that FMDS configuration is less dependent on the choice of the model hyperparameter, avoiding the false positiveness.

Related Work

Supervised multidimensional scaling

Consider a balanced design where the number of total observations is N, and each observation $\mathbf{x}_i$ is S-dimensional, pertaining to a binary label $y_i \in \{0,1\}$ for every $i = 1, \cdots N$. Using the set of observations $\mathbf{X} = (\mathbf{x}_1, \cdots \mathbf{x}_N)$, a pairwise distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be obtained using a metric, e.g., Euclidean, Bray-Curtis [7], UniFrac [22]. Let us define $\mathbf{D} = [d_{ij}] \in [0,\infty)^{N\times N}$ the matrix obtained from the pairwise distance.

In metric multidimensional scaling, a lower-dimensional configuration $\mathbf{Z} = (\mathbf{z}_1, \cdots \mathbf{z}_N) \in \mathbb{R}^{N\times 2}$ is sought in a way that preserves the pairwise distance while the dimension is reduced. This is enabled by minimizing the "raw stress" [6]. The metric MDS is unsupervised learning and it does not require a set of labels y_i for training.

In Supervised multidimensional scaling (SMDS), on the other hand, imposes an additional constraint on the configuration based on class labels [36], with its purpose to discriminate observations by the labels as well as to carry out the classical scaling. Witten et al. [36] proposed an objective function by adding a newly designed confirmatory term to the raw stress,

$$O(\mathbf{Z}) = \underbrace{\frac{1}{2}\sum_{i,j}(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2}_{\text{raw stress}} + \underbrace{\lambda \sum_{i,j:y_j>y_i}(y_j - y_i)\sum_{k=1}^{2}\left(\frac{d_{ij}}{\sqrt{2}} - (z_{jk} - z_{ik})\right)^2}_{\text{confirmatory term}}, \tag{1}$$

where the confirmatory term involves the labels set $[y_i]$. Minimizing the objective $O(\mathbf{Z})$ locates the configuration points closer to each other when they are within the same label group. Additionally, a hyperparameter λ balances the confirmatory term with the stress and controls the degree of classification. Selection of the hyperparameter is carefully driven by the data structure, so that the process avoids spurious group distinctions during the visualization.

Non-parametric multivariate hypothesis testing

Hypothesis testing for a group difference in multivariate data is commonly performed by calculating the F-statistic, a ratio comparing inter- and intra-group variances based on pairwise distances. However, a standard F-test requires an assumption that observations are normally distributed with the same variance. Because the assumption is not usually met for biological data, an alternative approach is proposed to derive a "pseudo" F-ratio [9], which is by permuting group labels to create an empirical distribution [1, 17]. To be specific, the pseudo F-ratio is defined as

$$F = \frac{\sum_{i,j} d_{ij}^2 - 2\sum_{i,j}\mathbb{I}\{y_i = y_j\}d_{ij}^2}{2\sum_{i,j}\mathbb{I}\{y_i = y_j\}d_{ij}^2} \cdot (N-2), \tag{2}$$

where $\mathbb{I}\{\cdot\}$ denotes an indicator function. While the pseudo F does not necessarily follow the F-distribution, an empirical distribution can be constructed instead by permuting the labels for a large enough size of dataset [15, 1]. In other words, by denoting F^{π_k} as a new F-ratio that is obtained from a permuted labels set $[y_i]^{\pi_k}$, p-value is obtained by repeating the permutation π_k for $k = 1, \cdots, K$:

$$p = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\{F^{\pi_k} \geq F\}. \tag{3}$$

The procedure is known as the permutational multivariate analysis of variance (PERMANOVA) [1].

Materials and methods

Our goal is to incorporate the multivariate hypothesis testing result for determining an MDS configuration of the dataset. To achieve the goal, we take a weakly supervised approach with the following two steps. First, a two-dimensional configuration is trained by performing an unweighted metric MDS (i.e., unsupervised) to initialize the configuration. Next, the configuration points are adjusted in a way that best represents a p-value calculated from the multivariate, binary-labeled dataset. This is enabled by adding a confirmatory term to the raw stress, a similar approach proposed by the supervised MDS [36],

$$O(\mathbf{Z}) = \underbrace{\sum_{i,j}(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2}_{\text{raw stress}} + \lambda \cdot \underbrace{\arg\min_{\mathbf{z}} |F_{\mathbf{z}} - f_{\mathbf{z}}(F)|}_{\text{confirmatory term}}, \quad (4)$$

in which the confirmatory term is designed to minimize a difference in p-value based on pseudo F's calculated under the original S-dimension and the two-dimension, referred to as F and $F_{\mathbf{z}}$, respectively. Again, each p-value is obtained based on the empirical, pseudo F-distribution via permutation [1]. Noting that a scale of these statistics can vary by the dimension and the hyperparameter, we introduce a regression function $f_{\mathbf{z}}(F) : \mathbb{R} \to \mathbb{R}$ that maps F-ratio from S-dimension onto two-dimension (Supplementary Note 1).

In detail, the regression is carried out by iterating permutations of the label set $[y_i]$ for each dimension to obtain two pseudo F's independently:

$$F^{\pi_1} = \left(\frac{\sum_{i,j} d_{ij}^2}{2\sum_{i,j} d_{ij}^2\, \mathbb{1}\{y_i^{\pi_1} = y_j^{\pi_1}\}} - 1\right) \cdot (N-2), \quad (5)$$

$$F_{\mathbf{z}}^{\pi_2} = \left(\frac{\sum_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2}{2\sum_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2\, \mathbb{1}\{y_i^{\pi_1} = y_j^{\pi_2}\}} - 1\right) \cdot (N-2), \quad (6)$$

with the superscript π_j denoting the permutation ($j = 1, 2$). Note that Equation 6 represents a pseudo F on a two-dimensional configuration $\mathbf{z}$. Using the sequence of pair $(F^{\pi_1}, F_{\mathbf{z}}^{\pi_2})$, the mapping function $f_{\mathbf{z}}$ is derived by a local regression.

In practice, the confirmatory term for our FMDS objective can be derived by seeking a configuration $\mathbf{Z}$:

$$\arg\min_{\mathbf{Z}} |F_{\mathbf{z}} - f_{\mathbf{z}}(F)| \quad (7)$$

$$\approx \arg\min_{\mathbf{Z}} \left| \sum_{i,j} \left[1 - 2\mathbb{1}\{y_i = y_j\}\left(1 + \frac{f_{\mathbf{z}}(F)}{N-2}\right)\right] \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \right|, \quad (8)$$

with a detailed derivation provided in Supplementary Note 1. Algorithm 1 describes a pseudocode that derives the mapping function $f_{\mathbf{Z}}$.

Algorithm 1 Mapping from F to $F_{\mathbf{z}}$ with random permutation.

```
          $\mathbf{Z} \in \mathbb{R}^{N\times 2}$         : configuration
Require:  $\mathbf{D} \in \mathbb{R}_+^{N\times N}$       : pairwise distance
          $\mathbf{y} = \{0,1\}^N$       : labels
Ensure: $f_{\mathbf{z}} : \mathbb{R} \to \mathbb{R}$   : mapping function
  $\mathcal{L} \Leftarrow$ empty list
  $\mathcal{L}_{\mathbf{z}} \Leftarrow$ empty list
  for $1 \le i \le 999$ do
      $\mathbf{y}^{\pi_1} \Leftarrow$ random permutation on $\mathbf{y}$
      $\mathbf{y}^{\pi_2} \Leftarrow$ random permutation on $\mathbf{y}$
      $F^{\pi} \Leftarrow$ $F$-ratio using $\mathbf{D}$ and $\mathbf{y}^{\pi_1}$ (Equation S1)
      $F_{\mathbf{z}}^{\pi} \Leftarrow$ $F$-ratio using $\mathbf{Z}$ and $\mathbf{y}^{\pi_2}$ (Equation S2)
      $\mathcal{L} \Leftarrow$ append $F^{\pi}$ to $\mathcal{L}$
      $\mathcal{L}_{\mathbf{z}} \Leftarrow$ append $F_{\mathbf{z}}^{\pi}$ to $\mathcal{L}_{\mathbf{z}}$
  end for
  $\mathcal{L} \Leftarrow$ sort $\mathcal{L}$ by an increasing order
  $\mathcal{L}_{\mathbf{z}} \Leftarrow$ sort $\mathcal{L}_{\mathbf{z}}$ by an increasing order
  $f_{\mathbf{z}} \Leftarrow$ local regression from $\mathcal{L}$ to $\mathcal{L}_{\mathbf{z}}$
```

Majorize-Minimization (MM) algorithm

Minimizing raw stress to perform MDS has been achieved by several iteration methods in the past few decades, part of which have originated in the field of graph drawing [18]. Because the stress is a non-convex function in terms of $\mathbf{Z}$, there is no global minimum and it may require a different initialization [12, 39]. This has also led with proposing several optimization methods including 2D Newton–Raphson [18], stochastic gradient descent [39, 21], divide-and-conquer [38, 31], and majorization [11].

In our problem setting, we implemented the majorization (or Majorize-Minimization) approach to minimize the FMDS objective (Equation 4) as similarly described by [6] and [36]. In each step of majorization, an optimal configuration point

$$\mathbf{z}_k^* = \arg\min_{\mathbf{z}_k} O(\mathbf{Z}). \quad (9)$$

is sought for every $k = 1, \cdots N$ while other points except for $\mathbf{z}_k$ are fixed. Majorizing with Equation 4 results in a quadratic expression in terms of $\mathbf{z}_k$, which can be locally minimized by taking a derivative with respect to z_{ks} and setting to zero, given that the expression is convex with assumptions. Detailed procedure is described in Supplementary Note 2 and the resulting update rule is given in Algorithm 2.

Algorithm 2 Majorize-Minimization for FMDS.

```
          $\lambda \in \mathbb{R}_+$           : hyperparameter
Require:  $\mathbf{D} \in \mathbb{R}_+^{N\times N}$       : pairwise distance
          $\mathbf{y} = \{0,1\}^N$       : labels
Ensure: $\mathbf{Z} \in \mathbb{R}^{N\times 2}$     : configuration
  $F \Leftarrow$ $F$-ratio using $\mathbf{D}, \mathbf{y}$ (Equation 2)
  for $1 \le t \le T$ do
      for $1 \le k \le N$ do
          $\delta(\mathbf{Z}) \Leftarrow$ sign of confirmatory term (Equation S13)
          $f_{\mathbf{z}}(F) \Leftarrow$ mapping function (Algorithm 1)
          $\mathbf{z}_k \Leftarrow \mathbf{z}_k$ using $\delta(\mathbf{z})$ and $f_{\mathbf{z}}(F)$ (Equation S17)
      end for
  end for
```

Results

In this section, we first characterize the behavior of FMDS across a range of hyperparameters and compare its performance to benchmark dimension reduction methods, including MDS, supervised MDS [36], UMAP [26], t-SNE [35], Isomap [34] and a self-supervised neural network such as SimCLR [8]. For the evaluation, we consider two types of datasets, simulated and experimental, both of which served as examples where the two-dimensional configurations from classical MDS were not consistent with the hypothesis testing results. Next, we demonstrate how FMDS can improve visualizing biological samples in a way that multivariate F-test results are addressed.

Evaluating performance of F-informed MDS

Inspired by the evaluation metric proposed by [33], we provide the following numerical procedure to evaluate the performance of each dimension reduction method. First, calculate the permuted F-ratio, using the original data while permuting the labels. Second, calculate the permuted F-ratio with the configurations. Third, compute the correlation between the two sets of F-ratios, namely F-correlation (F-cor). If the configurations show similar dispersion patterns to the original data with permuted labels, then the F-correlation would have high values. Finally, calculate p-ratio, the ratio between the permutation-based p-value of the original data and the embedding. The p-ratio being close to 1 implies that the configuration conveys an accurate inference regarding differences in labels and that the dimension reduction method is insensitive to false positives.

Additionally, we quantify a degree of deviation of the configuration $\mathbf{z}$ from the original distances d_{ij}. It is carried out by calculating two metrics; the first is the "normalized" stress or Stress-1 [6],

$$\text{Stress-1} = \frac{\sum_{i,j}(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2}{\sum_{i,j}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}, \qquad (10)$$

and the second using a Shepard diagram [13] and its correlation coefficient (Cor). Both terms measure how much distortion from the original global structure has occurred in the configuration from each reduction method.

Simulated dataset

We first consider a three-dimensional dataset of two balanced groups that follow normal distributions slightly different in means, μ_0, μ_1, but the same covariance matrix D. Such set of observations x_i can be constructed as, for example,

$$x_i \sim \begin{cases} \mathcal{N}(\mu_0, D), & i = 1, 2, \cdots 50 \\ \mathcal{N}(\mu_1, D), & i = 51, 52, \cdots 100, \end{cases} \qquad (11)$$

where $\mu_0 = [0,0,0]^\top$, $\mu_1 = [0,0,1]^\top$, and $D = [[3,0,0],[0,3,0],[0,0,1]]$.

Here the means are different at the third dimension, where the lowest variance was imposed among the principal diagonals of D (Supplementary Figure 1). Therefore, the MDS does not distinguish groups in a two-dimensional configuration, but performing a multivariate hypothesis testing on the datasets, e.g., PERMANOVA [1], indicates there is a statistically significant difference between the groups with $p = 0.005$.

Using the simulated dataset, we measure the normalized stress (Equation 10) and correlation coefficient in the Shepard diagram. Using Algorithm 2 we confirmed the configuration converged over a number of iterations for a range of hyperparameters up to $\lambda = 0.7$, a value that ensured the confirmatory term in Equation 4 becomes negligible compared to the MDS term (Figure 1).

We then compare the performance of FMDS to other dimensionality reductions such as MDS, SMDS, UMAP [26], t-SNE [35], and Isomap [34], which are widely used in visualizing multivariate data. For UMAP, the supervised learning mode is also implemented (UMAP-S), as well as the unsupervised version (UMAP-U). Table 1 presents a summary of performance results using the simulated dataset, showcasing the optimal outcomes across various hyperparameters for each method. The result suggests the proposed FMDS produces a 2D configuration with its distance structure less distorted than other methods (as evaluated by low stress and high correlation), as well as carrying a meaningful statistical inference as measured by a high p-ratio.

In addition to its performance evaluation, Figure 2 shows a detailed characteristic behavior of FMDS compared to supervised MDS. Regardless of a choice of the hyperparameter λ, FMDS visualization exhibits a consistent performance with its stress nearly as 0.2 and correlation coefficient higher than 0.9 (Figure 2a). The behavior is in contrast to supervised MDS [36] which was monotonically dependent to λ, as expected, because SMDS distinguishes groups at the expense of the original distance structure. Overall, the stress obtained from the proposed MDS is consistently lower than those from SMDS. Similarly, as displayed in Figure 2b, Shepard plot shows the proposed FMDS presents a higher correlation of the sample pair distance in between the original and two-dimensional space.

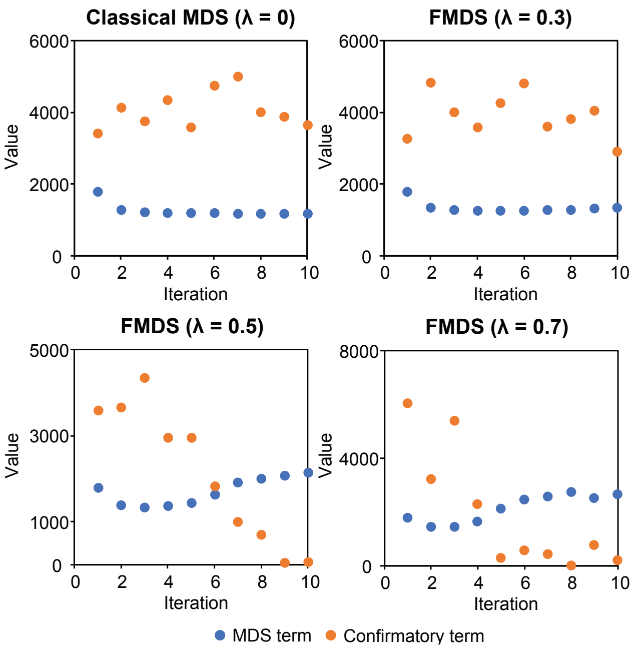

Fig. 1. Comparison of MDS and confirmatory terms by iteration of Algorithm 2 for a range of λ. Each term values were calculated using simulated dataset following Equation 11.

Table 1. Summary of performance of dimension reduction methods using simulated data with four different evaluation metrics, p-ratio, F-correlation, Stress-1, and Shepard diagram correlation coefficient.

Simulated	p-ratio	F-cor	Stress-1	Cor
FMDS	1	0.905	0.176	0.932
MDS	0.064	0.960	0.173	0.959
SMDS	1.002	0.646	0.238	0.875
UMAP-S	1.002	0.271	0.820	0.303
UMAP-U	0.958	0.776	0.550	0.768
t-SNE	0.998	0.778	0.975	0.801
Isomap	0.016	0.958	0.162	0.960

Algal microbiome

We next take an algal microbial community as another dataset to compare the performances of different dimension reductions. The dataset presents a compositional structure expressed by the abundance of 16S rRNA gene of 72 bacterial taxa, samples harvested with proximity to their algal host [20]. As a distance metric, the weighted Unifrac [23] is chosen to obtain the pairwise distance between individual community samples. We consider two datasets (Site 1 and Site 2) where each contains

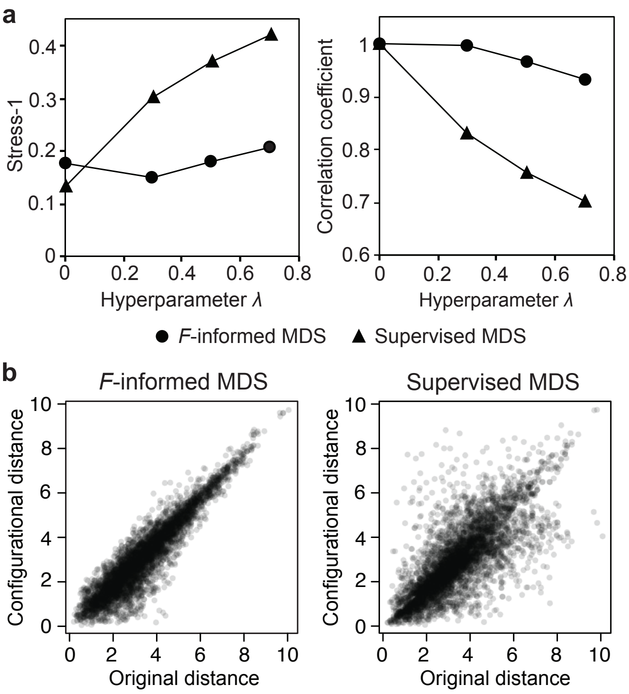

Fig. 2. Evaluation of *F*-informed MDS using simulated data. (a) Performance of FMDS compared to supervised MDS [36] by measuring stress and Pearson correlation coefficient from Shepard plot. (b) Shepard plot of FMDS compared to supervised MDS for a hyperparameter $\lambda = 0.5$. More Shepard plots are provided in Supplementary Figure 2.

thirty-six, balanced community samples with a binary label (e.g., with or without the presence of algal host).

The performance of FMDS is again evaluated by calculating four metrics, which is then compared to existing dimension reduction methods. Table 2 and Table 3 show the summarized performance of FMDS with the dataset. Notably, we observe that the proposed FMDS delivers a two-dimensional configuration with a less distorted distance structure when compared to self-supervised learning neural network (SimCLR, see Supplementary Note 3) or supervised UMAP from in Site 2 algal microbiome. In addition, compared with unsupervised methods including MDS, unsupervised UMAP, t-SNE, and Isomap, our visualization provides a more accurate inference, as represented by the *p*-ratio being close to unity. Overall, our evaluation suggests that FMDS outperforms existing dimension reduction tools by preserving the global structure in regard to the pairwise distances, as well as being less prone to false positives.

Table 2. Summary of dimensionality reduction tool performances using Site 1 algal microbiome dataset

Site 1	*p*-ratio	*F*-cor	Stress-1	Cor
FMDS	1.000	0.648	0.361	0.611
MDS	1.000	0.925	0.346	0.911
SMDS	1.000	0.880	0.215	0.909
UMAP-S	1.000	0.351	0.994	0.286
UMAP-U	1.000	0.791	0.955	0.694
t-SNE	1.000	0.622	1.000	0.551
Isomap	1.000	0.909	0.237	0.862
SimCLR	0.481	0.103	0.990	-0.022

Table 3. Summary of dimensionality reduction tool performances using Site 2 algal microbiome dataset

Site 2	*p*-ratio	*F*-cor	Stress-1	Cor
FMDS	0.989	0.806	0.244	0.857
MDS	0.583	0.903	0.399	0.877
SMDS	1.094	0.596	0.272	0.834
UMAP-S	1.094	0.103	0.990	0.069
UMAP-U	0.138	0.687	0.964	0.479
t-SNE	0.247	0.598	1.000	0.593
Isomap	0.354	0.849	0.275	0.828
SimCLR	1.094	0.057	0.987	0.013

Additionally, the detailed performance of FMDS is characterized by varying the hyperparameter as shown in Figure 3. Again we observe that FMDS configuration produced a low stress that is less dependent on the hyperparameter λ, which even decreases with nonzero λ (0.1, 0.3) than MDS (Figure 3a). Shepard plot and Pearson correlation also show that the configurations nicely preserve the original distance in the microbial community data, except for a case when the largest $\lambda = 0.5$ is applied to Site 1 community dataset (Supplementary Figure 4).

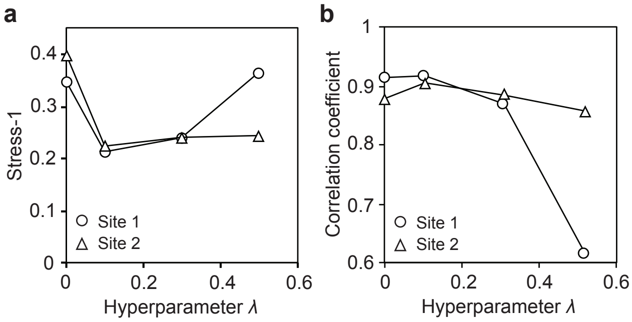

Fig. 3. Evaluation of FMDS using microbial community data, measured by (a) Stress-1 and Pearson correlation from Shepard plot. Shepard plot from each sample site and more results are provided in Supplementary Figures.

Self-supervised learning with SimCLR and PopPhy-CNN.

In addition to the traditional dimension reductions, we also sought to compare our FMDS with existing neural network models that are used for dimensionality reduction. For converting compositional microbial abundance with its phylogenetic information into a matrix, we implemented PopPhy-CNN [32] architecture. In brief, the encoder consists of one Gaussian noise filter, two 2D convolution layers (with kernel size of 5 by 3), and one fully connected layer with 32 output nodes. For the self-supervised learning framework we choose SimCLR [8], where the data augmentation is performed by applying random brightness and contrast filter. In a pretraining step of SimCLR, a model is constructed by compiling encoder, projection head (two dense layers each of 32 output nodes), and one dense layer (10 output nodes). Site 1 and 2 datasets were individually trained and evaluated (30, 6 data each) for the pretraining step, resulting a linear probing accuracy of 53.3% after 50 training epochs. In the following finetuning step, the encoder is added with linear probe, resulting in a validation accuracy higher than 83.3% after 50 epochs. The trained encoder is used to obtain a 32 nodes-sized feature for each microbial community sample in Site 1 and 2. Detailed parameters and steps for training and evaluating the neural network architecture are described in Supplementary Note 3.

Human microbiome dataset

In addition to evaluating FMDS using algal microbiome, we further summarize the performance of FMDS by applying it to broader microbiome datasets of larger sizes. We used two publicly available datasets [28, 32] that were generated using Shotgun Metagenomic sequencing. For our analysis, we specifically selected datasets that included information at the genus level. The first dataset is analyzed from a gut microbiome from 118 healthy subjects and 114 liver cirrhosis patients from a single study [29]. The second microbiome is sampled from a human gut of 223 healthy and 223 patients with type 2 diabetes (T2D) combined by two separate studies [30, 19]. In both cases, the PERMANOVA results indicate a significant difference between healthy subjects and patients at a significance level of 0.05, with respective p-values of 0.006 and 0. As shown in Tables 4, 5, all dimension reductions show a p-ratio close to unity, suggesting these data visualizations effectively convey statistical difference observed in the original dimension. It is notable, however, that FMDS presents the lowest stress among the methods, implying its minimal distortion of the microbiome structure towards its visualization. Indeed, visualization of the FMDS configuration (Figure 4) suggests that the global data structure is more effectively preserved by uniformly distributing the points, as well as preventing a false positive data interpretation. This is particularly evident when compared to UMAP-S, a supervised dimension reduction method.

Table 4. Summary of performance of dimension reduction methods using the cirrhosis dataset.

Cirrhosis	p-ratio	F-cor	Stress-1	Cor
FMDS	0.998	0.707	0.325	0.782
MDS	1.000	0.777	0.988	0.736
UMAP-S	1.006	0.105	0.968	0.133
UMAP-U	0.978	0.563	0.907	0.503
t-SNE	0.990	0.411	0.995	0.559
Isomap	1.004	0.671	0.421	0.658

Table 5. Summary of performance of dimension reduction methods using the T2D dataset.

T2D	p-ratio	F-cor	Stress-1	Cor
FMDS	1.000	0.677	0.332	0.768
MDS	0.998	0.727	1.040	0.716
UMAP-S	1.000	0.156	0.976	0.218
UMAP-U	0.998	0.554	0.917	0.520
t-SNE	1.000	0.579	0.995	0.566
Isomap	0.990	0.638	0.498	0.579

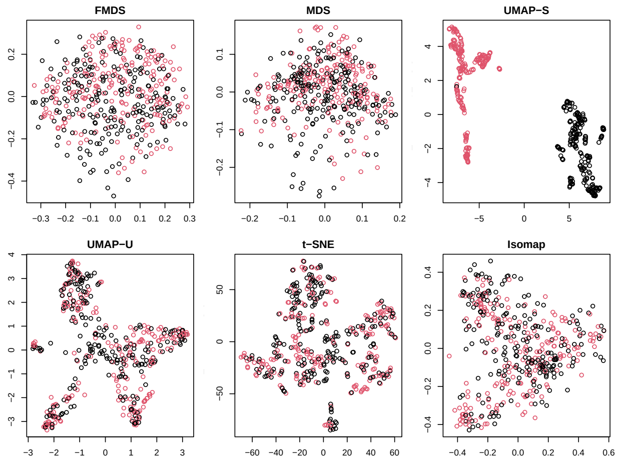

Fig. 4. Visualization of T2D patients and healthy human gut microbiome. The following hyperparameters were used: $\lambda = 0.3$ (FMDS), neighbors $= 5$ (UMAP-S), neighbors $= 30$ (UMAP-U), perplexity$=$ 10 (t-SNE), $K = 7$ (Isomap).

Discriminant analysis with FMDS

We next demonstrate how FMDS can handle two-dimensional configuration by addressing sample group differences. First, we consider the simulated dataset (Section 3.1.1) again where a classical MDS does not distinguish the binary groups in two-dimensional configuration ($p = 0.914$) because the group difference lies in the third dimension which is not identifiable (Figure 5a, Supplementary Figure 1). However, the group difference becomes more visible when FMDS is employed, as shown in Figure 5b. The distinctions are also verified by a low p-value resulting from the PERMANOVA test using the two-dimensional configurations.

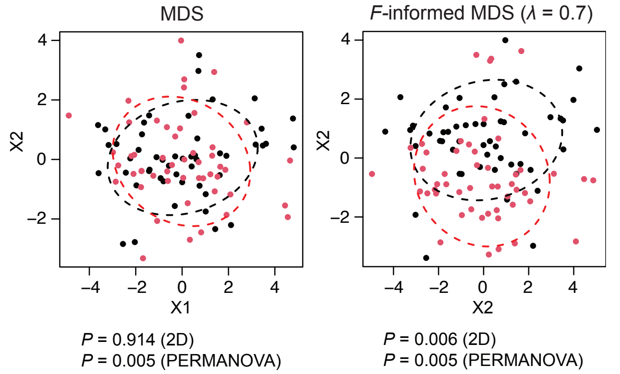

Fig. 5. Two-dimensional visualization of simulated data. Configurations are obtained by training with MDS (left) or FMDS with $\lambda = 0.7$ (right). For each configuration, an error ellipse (68% confidence) and p-value are presented and compared to PERMANOVA result [1].

Algal microbiome

Finally, we demonstrate how FMDS visualization can improve interpreting microbiome structure while addressing F-test results at the same time. Our datasets, Site 1 and 2, present a unique case where a two-dimensional MDS does not fully explain statistical test results on group differences. As shown in Figure 6a, groups in Site 1 are dispersed in a different location whereas Site 2 groups are not, when visualized using MDS. In both sites, however, moderately small p-values are obtained from multivariate hypothesis testing (< 0.1, Table 6), indicating the group difference in the community structure is, in fact, statistically significant.

We then visualize the microbiome data using MDS and FMDS and compare the configurations. Indeed, for Site 1 community samples the configuration retains its distinction between the class labels regardless of the choice of the hyperparameter λ (Figure 6a). Moreover, for Site 2 samples a higher distinction is observed between the groups with higher λ when compared to MDS (Figure 6b, Supplementary Figure 3). The observation with the visualizations is justified by a quantitative measure using p-value calculated on the 2D configurations (Table 6).

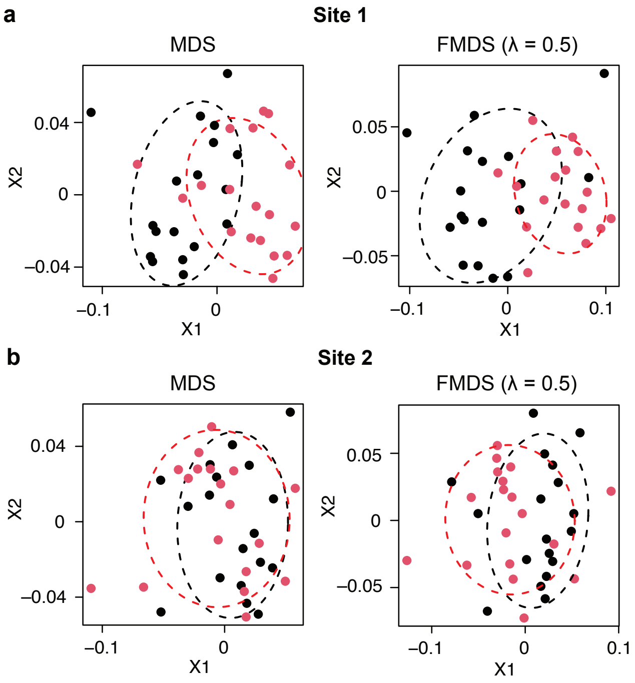

Fig. 6. Visualization of algal microbiome data using FMDS. (a) Classical and (b) proposed FMDS comparing two-dimensional configurations sampled at Site 1 and 2. More configurations with other values of hyperparameters are displayed in Supplementary Figure.

Table 6. Statistical significance on the group difference between two treatments using PERMANOVA test [1]

p-value	Site 1	Site 2
Original data	<0.001	0.093
MDS	<0.001	0.05
FMDS ($\lambda = 0.5$)	<0.001	0.099

Discussion

In this work, a weakly supervised multidimensional scaling is proposed based on the F-ratio and is characterized by comparing the configuration to hypothesis testing results. Using the simulated datasets and ecological samples, our results demonstrate that the proposed FMDS outperforms existing methods for addressing class labels, as evaluated by various measures. However, one limitation of FMDS is its computationally intensive nature, particularly in its algorithmic performance during iterations (see Supplementary Material).

Special attention has been given to its behavior, showcasing less dependency on the choice of hyperparameter to an extent where the algorithm effectively converges. The finding suggests that the new approach mediates the downside of typical confirmatory MDS as a dimensionality reduction tool, in which stress minimization has been underscored by a discriminatory purpose via heavier weight imposed by the hyperparameter choice. In practical terms, users may alleviate the burden of hyperparameter selection through techniques such as cross-validation.

Taken together, the proposed FMDS proves to be useful for visualizing microbiome datasets at a higher consistency delivering hypothesis testing results. The method presents a broader applicability of MDS in the contemporary biological and multivariate data analysis.

Competing interests

No competing interest is declared.

Acknowledgments

The work was supported by the Department of Energy's Genome Sciences Program grant SCW1039. H.K. was partly supported by the Kwanjeong Educational Foundation. S.K. was partly supported by Global Korea Scholarship (GKS). The authors thank the anonymous reviewers for their valuable suggestions. This work is supported in part by funds from the National Science Foundation (NSF: # 1636933 and # 1920920).

References

1. M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001.
2. G. Armstrong, C. Martino, G. Rahman, A. Gonzalez, Y. Vázquez-Baeza, G. Mishne, and R. Knight. Uniform manifold approximation and projection (umap) reveals composite patterns and resolves visualization artifacts in microbiome data. *mSystems*, 6(5):e0069121, 2021.
3. George Armstrong, Gibraan Rahman, Cameron Martino, Daniel McDonald, Antonio Gonzalez, Gal Mishne, and Rob Knight. Applications and comparison of dimensionality reduction methods for microbiome data. *Frontiers in Bioinformatics*, 2, 2022.
4. I. Borg and P. Groenen. *Confirmatory MDS*, pages 181–197. Springer New York, New York, NY, 1997.
5. I. Borg and P. Groenen. *A Majorization Algorithm for Solving MDS*, pages 169–197. Springer New York, New York, NY, 1997.
6. I. Borg and P. Groenen. *MDS Models and Measures of Fit*, pages 37–61. Springer New York, New York, NY, 2005.
7. J. Roger Bray and J. T. Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.
8. T Chen, S Kornblith, M Norouzi, and G Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference*

on Machine Learning, volume 119, pages 1597–1607. Proceedings of Machine Learning Research, 2020.

9. K. R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143, 1993.
10. Trevor F. Cox and Gillian Ferry. Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153, 1993.
11. Jan de Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5(2):163–180, 1988.
12. Erik Demaine, Adam Hesterberg, Frederic Koehler, Jayson Lynch, and John Urschel. Multidimensional scaling: Approximation and complexity, September 01, 2021 2021.
13. E. Dexter, G. Rollwagen-Bollens, and S. M. Bollens. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. *Limnology and Oceanography: Methods*, 16(7):434–443, 2018.
14. C. S. Ding. *Testing Pattern Hypotheses with MDS*, pages 165–173. Springer International Publishing, Cham, 2018.
15. I. Gijbels and M. Omelka. Testing for homogeneity of multivariate dispersions using dissimilarity measures. *Biometrics*, 69(1):137–145, 2013.
16. G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8, 2017.
17. A. P. Holmes, R. C. Blair, J. D. Watson, and I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16(1):7–22, 1996.
18. Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.
19. F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.
20. H. Kim, J. A. Kimbrel, C. A. Vaiana, J. R. Wollard, X. Mayali, and C. R. Buie. Bacterial response to spatial gradients of algal-derived nutrients in a porous microplate. *The ISME Journal*, 16(4):1036–1045, 2022.
21. J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964.
22. C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–35, 2005.
23. C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–85, 2007.
24. Himel Mallick, Siyuan Ma, Eric A. Franzosa, Tommi Vatanen, Xochitl C. Morgan, and Curtis Huttenhower. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*, 18(1):228, 2017.
25. A. P. Martin. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied Environmental Microbiology*, 68(8):3673–82, 2002.
26. L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, February 2018.
27. Robert J Moore and Dragana Stanley. Experimental design considerations in microbiota/inflammation studies. *Clinical & Translational Immunology*, 5(7):e92, 2016.
28. E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLoS Computational Biology*, 12(7):e1004977, 2016.
29. Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, Jiawei Zhou, Shujun Ni, Lin Liu, Nicolas Pons, Jean Michel Batto, Sean P. Kennedy, Pierre Leonard, Chunhui Yuan, Wenchao Ding, Yuanting Chen, Xinjun Hu, Beiwen Zheng, Guirong Qian, Wei Xu, S. Dusko Ehrlich, Shusen Zheng, and Lanjuan Li. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, 2014.
30. Qian Qin, Su Yan, Yang Yang, Jingfeng Chen, Tiantian Li, Xinxin Gao, Hang Yan, Youxiang Wang, Jiao Wang, Shoujun Wang, and Suying Ding. A metagenome-wide association study of the gut microbiome and metabolic syndrome. *Frontiers in Microbiology*, 12, 2021.
31. Taiguo Qu and Zixing Cai. A fast multidimensional scaling algorithm. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2569–2574, 2015.
32. D. Reiman, A. A. Metwally, J. Sun, and Y. Dai. PopPhy-CNN: A phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2993–3001, 2020.
33. J. S. Rhodes, A. Cutler, G. Wolf, and K. R. Moon. Random forest-based diffusion information geometry for supervised visualization and data exploration. In *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pages 331–335, 2021.
34. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–23, 2000.
35. L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
36. D. M. Witten and R. Tibshirani. Supervised multidimensional scaling for visualization, classification, and bipartite ranking. *Computational Statistics & Data Analysis*, 55(1):789–801, 2011.
37. F. Yang, W. Yang, R. Gao, and Q. Liao. Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(3):388–392, 2018.
38. Tynia Yang, Jinze Liu, Leonard McMillan, and Wei Wang. A fast approximation to multidimensional scaling. In *Proceedings of the ECCV Workshop on Computation Intensive Methods for Computer Vision (CIMCV)*, 2006.
39. J. X. Zheng, S. Pawar, and D. F. M. Goodman. Graph drawing by stochastic gradient descent. *IEEE Transactions on Visualization and Computer Graphics*, 25(9):2738–2748, 2019.