Shape-based Clustering of Multivariate Longitudinal Data

Alex Braksator, Yiyang Li, Connie Okasaki, Seth D Temple

18 March 2021

Abstract

We develop, implement, and evaluate a procedure for clustering mutivariate longitudinal data according to trajectory shapes. Our method extends a recent shape-respecting k-means algorithm that showed good performance for univariate outcomes. Both methods are based on the Frechét distance between two polygonal curves. Frechét distance generalizes any metric on points to a new metric on curves. We compare multivariate outcomes of similar range with the ℓ_2 norm and suggest a modification to the ℓ_2 norm for known zero inflation. We applied our method to study microbiome samples from transplant patients and identified some interesting cluster profiles. To our knowledge, no other paper has demonstrated longitudinal clustering of multivariate outcomes.

1 Introduction

Clustering techniques provide data-driven insight into subgroups with similar characteristics from within a larger population. With longitudinal data there are usually one or more *a priori* clustering variables which describe a set of known subgroups. Naïve clustering techniques applied to "long" data (i.e. data where each replicate is a separate row) may fail to identify novel subgroups, instead re-identifying known subgroups. Clustering techniques applied to "wide" data (i.e. data where replicates are assigned to separate columns of the same row) face other challenges. For instance, different subgroups may have different numbers of measurements, resulting in dimensionality mismatching between rows. Furthermore, basic clustering algorithms like k-means struggle with high-dimensional data in which Euclidean distances between points become less informative [Napoleon and Pavalakodi, 2011].

One way to address these problems is a "shape-based clustering" pioneered by Genolini et al. [2016]. This technique uses the Frechét distance metric, shown in Figure 1, rather than the Euclidean distance metric. It simplifies the problem of irregular measurements to one of choosing an interpolation strategy, alleviating many of the problems associated with

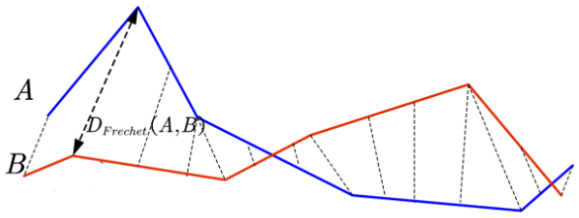

Figure 1: A demonstration of the Frechét distance. The red and blue curves are the trajectories $A(\alpha(t))$ and $B(\beta(t))$. The black lines are the "leash" that separates the two curves. The longest extent of that leash is $D(A, B)$. Image taken from [Guo et al., 2017].

high-dimensional data. The Frechét metric is formally defined as follows: given two curves $A(t) : [0, 1] \to S$ and $B(t) : [0, 1] \to S$ and an underlying distance metric $d(x, y) : S \times S \to \mathbb{R}$, the Frechét distance $D(A, B)$ is given by

$$D(A, B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d(A(\alpha(t)), B(\beta(t)))$$

This distance metric can also be defined to penalize deviations between α and β by defining

$$D(A, B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d\left(\begin{bmatrix}\alpha(t)\\ A(\alpha(t))\end{bmatrix}, \begin{bmatrix}\beta(t)\\ B(\beta(t))\end{bmatrix}\right),$$

in effect treating time as just another variable. Since the scale of these time variables relative to the metric space S may be unclear, we generalize this distance to allow for a scalar transformation of time, i.e.

$$D_\lambda(A, B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d\left(\begin{bmatrix}\lambda\alpha(t)\\ A(\alpha(t))\end{bmatrix}, \begin{bmatrix}\lambda\beta(t)\\ B(\beta(t))\end{bmatrix}\right).$$

This is called generalized Frechét distance [Genolini et al., 2016].

The Frechét distance is non-trivial to calculate. Naïve dynamic programming algorithms achieve runtime complexity $O(mn)$, where m and n are the number of discrete time points in curves A and B [Bringmann et al., 2019]. More sophisticated algorithms can improve runtime, but for large data sets this computation poses a challenge. Genolini et al. [2016] solve this by reducing the size of the data set. They calculate a smaller time series that matches the shape of the original using the Douglas-Peuker algorithm [1973].

2 Methods

The `R` package `kmlShape` from Genolini et al. [2016] offers an existing shape-based clustering algorithm for longitudinal data, but it is limited to univariate longitudinal observations and scales poorly for increasing sample sizes. We extend this approach by generalizing

shape-based clustering to longitudinal data where multivariate outcomes are compositional. Additionally, we achieve computational speedups by keeping constant the number of discrete longitudinal measurements for each cluster mean. Here we focus on how we implement a shape-respecting distance metric and a shape-respecting cluster mean, highlighting the advantages of our method over those of `kmlShape`.

2.1 Discrete Frechét Distance

Like in `kmlShape`, we use the generalized Frechét distance as our shape-respecting distance. We compute the discrete Frechét distance with a standard recursive algorithm [Eiter and Mannila, 1994, Figueira, 2020]. Let ℓ be a norm (e.g. ℓ_2 norm) and ℓ_f be the discrete Frechét distance. Consider multivariate longitudinal data $\mathbf{Y}_1 = (\mathbf{X}_1, \mathbf{t}_1)$ and $\mathbf{Y}_2 = (\mathbf{X}_2, \mathbf{t}_2)$, where $\mathbf{X}_1 \in \mathbb{R}^p \times \mathbb{R}^q$, $\mathbf{t}_1 \in \mathbb{R}^q$, $\mathbf{X}_2 \in \mathbb{R}^p \times \mathbb{R}^r$, and $\mathbf{t}_2 \in \mathbb{R}^r$. This data corresponds to p multivariate longitudinal observations for two subjects with q and r repeated measurements. Times for these measurements $\mathbf{t}_1$ and $\mathbf{t}_2$ may differ and the number of repeated measurements may vary between subjects ($q \neq r$). We denote $\mathbf{X}_{1,i}$ as the p observations for subject 1 at the i^{th} repeated measurement and $\mathbf{t}_{1,i}$ as the time of the measurement. Recursively, we compute:

$$\begin{aligned} d_0 &= \min(d_{(i-1,j)}, d_{(i,j-1)}, d_{(i-1),(j-1)}) \\ d_1 &= \ell((\mathbf{X}_{1,i}, \mathbf{t}_{1,i}), (\mathbf{X}_{2,j}, \mathbf{t}_{2,j})) \\ d_{i,j} &= \max(d_0, d_1) \end{aligned}$$

At each step (i, j) we determine the maximum distance between two polygonal curves $\mathbf{Y}_{1,[1,i]}$ and $\mathbf{Y}_{2,[1,j]}$ assuming we have traversed the trajectory with minimal distance thus far. This recursive algorithm generates a $q \times r$ matrix referred to as the free space diagram. The Frechét distance ℓ_f is the value $d_{q,r}$ from the free space diagram. For our purposes, we compute the discrete Frechét distance between subjects and clusters to assign subjects to the cluster for which they have the smallest discrete Frechét distance.

The free space diagram also offers a proposed trajectory for how two polygonal curves align. We backtrack through the free space diagram to match longitudinal observations between subjects and assigned clusters. For $d_{i,j}$, we move left, up, or diagonal based on the smallest value among $d_{(i-1),j}, d_{i,(j-1)}$, and $d_{(i-1),(j-1)}$. (For ties we move diagonal before left and left before up.) We start at $d_{q,r}$ and conclude at $d_{1,1}$. Below we provide an example of a free space diagram.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 3 & 4 \\ 3 & 2 & 2 & 2 \\ 4 & 2 & 3 & 3 \end{bmatrix}$$

Backtracking we attain the trajectory $[(1,1),(2,1),(3,2),(3,3),(4,4)]$. This aligns the subject's first time step to the cluster's first time step, the subject's second time step to the cluster's first time step, the subject's third time step to the clusters' second time step, and so on. We save these parameterizations in an object-oriented class to make use of when we compute a shape-respecting mean for each cluster.

k-means clustering necessitates computing a cluster mean, but computing a mean for longitudinal data is challenging. Genolini et al. [2016] suggest a shape-respecting cluster mean that is of polynomial order in the sample size. They address this runtime inefficiency with a data reduction step. We instead put forward a shape-respecting mean that leverages the discrete parameterizations from the free space diagram. Let C be a cluster and $c_1, \ldots, c_m$ be the subject assigned to it. For the i^{th} time step of C, we consider the multivariate data at time steps from $c_1, \ldots, c_m$ aligned to the i^{th} time step. We first average within subjects, and then we average between subjects. In doing so the number of discrete time steps remains the same for C. Moreover, our shape-respecting mean is of linear order in the sample size. Given that k-means clustering is known to converge slowly for high-dimensional data, this is an important improvement in runtime efficiency.

2.2 k-means Clustering

Clustering is a non-parametric method that groups similar individual trajectories into subgroups called clusters. k-means is a popular data partitioning related to an expectation-maximization algorithm for Gaussian mixtures. It involves iteratively assigning subjects to groups and computing group means until a convergence is achieved. We can here employ k-means, but with Frechét distances and Frechét means as in (2.1):

```
Data: n individuals, Y_1, ..., Y_n
Result: data partition
/* Initialization */
step <- 0
Random sample without replacement k centers from n individuals
repeat
  /*Assignment phase*/
  step <- step + 1
  for i in 1...n do
    for j in 1...k do
      Compute Frechet distance F_{i,j}
      C_i <- j such that F_{i,j} < F_{i,j'} for all j'!=j
  end
  /*Re-center phase*/
  for j in 1...k do
    Compute Frechet mean for cluster j
  end
until clusters are static
```

2.3 Zero Inflation

In our data analysis, we worked with genus counts from 16S sequencing data for which zero inflation is a known issue. To address this concern, we propose an alternative ℓ_z to the standard Euclidean norm ℓ_2. ℓ_z is the ℓ_2 norm, except if a value from either vector is zero at a coordinate we ignore that coordinate. For instance, let $\nu_1 = (1, 1, 1, 1)$ and $\nu_2 = (2, 2, 0, 2)$.

$$\ell_2(\nu_1, \nu_2) = \sqrt{3(2-1)^2 + 1(0-1)^2}$$
$$\ell_z(\nu_1, \nu_2) = \sqrt{3(2-1)^2 + 0}$$

We refer to ℓ_z as a zero-inflated seminorm. The motivation for this seminorm is to not base the computation on a coordinate value that may be a false zero. Silverman et al. [2020] enumerate zero processes present in 16S sequencing data. Here the assumption is that observed zeros are not true zeros, corresponding to either sampling errors (low sequencing depths) or laboratory technical errors (extracting DNA transcripts using specific materials). This assumption is reasonable for our study because we limited focus to the five most abundant and most prevalent bacteria genuses (4.1).

To study this zero-inflated seminorm, we introduced increasing zero inflation into our real data and compared ℓ_z against ℓ_2. Without introducing zeros, using ℓ_z and ℓ_2 as pointwise pseudometrics in shape-respecting k-means resulted in similar clusters. This is because zero inflation was minimal for genuses that appear at least 3/4 of the time. With probabilities 1/10, 1/4, and 1/2, we switched covariate observations to zeros and renormalized so that relative abundances summed to 1. (We removed rows if all covariates were changed to zero.) For increasing zero inflation, shape-based k-means clustering required more iterations to converge for both ℓ_z and ℓ_2. For probabilities 1/4 and 1/2, ℓ_2-based runs generated one very large cluster and sometimes resulted in clusters with zero assignments. In contrast, ℓ_z-based runs continued to identify multiple clusters of reasonable size. On the other hand, for severe zero inflation, shape-based k-means clustering with ℓ_z may find clusters based on which covariates appear despite zero processes, a clustering that may not be meaningful. More thought and experimentation is required to assess the utility of ℓ_z in shape-based clustering for zero-inflated data.

2.4 Scale Concerns

Calculating the Frechét distance for multivariate longitudinal observations depends on the scales of the p covariates. For example, if one covariate is observed in the thousands and another covariate is observed in the tens, the Euclidean norm ℓ_2 puts more weight on changes in the first covariate than changes in the second covariate. Our focus was on genus counts from microbiome fecal samples where a transformation from absolute abundance to relative abundance is a common practice in the field. By dealing with compositional covariates that have common range [0,1], we evaded this issue of different scales among the p covariates.

There is also concern that the longitudinal time covariate is on a different scale than the p covariates. Not accounting for this scale difference may result in a clustering that overemphasizes the longitudinal times, ignoring the p covariates of primary interest. `kmlShape` remedies this by scaling the longitudinal time by a hyperparameter λ [Genolini et al., 2016]. We likewise multiplied the time vectors by a scalar and performed a grid search to optimally choose λ.

2.5 Cluster Stability and Separation

Evaluating the quality of data partitioning is of principal importance to any clustering method. We implemented two cluster statistics from the literature: the Rand index [Rand, 1971] for cluster stability and a Dunn-like index [Dunn, 1974] for cluster separation. For our study, we initialized clusters by sampling from our subjects, making our analysis subject to random initialization. The Rand index measures similarity between data partitions by summarizing among all pairings the agreement or disagreement on whether the pair (i, j) is in the same cluster. Let S be labeled subjects in the study and let X and Y be two data partitions.

$$\begin{aligned}
\text{TP} &= \text{number of pairs from S in the same cluster in both } X \text{ and } Y \\
\text{TN} &= \text{number of pairs from S in different clusters in both } X \text{ and } Y \\
\text{F} &= \text{number of pairs from S in same cluster and different clusters} \\
R &= \frac{TP + TN}{TP + TN + F}
\end{aligned}$$

For each random initialization, we computed the Rand index for the current data partitioning relative to the previous data partitioning. Next, we averaged these Rand indices. This average illustrates how stable cluster assignments with regard to different random initializations.

Dunn-like indices determine how well-separated clusters are. There is considerable flexibility in defining a Dunn-like index, but the general principle is to divide a numerator statistic based on between cluster variation by a denominator statistic based on within cluster variation. For the denominator statistic, for each cluster we calculated the average Frechét distance among assigned subjects with respect to the cluster mean and then we took the maximum over clusters. That is, the denominator statistic is the maximum average Frechét distance within clusters. A large denominator indicates a cluster that is ill-defined. For the numerator statistic, we determined the minimum Frechét distance with respect to cluster means. A small numerator indicates a cluster pair that is not well-separated. This index measures worst-case performance and can have poor results if the denominator is too large or if the numerator is too small. We prioritized the Rand index for cluster stability when optimally choosing hyperparameters k and λ.

3 Simulation Study

We evaluated our longitudinal clustering method on simulated multivariate data of differing shapes. Each simulation involved forty individuals with ten longitudinal measurements taken at the same time points. Each simulated data set exhibited a balanced number of data points in each simulated cluster. We assessed performance based on counting the number of misclassified individuals at each iteration of the simulation. Each of the following scenarios was simulated 100 times.

We first explored simulations of two clusters. Figure 2 shows a representative individual from each of the two simulated clusters where one grouping experiences an increase in one genus and a decrease in another and the other grouping experiences the opposite. Shape-based clustering correctly clustered all individuals for all iterations. Figure 3 shows a representative individual from each of the two simulated clusters where one of the five genuses dominates in each cluster. Shape-based clustering correctly clustered all individuals for all iterations. Figure 4 shows a representative individual from each of the two simulated clusters where one of the five genuses spikes in relative abundance in each cluster. Shape-based clustering correctly clustered 97 out of 100 iterations. Due to spikes caused by simulated random noise, this simulation achieved slightly less accuracy.

Next we considered simulations of four clusters. Figure 5 shows a representative individual in each of four simulated clusters where each cluster exhibits one of the patterns tested above. Shape-based clustering correctly classified in 47 out of 100 iterations. The overall simulation performance of the algorithm decreased as k increased because of random initialization of mean trajectories. When all initializations were from the same cluster, the algorithm led to inconvenient convergences like grouping two distinct clusters into one and leaving behind a near-empty cluster. When manually specifying the clustering method to initialize mean trajectories with a representative individual from each of the four clusters, shape-based clustering correctly partitioned the data in all 100 iterations.

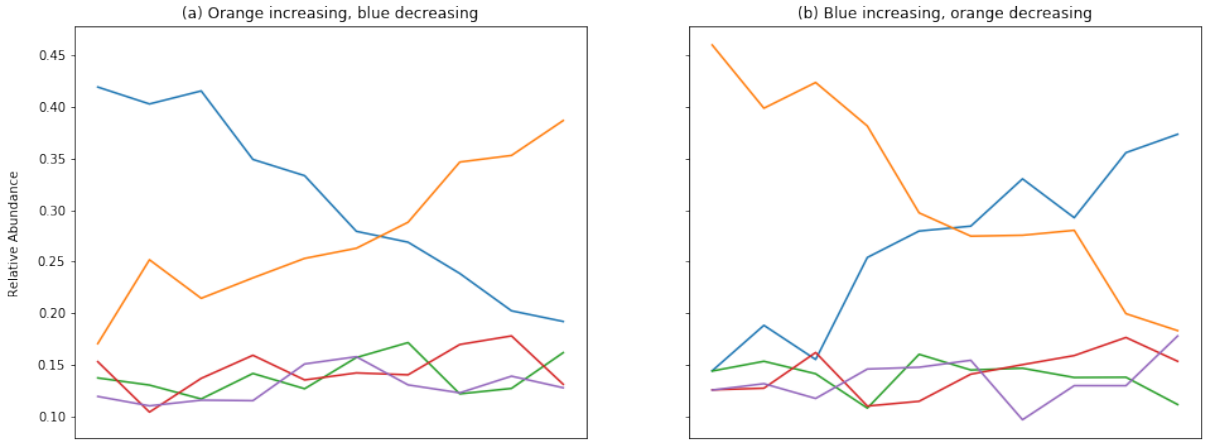

Figure 2: Sample of simulated data for each of two clusters in first simulation

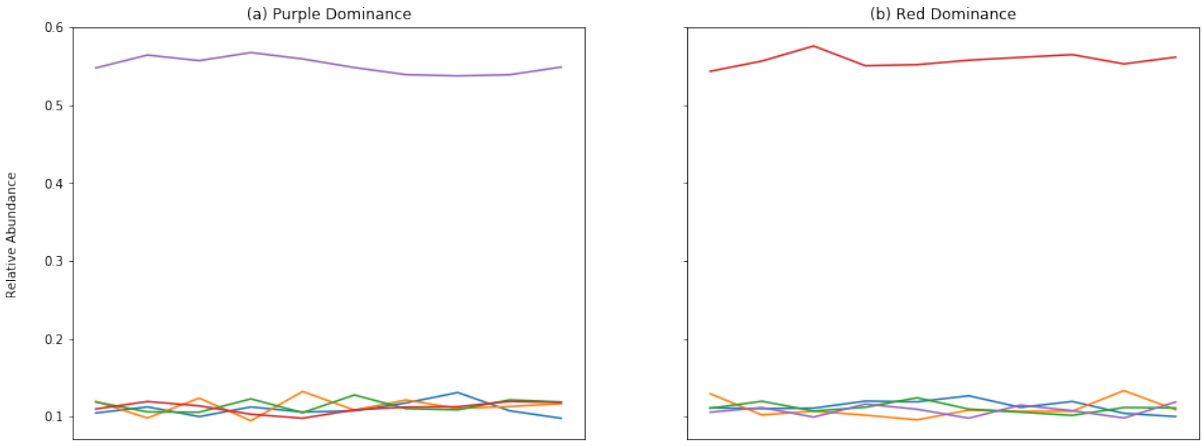

Figure 3: Sample of simulated data for each of two clusters in second simulation

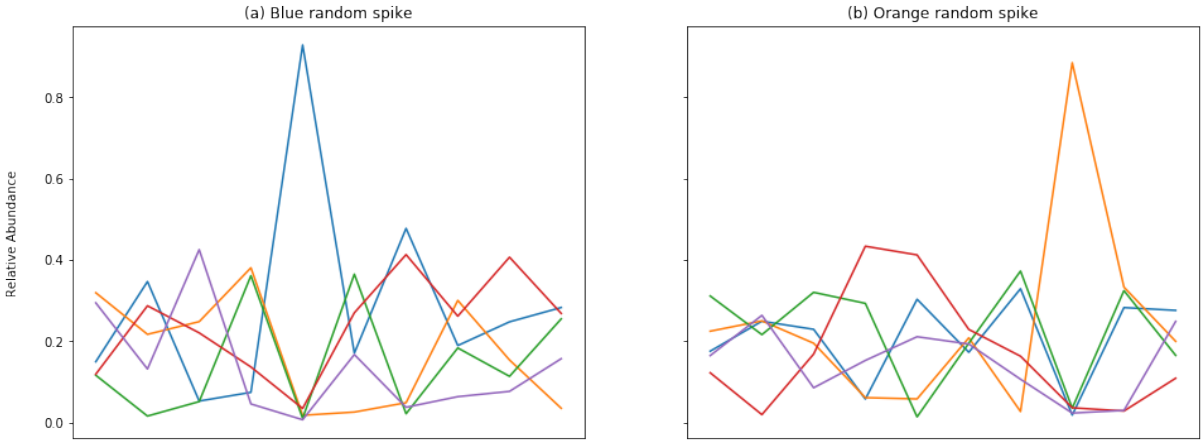

Figure 4: Sample of simulated data for each of two clusters in third simulation

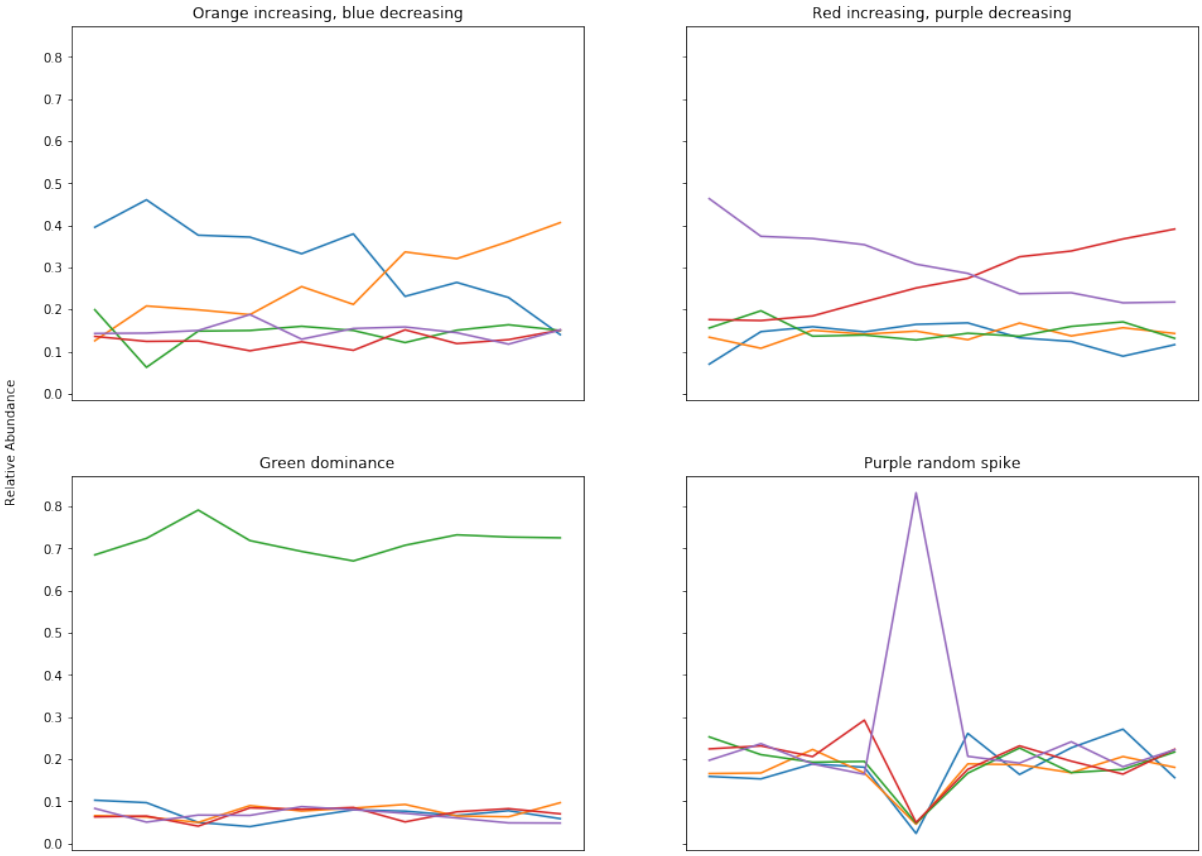

Figure 5: Sample of simulated data for each of four clusters in fourth simulation

4 Microbiome Data Analysis

4.1 Data Pre-Processing

We applied our method to a data set consisting of 2046 longitudinal microbiome fecal samples from 326 transplant patients. The objective was to learn more about how the microbiome is implicated in the development of graft-versus-host disease (GvHD) after a transplant. Each observation has 17 baseline covariates and counts from over 800 bacteria species. The original data contains many zero entries. This sparsity makes modeling difficult. In some exploratory model fitting, we found that generalized linear mixed models (GLMMs) encountered computational difficulties and that generalized estimating equations (GEEs) failed to identify statistically significant associations between GvHD presence or severity and specific microbes. We tried to trim the original data set and run regressions on microbes that appears in more than 1/2 or 3/4 of observations, taking the response to be GvHD presence or severity. Results on different taxonomic ranks (species, genus, family) offered conflicting findings. For example, genus *Odoribacter* was significant when regressing genus abundance on GvHD presence while insignificant when regressing genus presence on GvHD presence

($\alpha = 0.05$). In a trimmed data set with genuses that appear in more than 1/2 of the observations, *Odoribacter* was removed. On the other hand, family-level counts or presences were too vague to distinguish GvHD cases from controls.

After this exploratory data analysis, we decided to study further genuses that appear in 3/4 of all observations. Next, we restricted attention to the five most abundant and prevalent genuses for clustering: *Enterocloster*, *Lactobacillus*, *Blautia*, *Bacteroides*, and *Streptococcus*. In retrospect, these highly prevalent genuses are remarkably relevant: *Blautia* has been associated with reduced GvHD mortality [Jenq et al., 2015]; *Lactobacillus* is associated with immune responses in the form of allergies [Cox et al., 2010, Fujimura et al., 2014]; *Bacteroides* is one of the most common gut microbes overall and is associated with diet and lifestyle [Gorvitovskaia et al., 2016]; and *Streptococcus* is associated with diet [Jones et al., 2019] and health outcomes such as premature birth [Dahl et al., 2017]. With relative abundances for these five key genuses, non-parametric clustering became attractive for finding longitudinal patterns among our sample. For a longitudinal record of one individual, we focused on data points measured after the transplant surgery, as the operation might have imposed changes to microbes shown in the patient's body.

4.2 Parameter Grid Search

Data partitioning algorithms require specification of unknown hyperparameters [Kodinariya and Makwana, 2013]. For shape-based k-means clustering, the number of clusters k and the scale parameter λ exercise considerable influence on the procedure. For inappropriately small k, we may observe ill-defined clusters; for inappropriately large k, we may observe clusters not well-separated. Selection of k is important for cluster interpretability. λ controls how much impact the time covariate has on the Frechét distance. We performed two small grid searches to determine reasonable k and λ inputs. A more exhaustive grid search was prohibitively expensive because each run of k-means clustering took between 30-120 seconds to converge.

For each grid search we considered 11 random cluster initializations and ran k-means until subjects stopped changing cluster assignments. From (2.5) we calculated our Dunn-like index for each clustering run and computed the Rand index between the current and previous iteration assignments. In Table 1 we report the Dunn-like and Rand index averages. We favored parameter choices with high average Rand index as these data partitionings were most stable with regard to random initializations. High average Dunn-like index is ideal as well, but we emphasized this less since it measures worst case performance. For choice of λ, we also plotted cluster mean trajectories to evaluate if any cluster assignments were driven by longitudinal time. Ultimately, we selected $\lambda = 0.001$ and $k = 6, 10$ as parameter choices to move forward with.

Table 1: Parameter Grid Search

k	Rand Index	Dunn Index
4	0.78457	0.51163
6	0.84207	0.47778
8	0.85749	0.40582
10	0.86169	0.42176

λ	Rand Index	Dunn Index
0.01	0.83686	0.58811
0.001	0.84952	0.49079
0.0001	0.80240	0.46471

4.3 Results

To investigate the effects of varying k, we ran our clustering algorithm for $k = 6$ and $k = 10$. We found several interesting microbiome profiles, shown in Figure 6 for $k = 6$. For 10 clusters, we observed similar overall trajectories with some clusters appearing visually to be repeats of one another. The qualitatively distinct patterns visible in these clusters for $k = 6$ is promising. We uncovered clusters with distinct genus compositions overall and across time.

Table 2: Summary statistics for six clusters

Cluster	Mean Grade	Mean Occurrence
1	1.265	0.588
2	2.061	0.818
3	1.538	0.692
4	1.353	0.618
5	1.800	0.780
6	1.225	0.625

Our motivating hypothesis was that different clusters would be associated with significantly different GvHD outcomes. To investigate whether our clusters were indeed associated with GvHD, we calculated three statistics for each cluster: the mean occurrence of GvHD, the mean day of onset for those with GvHD, and the mean severity (grade) of disease for those with GvHD. For our test statistics, we calculated the inter-cluster range of each of these three statistics. We then conducted bootstrap simulations, replicating the sizes of the clusters and recalculating these three test statistics in each bootstrapped replicate. We computed p-values for our actual clustering test statistics based on the empirical quantile function from bootstrapped replicates. In essence, we quantified the association between clusters and GvHD based on the difference between the cluster with the highest and lowest GvHD outcomes. These bootstraps were run with a total of 10,000 simulations.

Our bootstrap analysis revealed that for $k = 6$ the range of mean GvHD grades was abnormally large ($p = 0.04$) and for $k = 10$ the range of GvHD occurrence was abnormally large ($p = 0.02$). Neither effect was robust to a Bonferroni correction for our three-fold

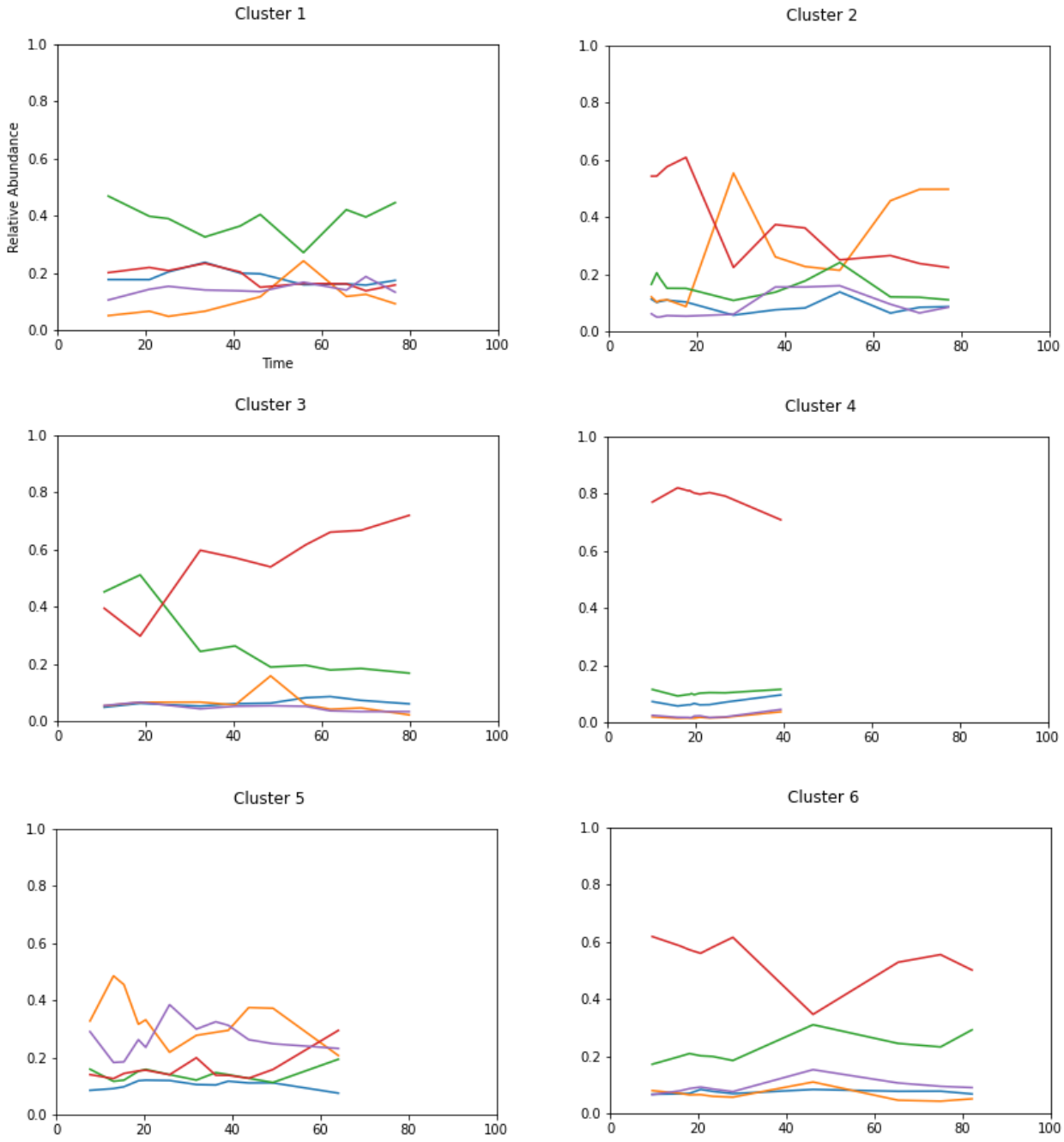

Figure 6: Longitudinal profiles found via shape-respecting k-means. Colors correspond to common genuses. Blue = *Enterocloster*, Orange = *Lactobacillus*, Green = *Blautia*, Red = *Bacteroides*, Purple = *Streptococcus*.

multiple testing; however, in this final analysis and in our initial explorations, we consistently saw similar small but significant effects.

Although the range in occurrence and date-of-onset were not significant for $k = 6$, the identified cluster with maximum GvHD grade also had the earliest average date of onset and the highest average occurrence of GvHD. This was despite a relatively low correlation between grade and date (-0.19) and no association between occurrence and either grade or date (since neither is defined if GvHD does not occur). Therefore, this association cannot be due to inherent correlation between GvHD metrics but must instead be due to association with some unmeasured variable, variation in which is captured by our clusters. To investigate this association further, we examined the correlation between mean metrics at the cluster level, specifically the high correlation we discovered between occurrence and grade at the cluster level, to determine whether this high correlation could be attributed to random chance. We discovered that for 6 clusters our correlation of 0.834 was significant ($p = 0.022$), but for 10 clusters our correlation of 0.593 was not ($p = 0.055$).

5 Conclusions

Our method offers a promising technique for data exploration in longitudinal data sets. We have developed an extension of univariate shape-based clustering [Genolini et al., 2016] to allow for the extremely common practical case of multivariate data. Moreover, we demonstrated that our method can identify sub-populations with differential outcomes for both GvHD occurrence and GvHD severity on the basis of only relative abundance of prominent gut microbiota.

In particular, for the more interpretable $k = 6$ case, we found clusters that represented a spectrum from "high GvHD" clusters 2, 3, and 5 to "low GvHD" clusters 1, 4 and 6. Interpretation of the genuses in these clusters is difficult; however, high levels of *Blautia* have previously been implicated in improved GvHD outcomes [Jenq et al., 2015]. One interesting pattern to note is that clusters 1 and 5, two low-GvHD clusters, both had consistently high levels of *Blautia*, while cluster 3, a high-GvHD cluster, showed a decline in *Blautia* that could coincide with onset of GvHD. Another interesting observation is that high GvHD clusters 2 and 5 both experience spikes between days 20-40 for genuses *Lactobacillus* and *Streptococcus* respectively.

Our method can be expected to work well when there are complex relationships between variables through time. On the other hand, when data conforms to a simpler parametric model, such as a GLMM or GEE, these models would be expected to produce more useful results. Our method also suffers from the problem of local optima, and may have poor data partitioning when random initializations come from the same latent subgroup. This is a common issue with k-means algorithms, and there are many approaches to solving it which could be investigated in future work [Steinley and Brusco, 2007]. Finally, since this is a multivariate method, when the scale of variables differs dramatically, pre-processing is

needed in order to ensure good model performance. We solved this problem by clustering on relative abundance, but other approaches such as data normalization could be used in other contexts.

In future work, there is much room for further expansion of this method. We make available our work for these purposes (`https://github.com/sdtemple/mvtraj`). One avenue for expansion would be to modify the underlying metric space S. This would allow easy incorporation of different forms of data. For example, binary data could be incorporated using the Hamming distance [Wu et al., 2015]. By incorporating more sophisticated versions of k-means for high dimensional data, the Frechét method for longitudinal clustering could also be extended to identify shape-based subspace clusters [Jing et al., 2007]. Finally, we doubt that there is a strong causal effect of the exact genuses we used for our analysis here. Further work should focus on verifying the association between GvHD and microbiome trajectories, elucidating plausible mechanistic pathways through which the two may be associated.

References

Karl Bringmann, Marvin Künnemann, and André Nusser. Walking the dog fast in practice: Algorithm engineering of the fr\'echet distance. *arXiv preprint arXiv:1901.01504*, 2019.

Michael J Cox, Yvonne J Huang, Kei E Fujimura, Jane T Liu, Michelle McKean, Homer A Boushey, Mark R Segal, Eoin L Brodie, Michael D Cabana, and Susan V Lynch. Lactobacillus casei abundance is associated with profound shifts in the infant gut microbiome. *PloS one*, 5(1):e8745, 2010.

Cecilie Dahl, Maggie Stanislawski, Nina Iszatt, Siddhartha Mandal, Catherine Lozupone, Jose C Clemente, Rob Knight, Hein Stigum, and Merete Eggesbø. Gut microbiome of mothers delivering prematurely shows reduced diversity and lower relative abundance of bifidobacterium and streptococcus. *PloS one*, 12(10):e0184336, 2017.

David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2):112–122, 1973.

Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. Technical report, Citeseer, 1994.

João Paulo Figueira. How long should your dog leash be?, 2020. URL `https://medium.com/tblx-insider/how-long-should-your-dog-leash-be-ba5a4e6891fc`.

Kei E Fujimura, Tine Demoor, Marcus Rauch, Ali A Faruqi, Sihyug Jang, Christine C Johnson, Homer A Boushey, Edward Zoratti, Dennis Ownby, Nicholas W Lukacs, et al. House dust exposure mediates gut microbiome lactobacillus enrichment and airway immune defense against allergens and virus infection. *Proceedings of the National Academy of Sciences*, 111(2):805–810, 2014.

Christophe Genolini, René Ecochard, Mamoun Benghezal, Tarak Driss, Sandrine Andrieu, and Fabien Subtil. kmlshape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *Plos one*, 11(6):e0150738, 2016.

Anastassia Gorvitovskaia, Susan P Holmes, and Susan M Huse. Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome*, 4(1):1–12, 2016.

Ning Guo, Mengyu Ma, Wei Xiong, Luo Chen, and Ning Jing. An efficient query algorithm for trajectory similarity based on fréchet distance threshold. *ISPRS International Journal of Geo-Information*, 6(11):326, 2017.

Robert R Jenq, Ying Taur, Sean M Devlin, Doris M Ponce, Jenna D Goldberg, Katya F Ahr, Eric R Littmann, Lilan Ling, Asia C Gobourne, Liza C Miller, et al. Intestinal blautia is associated with reduced death from graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 21(8):1373–1383, 2015.

Liping Jing, Michael K Ng, and Joshua Zhexue Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8):1026–1041, 2007.

Roshonda B Jones, Tanya L Alderete, Jeniffer S Kim, Joshua Millstein, Frank D Gilliland, and Michael I Goran. High intake of dietary fructose in overweight/obese teenagers associated with depletion of eubacterium and streptococcus in gut microbiome. *Gut microbes*, 10(6):712–719, 2019.

Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

D Napoleon and S Pavalakodi. A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, 13(7):41–46, 2011.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, 18:2789, 2020.

Douglas Steinley and Michael J Brusco. Initializing k-means batch clustering: A critical evaluation of several techniques. *Journal of Classification*, 24(1):99–121, 2007.

Guohua Wu, Hairong Lin, Ershuai Fu, and Liuyang Wang. An improved k-means algorithm for document clustering. In *2015 international conference on computer science and mechanical automation (CSMA)*, pages 65–69. IEEE, 2015.