TITLE

Multidimensional scaling informed by *F*-ratio for hypothesis testing in microbial community analysis

AUTHORS AND AFFILIATIONS

Hyungseok Kim[*1,2], Soobin Kim[*3], Megan M. Morris[4], Jeff 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

[4] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore CA USA

[*] Equal contribution

SUMMARY

Multidimensional scaling (MDS) is a common dimension reduction method to capture a global pattern and structure in microbial community. While a confirmatory analysis with the classical MDS is a promising approach in broadening an interpretation of the multivariate data (e.g., building a classifier, hypothesis testing), little attention is gained because the configuration can be biased towards the external information. Here we propose a confirmatory MDS informed by an *F*-ratio addressing a community-structural hypothesis testing result under a binary setting. Using a simulated or a 16S rRNA microbial community example, we present how the proposed configuration incorporates the testing result. We evaluate a performance by comparing to the previous confirmatory MDS, demonstrating that our method is less dependent on a selection of a model hyperparameter and minimally alters the classical MDS configuration. Our method proposes a broader applicability of MDS in modern biological and multivariate data analysis.

INTRODUCTION

Biotechnological advances in the past several decades have expanded a size and features of the multivariate data, necessitating a dimensionality reduction as a tool for the interpretation. By extracting an essential information from the biological data, the dimensionality reduction seeks a visual representation of the multivariate in a lower dimensional space. To retain a consistent data structure while performing the dimension reduction, a configuration is sought in a way that preserves dispersion or dissimilarity between samples, a process called as the multidimensional scaling (MDS). Compared to other nonlinear methods in dimensionality reduction [1-3], MDS is known to retain a global structure and represent a long-range interaction between samples, allowing its popularity for a long time since its inception.

In a classical MDS, the configuration is determined in a way that minimizes the difference between the dissimilarity in the original and the low dimensional space, a measure termed the stress. In microbial ecology community where the input data are compositional (i.e., a sample is comprised of different species expressing a level of abundance), the dissimilarity is measured by a difference in expression level of species within samples. For example, when interpreting a microbial community dataset such as 16S rRNA gene expression, a distance metric so-called the Unifrac [4] is used, and it allows to incorporate phylogenic diversity as well as the compositional differences. The choice of an appropriate dissimilarity metric remains an important criterion for processing the input data and represent them in the low dimensional space.

In addition to visualizing the community structure with the dimensionality reduction, a quantitative analysis is carried out using the statistical inference such as hypothesis testing. Because the compositional abundances do not assume a probabilistic distribution *a priori* (highly skewed, zero-inflated [5]), a nonparametric model is preferred over parametric models for constructing a statistic. When the response is assumed to be independent between the samples, the statistic can be readily obtained by permuting the data labels [6]. For testing a difference in the ecological groups based on the sample dispersion, a (pseudo) F statistic has widely been used for each permutation [5, 7].

While hypothesis testing provides a quantitative perspective to understand the multivariate data, it should be noted that the testing result does not account for the MDS configuration. This is because most MDS (e.g., PCA, PCoA) do not consider data labels, whereas the hypothesis testing aims to infer whether dependent variables (e.g., class, label) are influenced by distribution of each sample group. For example, a configuration from the classical MDS is not able to explain a small but statistically meaningful difference between groups of different treatment.

The insufficient explanation by the classical MDS encourages to revise the approach and to address a structural hypothesis by including an external information, conferred by the responses or class [8] labels. Because the stress function is non-convex and an optimization algorithm can produce several local configurations, there is a notion allowing an altered configuration up to a point where its stress does not deviate too much from the classical configuration [9]. Broadly termed as the confirmatory MDS, it imposes an external constraint to the classical MDS carrying over an additional task to minimizing the stress.

In the recent confirmatory MDS methods, an objective function is constructed by adding a confirmatory term to the stress, former of which can be quantified by the labels then multiplied by a hyperparameter. While these confirmatory MDS methods have successfully visualized the multivariate structure in a way that differentiates each sample group with a discriminative purpose [10, 11], choosing a proper hyperparameter remains as a bottleneck towards a broader application of these inventive methods. For example, setting a high hyperparameter results in an undesirable stress and a misleading configuration distorted from the original.

In this study we propose an alternative MDS informed by a hypothesis testing inferences under a binary class setting. Our approach is motivated by a purpose to explain a statistical difference between groups, if any, using the MDS combined with a confirmatory analysis. Because the method does not target to directly discriminate between groups, the motivation distinguishes itself from previously proposed confirmatory MDS. Furthermore, by characterizing the proposed framework we show that the revised configuration is less dependent to the choice of the model hyperparameter, mediating the previous issue with the distortion.

METHODS

Problem formulation

Consider a balanced design where the number of total observations is N, and each observation x_i is S-dimensional, pertaining to a set of labels $y_i \in \{0,1\}$ for every $i = 1 \ldots N$. Based on the observations ($x_1, \ldots x_N$), a distance matrix is obtained as $\mathbf{d} = [d_{ij}] \in \mathrm{R}^{N \times N}$. Now given the distance $\mathbf{d}$, we seek a two-dimensional configuration $\mathbf{z} = (z_1, \ldots z_N) \in \mathrm{R}^{N \times 2}$, that best represents the original dimension by the following criteria.

Classical MDS

In classical MDS, a configuration is realized by minimizing the following objective function.

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

In other words, the configuration $\mathbf{z}$ is obtained in a way that tries to preserve a distance between a pair of observations (x_i, x_j) for each $i, j \in N$. Note that Equation (1) does not contain any terms related to y_i, meaning that classical MDS does not consider the class labels.

Confirmatory MDS for F-informed hypothesis testing

Hypothesis testing for non-parametric multivariate analysis of variance

When testing a statistical difference between groups in multivariate analysis, the group variance is the measure of interest. It is represented by the F-statistic, a ratio between two group variances, each respectively derived from across- and within-group. While the conventional *F*-test requires an assumption that each observation follows a normal distribution, the convention is generalized by introducing an analogous statistic (pseudo *F*-ratio) which is combined with label permutation for quantifying a statistical significance. In this non-parametric approach [7], the pseudo *F*-ratio is defined as

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

where $1\{\cdot\}$ denotes an indicator function. Since the pseudo *F*-ratio does not follow an *F*-distribution under the relaxed model assumption, it is instead evaluated by an empirical distribution that is created by ‘permuting’ the labels. That is, in every permutation a new *F*-ratio, F^{π}, is obtained from the data structure, and by repeating this, we have a *P*-value written as

$$P = \frac{\text{Number of case where } (F^{\Pi} \geq F)}{\text{Number of total repeat}}. \qquad (3)$$

Known as the permutational multivariate analysis of variance (PERMANOVA), the hypothesis testing method has a broad application in microbial community analysis.

Proposed MDS

Now we propose a new multidimensional scaling that incorporates a hypothesis testing result in the multivariate setting. This is enabled by adding a confirmatory term to the classical MDS (Eq. (1)), giving an objective function as

$$O(\mathbf{z}) = \underbrace{\frac{1}{2}\sum_{i,j}(d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2}_{\text{MDS term}} + \lambda \cdot \underbrace{\frac{1}{2}\left|\sum_{i,j}[1 - (f_{\mathbf{z}}(\Phi_o) + 1)\mathbb{1}\{y_i = y_j\}]\|\mathbf{z}_i - \mathbf{z}_j\|_2^2\right|}_{\text{Confirmatory term}} \qquad (4)$$

where Φ_o is a ratio constant expressed in terms of the distance **d** and labels y_i and is defined as

$$\Phi_o(\mathbf{d}, y) := \frac{\sum_{i,j} \mathbb{1}\{y_i \neq y_j\} d_{ij}^2}{\sum_{i,j} \mathbb{1}\{y_i = y_j\} d_{ij}^2},$$

and $f_\mathbf{z}$ (Φ_o) : R → R is a mapping function which is determined by the configuration **z**. An exact derivation of $f_\mathbf{z}$ and a detailed description on the confirmatory term in Equation (4) is described in Appendix B. Given Equation (4), we want to find an optimal configuration $\mathbf{z}^*$ such that $\mathbf{z}^* = \text{argmin}_{(z1,\cdots,zN)}\, O(\mathbf{z})$.

Algorithm for the proposed MDS

Because the confirmatory term in (4) is strictly convex in terms of **z**, we are able to minimize *O*(**z**) by using the Majorize-Minimization (MM) algorithm, a typical approach in the MDS optimization task [12]. While the implementation of MM algorithm is described in detail in Appendix C, we provide its update rule as below.

Algorithm 1 MM algorithm for pseudo F-informed MDS

For epoch t and every $i = 1, \cdots N$,

$$\mathbf{z}_i^{[t+1]} \leftarrow \frac{2}{2(N-1) + \lambda\delta(\mathbf{z}_i^{[t]})(N - (N-2)f_\mathbf{z}(\Phi))} \times \left[(1 + \lambda\delta(\mathbf{z}_i^{[t]})) \sum_{\substack{j=1 \\ \epsilon_{ij}=0}}^{N} \mathbf{z}_j^{[t]} + (1 - \lambda f_\mathbf{z}(\Phi)\delta(\mathbf{z}^{[t]})) \sum_{\substack{j=1 \\ \epsilon_{ij}=1}}^{N} \mathbf{z}_j^{[t]} + \sum_{j=1}^{N} d_{ij} \frac{\mathbf{z}_i^{[t]} - \mathbf{z}_j^{[t]}}{\|\mathbf{z}_i^{[t]} - \mathbf{z}_j^{[t]}\|_2} \right],$$

where $\epsilon_{ij} = \mathbb{1}\{y_i = y_j\}$, $\delta_i(\mathbf{z}) = \text{sign} \sum_{j=1}^{N} [1 - (f_\mathbf{z}(\Phi) + 1)\epsilon_{ij}] \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$, with an initial value obtained from a classical MDS.

RESULTS AND DISCUSSION

Using Algorithm 1, we sought to determine how well our proposed MDS approach can produce a two-dimensional configuration in simulated and experimental data. We then assess its performance by comparing ours to a recent method in confirmatory MDS [13].

Simulated data

We first provide a representing case where our method can be useful in visualizing multidimensional data.

To do this we consider a binary labeled dataset where each group originates in a different multivariate Gaussian distribution. In a three-dimensional setting, for example, consider a balanced design where an observation expressed as

$$x_i \sim \begin{cases} \mathcal{N}\left([0,0,0]^\top, \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right), & i = 1, 2, \cdots 50 \\ \mathcal{N}\left([0,0,1]^\top, \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right), & i = 51, 52, \cdots 100. \end{cases} \tag{5}$$

As expected, PERMANOVA testing result indicates there is a statistically significant difference between the groups with pseudo $F = 5.402$ and $p = 0.005$. However, a classical MDS does not distinguish groups in two-dimensional configuration (Figure 1a) with $p = 0.914$, because the difference is in the third dimension with the lowest variance among the principal diagonals (Figure 6 in Appendix C).

On the other hand, our MDS configuration is able to display the difference between the groups, and it becomes clearer when a hyperparameter λ is large (Figure 1b,c). The distinctions are also verified by a low p-value resulting from PERMANOVA test using the two-dimensional configurations.

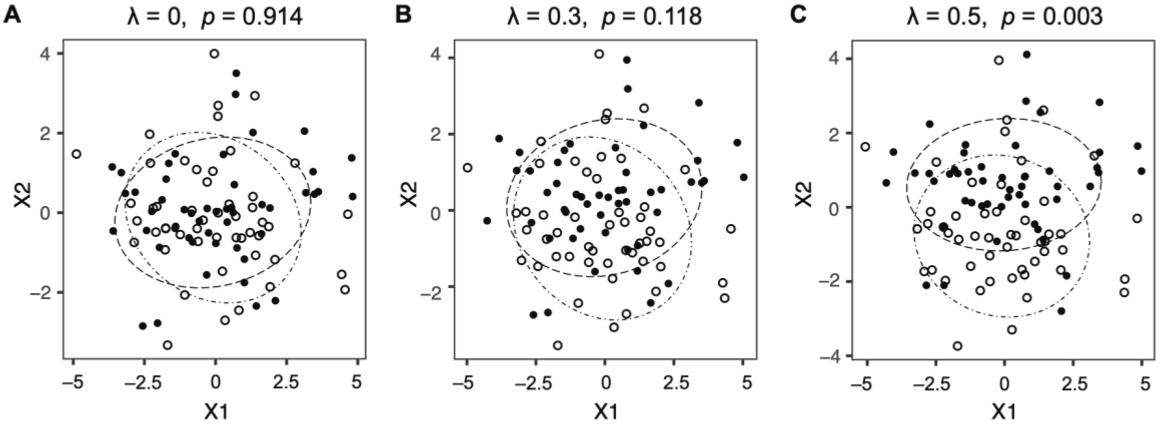


Figure 1: Two-dimensional visualization of proposed MDS with a hyperparameter (a) λ = 0 (classical MDS), (b) λ = 0.3, and (c) λ = 0.5. For each configuration, a p-value is given based on PERMANOVA test. An ellipse is drawn for each group with a confidence interval of 80%.

We next evaluate the performance of our method to existing confirmatory MDS by calculating a stress, which is defined as

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

or by calculating a Spearman correlation from Shepard diagram, as measures of evaluating the performance of MDS [14]. As a result, our proposed MDS visualization presents a Stress ~ 0.2 regardless of a choice of a hyperparameter λ (Figure 2a), suggesting either configuration can be used for visualizing the simulated data [15]. This is in contrast to the previous method [13] where the stress increases as λ becomes large, implying their approach distinguishes groups at the expense of the original distance structure. Similarly, in Shepard plot with a choice of λ, our method presents a more consistent correlation of the sample pair distance in between three- and two-dimension (Figure 2b,2c, and 7 in Appendix A).

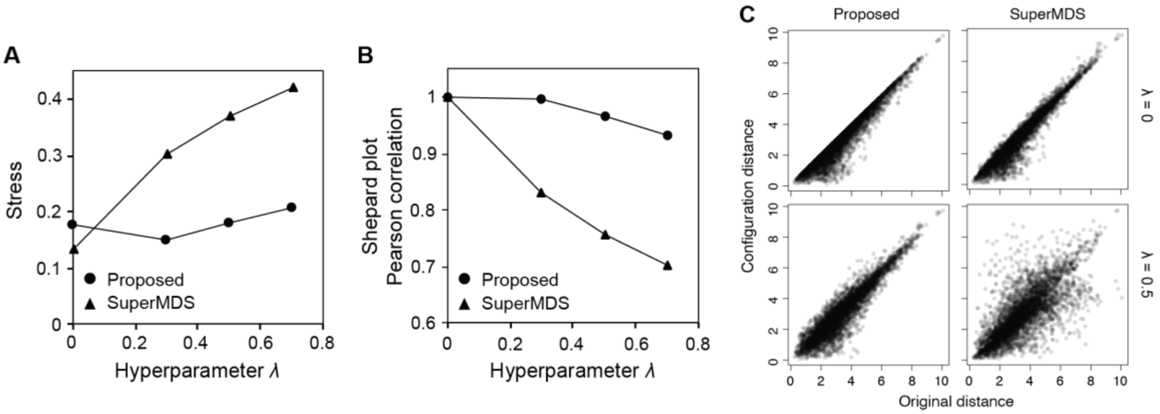


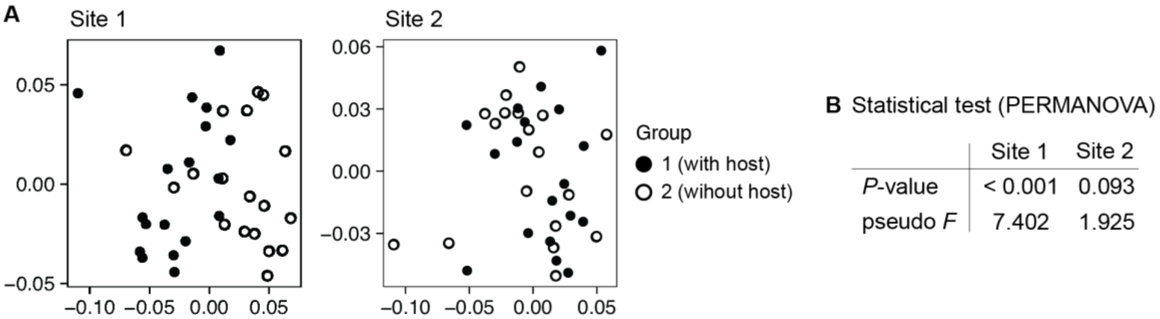
Figure 2: Performance of proposed MDS compared to an existing label-informed confirmatory MDS (SuperMDS [13]) by using (a) Stress and (b) Spearman correlation from the simulated data. (c) Shepard

plot of the proposed MDS comparing to SuperMDS for a hyperparameter λ = 0.5. More results are displayed in Figure 7.

It is worth noting that proposed MDS is invariant by the choice of the hyperparameter, which has not been observed in the existing confirmatory MDS methods.

Microbial community dataset

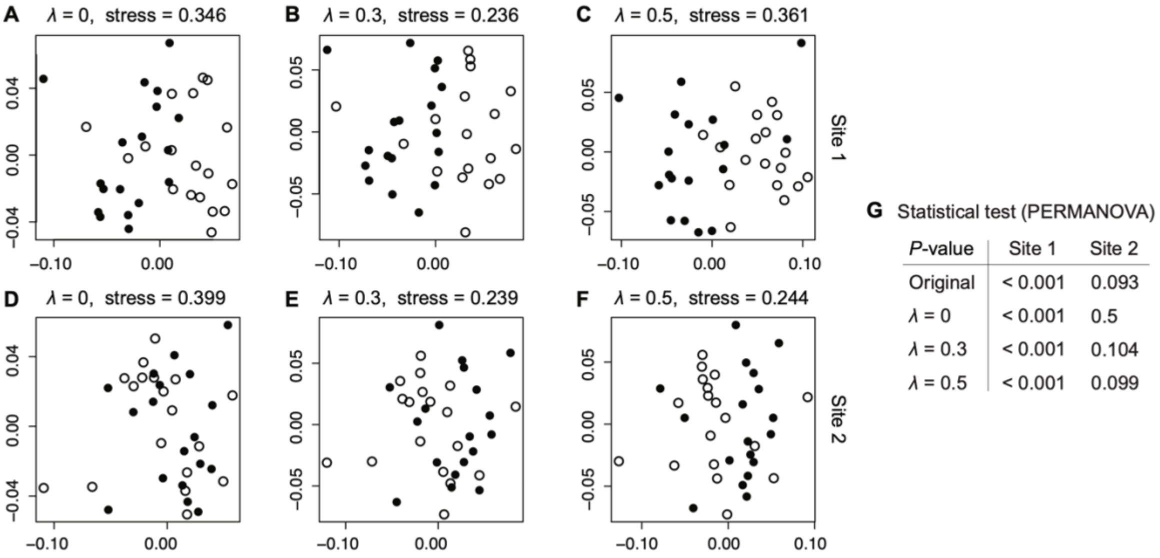
Next, we provide an example where biological hypothesis testing result is conveyed to the MDS configuration. We take microbial community dataset containing thirty-six, balanced samples of a binary label (e.g., with or without a presence of microbial host) [16]. In detail, each data represents expression levels 16S rRNA gene of 72 bacterial taxa, and the distance between samples is measured using the weighted Unifrac [17]. Particular attention is made on these datasets, that the classical two-dimensional MDS configuration does not explain PERMANOVA test results on a group differences. As shown in Figure 3a, groups in site 1 are dispersed in a different location whereas site 2 groups are not, when visualized using the classical MDS. In both sites, however, moderately small P-values are obtained (< 0.1, Figure 3b), indicating the group difference in the community structure is, in fact, statistically significant.



	Site 1	Site 2
P-value	< 0.001	0.093
pseudo *F*	7.402	1.925

Figure 3: (a) Multidimensional scaling and (b) statistical test result on a group difference between two sample groups for each site. The sample presents a microbial community mea- sured by 16S rRNA gene expression.

Using the community dataset, we present a configuration with the proposed MDS visualization. As expected, for site 1 community samples the configuration retains its distinction between the class labels regardless of the choice of the hyperparameter λ (Figure 4a-c). Moreover, for site 2 samples we observe a higher distinction between the groups with increasing λ (Figure 4d-f). The observation with the visualizations is justified by a quantitative measure using P -value calculated on the 2D configurations (Figure 4g).



P-value	Site 1	Site 2
Original	< 0.001	0.093
λ = 0	< 0.001	0.5
λ = 0.3	< 0.001	0.104
λ = 0.5	< 0.001	0.099

Figure 4: Two-dimensional configuration of microbial community samples using the proposed MDS method, where samples are collected from (a-c) site 1 and (d-f) site 2. (g) Statistical significance on the group difference between two treatments using PERMANOVA test.

We then evaluate the performance of the proposed MDS using stress measurement and Shepard plot. Again, we observe stress does not strictly depend on the hyperparameter λ or even show a decreased value when λ is nonzero (e.g., 0.1, 0.3) compared to the classical MDS (λ = 0, Figure 5a). Shepard plot and Spearman correlation also show that the configurations nicely preserve the original distance in the microbial community data, except for a case when the largest λ is set to site 1 community.

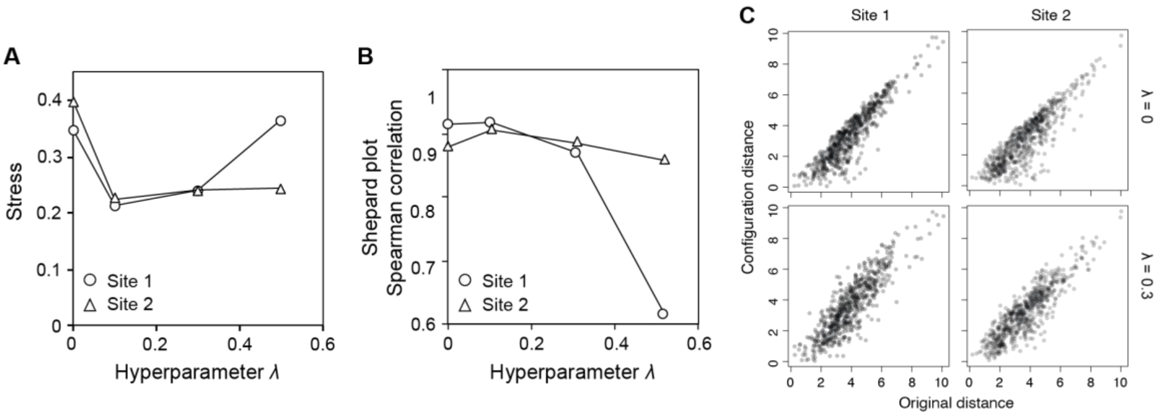


Figure 5: Evaluation of the proposed MDS using microbial community data, measured by (a) stress, (b) Spearman correlation, and (c) Shepard plot from each sample site. More results are displayed in Figure 8 (Appendix A).

CONCLUSION

- A new multidimensional scaling method which incorporates an F-statistic-informed hypothesis testing is proposed.
- We find that the performance of the proposed MDS excels existing MDS methods for addressing class labels, as evaluated by its stress and Shepard plot, validated using both simulated and real datasets.
- The proposed MDS can be useful when analyzing a biological dataset with F -informed hypothesis testing, providing informative and precise dimension reduction, especially for visualization.
- The method is less dependent on the choice of hyperparameter when producing the configuration. This lessens the risk of overreliance on the class labels in that the data are automatically grouped to a suitable degree. Also, users may avoid the hassle of hyperparameter selection using such as cross-validation.

REFERENCES

1. Tenenbaum JB, Silva Vd, Langford JC: **A Global Geometric Framework for Nonlinear Dimensionality Reduction**. *Science* 2000, **290**(5500):2319-2323.

2. McInnes L, Healy J, Melville J: **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. In.; 2018: arXiv:1802.03426.

3. van der Maaten L, Hinton G: **Visualizing Data using t-SNE**. *Journal of Machine Learning Research* 2008, **9**(86):2579-2605.

4. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities**. *Appl Environ Microbiol* 2005, **71**(12):8228-8235.

5. Gijbels I, Omelka M: **Testing for Homogeneity of Multivariate Dispersions Using Dissimilarity Measures**. *Biometrics* 2013, **69**(1):137-145.

6. Holmes AP, Blair RC, Watson JD, Ford I: **Nonparametric analysis of statistic images from functional mapping experiments**. *J Cereb Blood Flow Metab* 1996, **16**(1):7-22.

7. Anderson MJ: **A new method for non-parametric multivariate analysis of variance**. *Austral Ecology* 2001, **26**(1):32-46.

8. Ding CS: **Testing Pattern Hypotheses with MDS**. In: *Fundamentals of Applied Multidimensional Scaling for Educational and Psychological Research.* Edited by Ding CS. Cham: Springer International Publishing; 2018: 165-173.

9. Borg I, Groenen P: **Confirmatory MDS**. In: *Modern Multidimensional Scaling: Theory and Applications.* Edited by Borg I, Groenen P. New York, NY: Springer New York; 1997: 181-197.

10. Cox TF, Ferry G: **Discriminant analysis using non-metric multidimensional scaling**. *Pattern Recognition* 1993, **26**(1):145-153.

11. Yang F, Yang W, Gao R, Liao Q: **Discriminative Multidimensional Scaling for Low-Resolution Face Recognition**. *IEEE Signal Processing Letters* 2018, **25**(3):388-392.

12. Borg I, Groenen P: **A Majorization Algorithm for Solving MDS**. In: *Modern Multidimensional Scaling: Theory and Applications.* Edited by Borg I, Groenen PJF. New York, NY: Springer New York; 1997: 169-197.

13. Witten DM, Tibshirani R: **Supervised multidimensional scaling for visualization, classification, and bipartite ranking**. *Computational Statistics & Data Analysis* 2011, **55**(1):789-801.

14. Dexter E, Rollwagen-Bollens G, Bollens SM: **The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling**. *Limnology and Oceanography: Methods* 2018, **16**(7):434-443.

15. Kruskal JB: **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis**. *Psychometrika* 1964, **29**(1):1-27.

16. Kim H, Kimbrel JA, Vaiana CA, Wollard JR, Mayali X, Buie CR: **Bacterial response to spatial gradients of algal-derived nutrients in a porous microplate**. *The ISME Journal* 2022, **16**(4):1036-1045.

17. Lozupone CA, Hamady M, Kelley ST, Knight R: **Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities**. *Appl Environ Microbiol* 2007, **73**(5):1576-1585.