Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 5284145, 7 pages
https://doi.org/10.1155/2017/5284145

Research Article

2-Way *k*-Means as a Model for Microbiome Samples

Weston J. Jackson,[1] Ipsita Agarwal,[2] and Itsik Pe'er[1]

[1]*Department of Computer Science, Columbia University, New York, NY 10027, USA*
[2]*Department of Biological Sciences, Columbia University, New York, NY 10027, USA*

Correspondence should be addressed to Weston J. Jackson; weston.j.jackson@gmail.com

Received 20 May 2017; Accepted 17 July 2017; Published 5 September 2017

Academic Editor: Ahmad P. Tafti

Motivation. Microbiome sequencing allows defining clusters of samples with shared composition. However, this paradigm poorly accounts for samples whose composition is a mixture of cluster-characterizing ones and which therefore lie in between them in the cluster space. This paper addresses unsupervised learning of 2-way clusters. It defines a mixture model that allows 2-way cluster assignment and describes a variant of generalized *k*-means for learning such a model. We demonstrate applicability to microbial 16S rDNA sequencing data from the Human Vaginal Microbiome Project.

1. Introduction

Microbiome analysis [1] by sequencing of ubiquitous genes, most commonly 16S rRNA, is a standard, cost-effective way to characterize the composition of a microbial sample. Standard analysis tools facilitate quantifying the fraction of sequence reads from each bacterial species in a sample [2]. Interpretation of composition vectors across a collection of samples typically relies on dimensionality reduction followed by clustering in the lower-dimensionality space [3]. This allows identification of functionally meaningful subsets of samples with characteristic microbiota. The Human Microbiome Project [4] and its derivatives such as the Human Vaginal Microbiome Project [5] have collected and thus analyzed large numbers of samples towards elucidating the structure and composition of microbiota across physiological and pathological states.

Similar to variation in microbial genomes across different human individuals, variants along the nuclear genomes have been summarized by a small number of dimensions [6]. However, in contrast to analyses of microbiome samples, analyses of inherited genetic variation standardly assume and observe samples to be spread across a continuum in the reduced space, rather than be clustered [7]. Samples in between clusters are interpreted as originating from intermediate locales along a geographic cline [8] or as representing different levels of a mixture between cluster-specific populations.

In this paper, we formally tackle the problem of clustering while allowing elements to belong to two clusters. Specifically, we will describe in detail a model for clustering in $\mathbb{R}^d$. We construct a model that generalizes *k*-means clustering by allowing data points to be assigned to a point in the space along the line between two assigned clusters [9]. Each cluster is still modeled as a Gaussian with uniform, spherical covariances; the key difference is the presence of a parameter $u \in [0, 1]$ for each 2-way-assigned data point x_i, which determines the proportional assignment of x_i between its two cluster representatives. We first describe the 2-way model's inputs, parameters, and outputs. We then give the objective function, an algorithmic description, and a series of performance metrics. Next, we evaluate the performance on simulated data, describing benchmarks for optimal performance. Finally, we apply the model to real data of 16S rDNA sequencing from 1500 midvaginal bacterial samples by the Vaginal Human Microbiome Project.

2. Methods

2.1. 2-Way k-Means. The model characterizes a mixture where points are each sampled either from a *k*-mixture of

uniform, spherical Gaussian distributions or from pairwise weighted averages of these Gaussians.

Formally, we describe a generative model for a set X of data points $\{x_i\}_1^n \in \mathbb{R}^d$. The model involves $k \in \mathbb{Z}^+$ clusters. The jth cluster is parametrized by its mean $\mu_j \in \mathbb{R}^d$. To simulate x_i, the model first chooses a pair of cluster indices (j, j') along with a weighting $u_i \in [0, 1]$. x_i is drawn from a Gaussian distribution whose parameters are u_i-weighted averages of two representative clusters. Specifically, $x_i \sim N(x_i; \widetilde{\mu}_{ijj'}, \Sigma)$ such that $\widetilde{\mu}_{ijj'} = u_i\mu_j + (1 - u_i)\mu_{j'}$ and $\Sigma \in \mathbb{R}^{d\times d}$ is the given uniform, spherical covariance matrix.

The inference problem involves the inputs of data X and number of clusters k, seeking output of the generative model parameters, that is, the vectors of assignments $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$ and weights $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$.

2.2. Generalized k-Means. Given input $x_1, \ldots, x_n \in \mathbb{R}^d$ and cardinality $k \in \mathbb{N}$, k-means traditionally provides us with the following objective:

$$\sum_{i=1}^{n} \min_{j\in[k]} \left\|x_j - c_j\right\|_2^2, \tag{1}$$

where $c_1, \ldots, c_j$ are the cluster representatives. The k-means objective can be generalized as the following:

$$\min_{C,\Phi} \sum_{i=1}^{n} \|x_i - C\phi_i\|_2^2, \tag{2}$$

where $\Phi = [\phi_1|\phi_2|\cdots|\phi_n] \in \{0, 1\}^{k\times n}$ are the cluster assignments and $C = [c_1|c_2|\cdots|c_k] \in \{0, 1\}^{d\times k}$ are the cluster representatives.

A common generalization of k-means is to permit each ϕ_i to have s nonzero entries (in our case, we set $s = 2$). An algorithm for this generalized objective is simply to hold C fixed while performing sparse regression on Φ and then hold Φ fixed and use ordinary least squares (OLS) to find C.

In our case, because we only allow points x_i to lie uniformly between two cluster representatives, the two nonzero entries in a given ϕ_i are restricted to some $u_i \in [0, 1]$ and $1 - u_i \in [0, 1]$. Our problem is instead the following:

$$\min_{C,\Phi} \sum_{i=1}^{n} \|x_i - C\phi_i\|_2^2, \tag{3}$$

subject to

$$\|\phi_i\|_0 \le 2, \quad \|\phi_i\|_1 = 1, \quad \phi_i \ge 0. \tag{4}$$

2.3. 2-Way k-Means Algorithm. Our goal is to find a nonnegative 2-sparse solution for each ϕ_i. To do so, we can minimize over all $\binom{k}{2}$ cluster representative possibilities. This 2-sparse solution gives us indices (j, j') which correspond with the two cluster representatives. This corresponds with the following objective:

$$\min_{u_i\in\mathbb{R}, c_{j'}, c_j\in C} \left\|x_i - \left(u_i c_j + (1 - u_i)c_{j'}\right)\right\|_2^2, \tag{5}$$

subject to $u_i \in [0, 1]$.

For a given c_j and $c_{j'}$, minimizing with respect to $u_{ijj'}$ reveals a global minimum at

$$\frac{\left(c_{j'} - c_j\right)^T \left(c_{j'} - x_i\right)}{\left\|c_{j'} - c_j\right\|_2^2}. \tag{6}$$

After minimizing with respect to $u_{ijj'}$, we project $u_{ijj'}$ to the region $[0, 1]$. We set $u_{ijj'} = 0$ if the minimizer is less than 0 and set $u_{ijj'} = 1$ if the minimizer is greater than 1. This allows us to achieve the minimum value of u_i over the domain $[0, 1]$ for x_i.

After minimizing the assignment Φ, we then use OLS to pick optimal C as specified before. Formally, OLS produces a vector $\mathbf{c}_i^T$ that minimizes the squared residual error between an input matrix $\mathbf{\Phi}^T$ and vector $\mathbf{x}_i^T$.

$$\min_{\mathbf{c}_i^T} \left\|\mathbf{x}_i^T - \mathbf{\Phi}^T \mathbf{c}_i^T\right\|_2^2. \tag{7}$$

Taking the gradient and setting equal to zero yields the following formula:

$$\mathbf{c}_i^T = \left(\mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}\mathbf{x}_i^T. \tag{8}$$

Thus, we perform OLS for all vectors $\mathbf{c}_i^T$ at once with matrix multiplication:

$$\mathbf{C}^T = \left(\mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}X. \tag{9}$$

Thus, this gives us representatives $c_1, \ldots, c_k$ that minimize the residual error between the cluster representatives and data points subject to Φ. We then alternate this process for r rounds until convergence.

2.4. Performance Metrics. We use the 2-way k-means objective as a performance metric in measuring the accuracy of a model in unsupervised examples.

$$\text{obj}(X, k, r) = \min_{C,\Phi} \sum_{i=1}^{n} \|x_i - C\phi_i\|_2^2, \tag{10}$$

where Φ has at most two nonzero entries with values $u_i \in [0, 1]$ and $1 - u_i \in [0, 1]$.

Additionally, we also use four different error rates to measure the accuracy of 2-way k-means on test cases. Let c_i^*, μ_j^*, and $u_{ijj'}^*$ be the ground truth instance parameters, that is, respectively, true 2-way cluster assignment of x_i, center of cluster j, and 2-way weighting for x_i between clusters (j, j').

$\text{err}_{f(x)}$ defines the 0-1 error rate for 2-way cluster assignment:

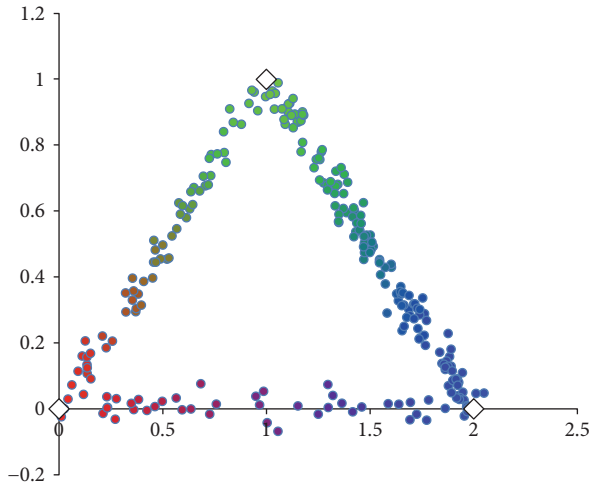

Figure 1: $n = 500$ simulated data points. The white diamonds are cluster centers for three simulated clusters (red cluster, bottom left; green cluster, top; and blue cluster, bottom right). Points are colored as a linear combination of the clusters they lie between (according to u).

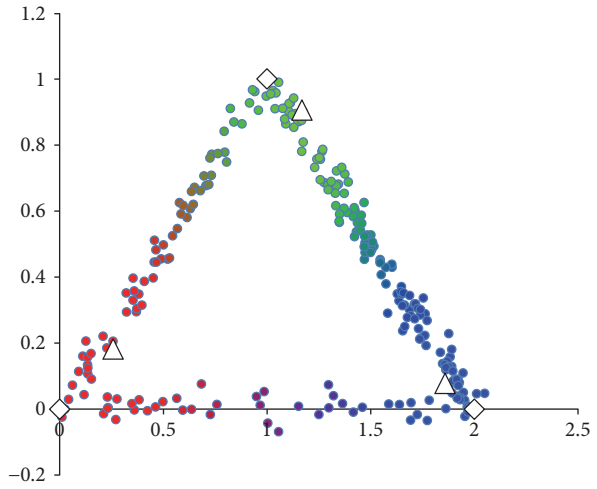

Figure 2: $n = 500$ simulated data points after k-means. The white diamonds are cluster centers for three simulated clusters (red cluster, bottom left; green cluster, top; and blue cluster, bottom right). White triangles are cluster centers determined by u means. Colors are u values determined by k-means.

$$\text{err}_{f(x)} = \frac{\sum_{i=1}^{n} \mathbb{1}_{c_i \neq c_i^*}}{n}. \tag{11}$$

err_{μ} defines the squared deviation from optimal μ^*:

$$\text{err}_{\mu} = \sum_{\forall j} \left\| \mu_j - \mu_j^* \right\|_2. \tag{12}$$

err_u defines the squared deviation from optimal $u^*_{ijj'}$. WLOG, we assume $u_{ijj'} = \max(u, 1-u)$, where u is the variable drawn from $[0, 1]$:

$$\text{err}_u = \frac{\sum_{i=1}^{n} \left\| u^*_{ijj'} - u_{ijj'} \right\|_2}{n}. \tag{13}$$

3. Results

3.1. Example Run for 2-Way k-Means. We find it illuminating to demonstrate the performance of 2-way k-means versus vanilla k-means on a cartoon example.

In Figure 1, we simulated $n = 500$ data points in $\mathbb{R}^2$ from three clusters, with respective means $\mu_1 = [0, 0]$, $\mu_2 = [1, 1]$, and $\mu_3 = [2, 0]$ and covariance matrices $\Sigma = 0.001I$. Data points are drawn into pairwise clusters by choosing two cluster representatives without replacement from the following prior probabilities:

$$\begin{aligned} P(c_1) &= 0.2, \\ P(c_2) &= 0.5, \\ P(c_3) &= 0.3. \end{aligned} \tag{14}$$

We initialize the cluster representatives with vanilla k-means. Vanilla k-means achieves the results in Figure 2. Statistics for vanilla k-means are given as follows:

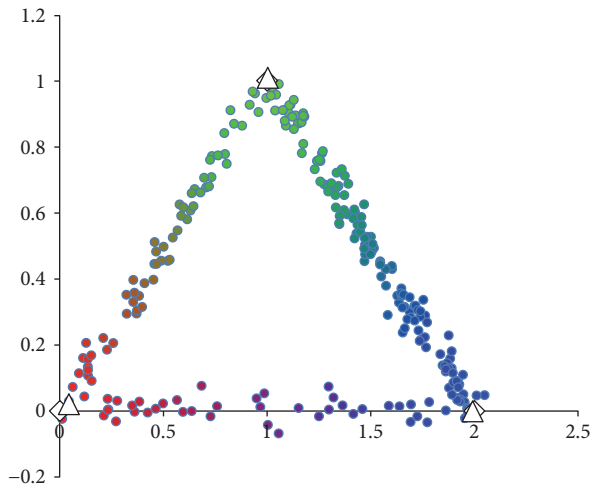

Figure 3: $n = 500$ simulated data points after 10 rounds of 2-way k-means. White diamonds are cluster centers determined by 2-way k-means (10 rounds). Colors are u values determined by 2-way k-means (10 rounds).

$$\begin{aligned} \text{obj} &: 41.1025, \\ \text{err}_{f(x)} &: 0.112, \\ \text{err}_{\mu} &: 0.4039, \\ \text{err}_u &: 0.2435. \end{aligned} \tag{15}$$

k-means predicts the 2-way cluster assignments of ≈11% points incorrectly due to skewing the cluster means toward the middle of the graph. 2-way k-means, however, significantly improves on all error rates. After 10 rounds of 2-way k-means, we achieve the results in Figure 3.

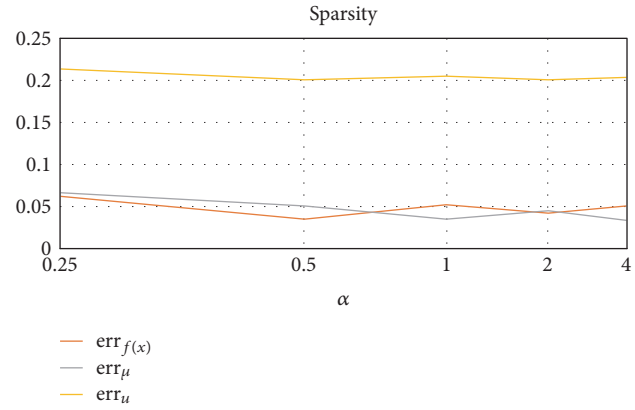

Figure 4: Error rates when $n = \alpha 500$, and cluster priors and centers are fixed.

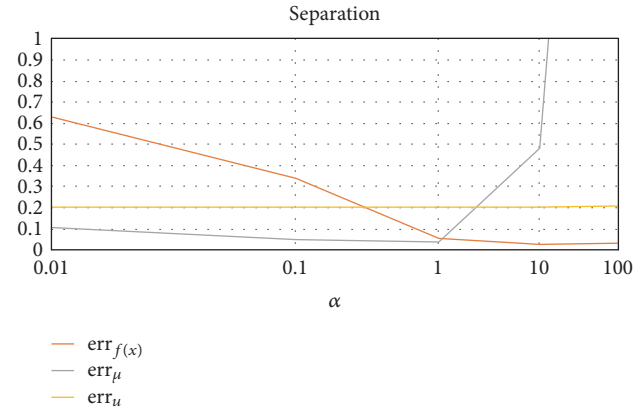

Figure 5: Error rates as a function of the Euclidean distances of μ, where $\mu_1 = \alpha[0, 0]$, $\mu_2 = \alpha[1, 1]$, and $\mu_3 = \alpha[2, 0]$.

$$\begin{aligned} \text{obj} &: 5.822, \\ \text{err}_{f(x)} &: 0.032, \\ \text{err}_{\mu} &: 0.0495, \\ \text{err}_{u} &: 0.2084. \end{aligned} \tag{16}$$

For every statistic, the results are clearly an improvement on standard k-means. The ≈3% error rate on cluster assignment still exists because 2-way k-means points closest to cluster representatives may be assigned to an incorrect secondary cluster.

3.2. Benchmarks

3.2.1. Sparsity (Avg. of 10 Trials, 10 Rounds Each). Our sparsity test was conducted by keeping cluster prior probabilities and cluster centers μ constant while varying the number of data points (ratio of α means $n = 500\alpha$). From Figure 4, we see that the algorithm performs consistently well under a variety of conditions, but too few data points can hurt performance to an extent.

3.2.2. Cluster Separation (Avg. of 10 Trials, 10 Rounds Each). We test the error rate as a function of the Euclidean distances of μ (ratio of α means $\mu_1 = \alpha[0, 0]$, $\mu_2 = \alpha[1, 1]$, and $\mu_3 = \alpha[2, 0]$). From the results in Figure 5, we can see that a certain threshold is required for proper performance of the algorithm. This makes sense, as when $\alpha = 0.01$, the clusters are almost on top of each other and difficult to distinguish. Additionally, as the cluster centers are moved farther apart, the ℓ_2 norm between the cluster representative determined by the algorithm and the actual cluster representative increases (but this is to be expected).

3.2.3. Variance (Avg. of 10 Trials, 10 Rounds Each). We increase the variance of the clusters while fixing cluster prior probabilities, data points, and cluster centers (ratio of α means $\Sigma = \alpha[[0, 0.0001], [0, 0.001]]$). From the results in Figure 6, we can see that large variance hurts proper performance of the algorithm. Analogous to cluster separation, as when $\alpha = 100$, the clusters are too close to distinguish.

3.3. Real Data. Publicly available sequence data for the Human Microbiome Project (HMP) study SRP002462,

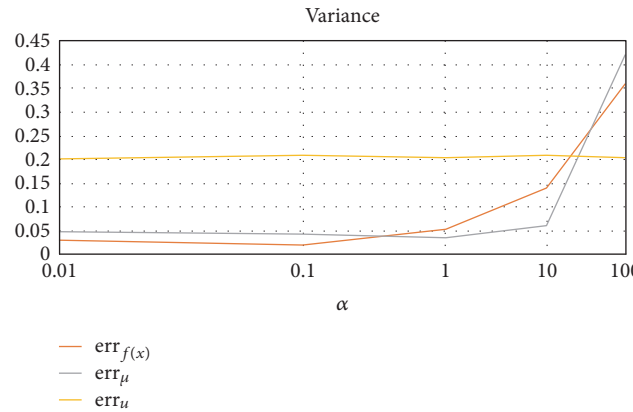

Figure 6: Error rates as a function of cluster variance Σ, where $\Sigma = \alpha[[0, 0.001], [0, 0.001]]$.

described as metagenomic sequencing of 16S rDNA from vaginal and related samples from clinical and twin subjects, was downloaded from the NCBI SRA database [10]. The downloaded sets of data correspond to two separate submissions: SRA169809 (1608/1608 samples were downloaded) and SRA273234 (34/133 samples were downloaded), for a total of 1642 samples.

The SFF files were processed and cleaned using the microbial community analysis software mothur [11], based on a standard protocol developed for 454 sequence data processing and quality control [12]. The dissimilarities between the samples were calculated using the Clayton-Yue dissimilarity measure. The data was subsampled to 5000 sequences per sample (this step results in dropping out 136 samples that had less than 5000 reads in total) 500 times to produce the distance file, which was used to calculate principal coordinates. Figure 7 shows the graph of ~1500 data points after PCoA. After implementing the 2-way k-means algorithm [13], we initialized with k-means $k = 5$ and ran 2-way k-means for 5 rounds on the data.

Unfortunately, the nonlinear arches between the clusters pushed the cluster representatives slightly outside the clusters. Nonetheless, the algorithm was still an improvement

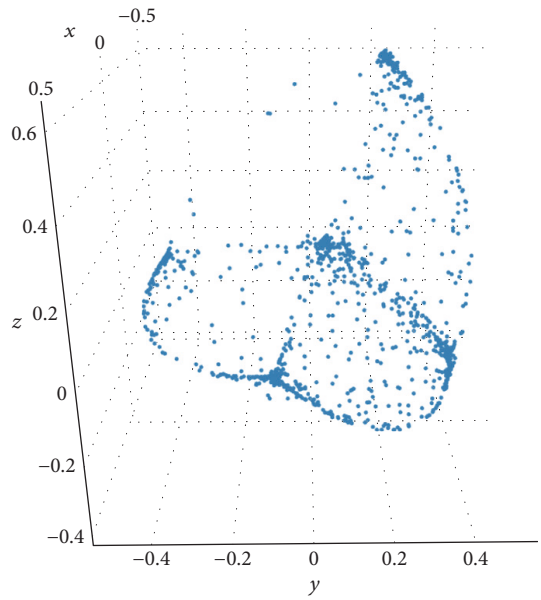

Figure 7: 1500 data points graphed after PCoA.

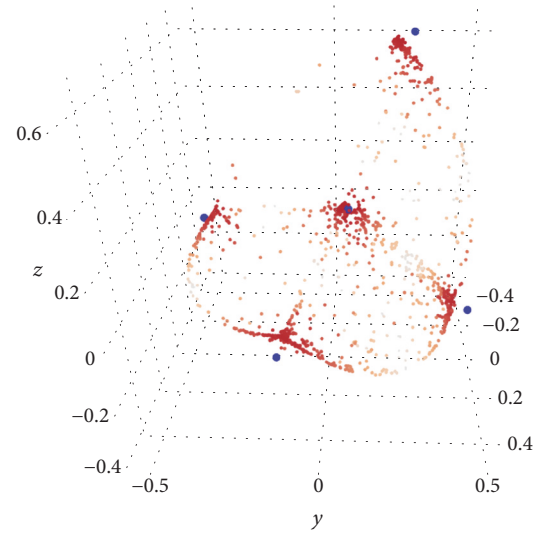

Figure 8: ~1500 data points after 5 rounds of 2-way k-means. The dark points are closer to the cluster representatives while the lighter points are more between them. The cluster representatives are the larger blue points that are slightly outside the clusters (compensating for the nonlinear arches between clusters).

over k-means. We note that after k-means, the 2-way objective had a value of 108.0 while our 2-way k-means algorithm converged on an objective of ≈51.0 after 5 rounds. Additionally, the algorithm gives us a characterization of the samples lying between two clusters. The results can be seen in Figure 8.

3.4. Discussion. We first get the most abundant operational taxonomic unit (OTU) in each sample (down to the genus level) and the closest cluster assignment for each sample. We use this to observe which OTUs are the most common

Table 1: The most abundant OTU per cluster. Because the 16S rDNA data maps multiple sequences to the same genus level, we use subscripts to denote different OTUs with the same genus.

Family; genus	c_1	c_2	c_3	c_4	c_5
Lactobacillaceae; *Lactobacillus*$_1$	5	534	17	0	1
Lactobacillaceae; *Lactobacillus*$_2$	0	0	3	269	0
Bifidobacteriaceae; *Gardnerella*	227	0	13	0	1
Lachnospiraceae; unclassified$_1$	0	0	4	0	150
Lactobacillaceae; *Lactobacillus*$_3$	2	0	44	2	0
Lactobacillaceae; *Lactobacillus*$_4$	2	4	32	1	1
Leptotrichiaceae; *Sneathia*$_1$	8	0	33	0	0
Prevotellaceae; *Prevotella*$_1$	2	0	30	0	0
Prevotellaceae; *Prevotella*$_2$	3	0	16	0	1
Unclassified; unclassified$_2$	2	1	15	0	2
Prevotellaceae; *Prevotella*$_3$	0	0	4	0	1
Leptotrichiaceae; *Sneathia*$_2$	1	0	2	0	1
Lachnospiraceae; unclassified$_3$	0	0	1	0	3
Streptococcaceae; *Streptococcus*$_1$	0	0	18	0	0
Veillonellaceae; unclassified$_4$	0	0	0	0	0
Streptococcaceae; *Streptococcus*$_2$	0	1	15	0	0
Mycoplasmataceae; *Mycoplasma*	0	1	7	0	0
Bifidobacteriaceae; *Bifodobacterium*	0	0	9	0	0
Fusobacteriaceae; *Fusobacterium*	0	0	7	0	0
Enterobacteriaceae; unclassified$_5$	0	0	8	0	0

to each cluster. We can find the closest sample to each data point by simply taking the argmax(u) for each data point x_i.

From Table 1, we see that four of the five clusters have a unique and most abundant OTU, while cluster c_3 has a variety of abundant types. Aside from the top four OTUs, separating the data into discrete clusters obscures how the rest of the OTUs can be characterized.

By using each data point's cluster-pair assignment, we further separate the data into $k^2 - k$ clusters. Let $c_{jj'}$ designate the data points that are between clusters j and j' but are nearer to cluster j than cluster j'. We take the most abundant OTUs in each sample and the cluster pair for each sample. We can then find the most abundant OTUs for each cluster pair.

Table 2 shows the structure of the most abundant OTU types for each 2-way cluster $c_{jj'}$ defined before. Once again, we find that clusters c_{1j}, c_{2j}, c_{4j}, and c_{5j} are all dominated by the same single OTU from before. Yet, observing clusters $c_{3j'}$ provides us with a more in-depth understanding of the diverse cluster c_3.

Interestingly, we see that the makeups of c_{31}, c_{32}, c_{34}, and c_{35} are remarkably different. We immediately see that the top four OTUs are all predominantly contained in cluster pairs that include their single main cluster. In addition, we notice that the samples with abundant *Sneathia*$_1$, *Prevotella*$_2$, and unclassified types are predominantly contained in c_{31}. c_{32} contains samples with a variety of abundant OTUs. *Lactobacillus*$_3$, *Lactobacillus*$_4$, *Prevotella*$_1$, *Streptococcus*$_1$, *Streptococcus*$_2$, and *Bifodobacterium* are abundant in samples that are

Table 2: The most abundant OTUs per cluster pair.

Genus	c_{12}	c_{13}	c_{14}	c_{15}	c_{21}	c_{23}	c_{24}	c_{25}	c_{31}	c_{32}	c_{34}	c_{35}	c_{41}	c_{42}	c_{43}	c_{45}	c_{51}	c_{52}	c_{53}	c_{54}
Lactobacillus$_1$	5	0	0	0	113	69	57	295	0	15	2	0	0	0	0	0	0	1	0	0
Lactobacillus$_2$	0	0	0	0	0	0	0	0	0	0	3	0	14	60	27	168	0	0	0	0
Gardnerella	75	32	95	25	0	0	0	0	10	0	3	0	0	0	0	0	1	0	0	0
Unclassified$_1$	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	54	23	23	50
Lactobacillus$_3$	0	2	0	0	0	0	0	0	6	7	31	0	0	0	2	0	0	0	0	0
Lactobacillus$_4$	1	1	0	0	0	4	0	0	4	4	23	1	0	1	0	0	1	0	0	0
Sneathia$_1$	0	8	0	0	0	0	0	0	22	2	4	5	0	0	0	0	0	0	0	0
Prevotella$_1$	0	2	0	0	0	0	0	0	9	4	17	0	0	0	0	0	0	0	0	0
Prevotella$_2$	0	0	0	3	0	0	0	0	15	0	0	1	0	0	0	0	0	0	1	0
Unclassified$_2$	2	0	0	0	0	1	0	0	8	6	0	1	0	0	0	0	1	0	1	0
Prevotella$_3$	0	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	1	0	0	0
Sneathia$_2$	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
Unclassified$_3$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0
Streptococcus$_2$	0	0	0	0	0	0	0	0	2	2	14	0	0	0	0	0	0	0	0	0
Unclassified$_4$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Streptococcus$_2$	0	0	0	0	0	1	0	0	2	2	11	0	0	0	0	0	0	0	0	0
Mycoplasma	0	0	0	0	1	0	0	0	5	2	0	0	0	0	0	0	0	0	0	0
Bifodobacterium	0	0	0	0	0	0	0	0	1	1	7	0	0	0	0	0	0	0	0	0
Fusobacterium	0	0	0	0	0	0	0	0	1	2	4	0	0	0	0	0	0	0	0	0
Unclassified$_5$	0	0	0	0	0	0	0	0	0	2	5	1	0	0	0	0	0	0	0	0

predominantly contained in c_{34}. Finally, almost no samples are in the cluster pair c_{35}, aside from a few *Sneathia*$_1$ types.

In this way, 2-way k-means also opens up a wealth of information on the relationships between samples. In particular, it now makes more sense to characterize the samples as being in 6 different clusters: c_1, c_2, c_{31}, c_{34}, and c_5. We also see that certain clusters have mixed relationships, while others have almost no interaction. Without 2-way k-means, this would not be immediately obvious.

4. Conclusion

The complexity of microbial populations is unfolding as microbiome data becomes increasingly available. Yet, standard methodologies oversimplify microbial compositions by pigeonholing them into discrete clusters. This paper further refines the models for microbial abundance across groups of samples. We allow samples to be presented as a weighted average of two clusters, rather than belonging to only one. This may be motivated biologically, as the sample often reflects a mixture of two sources of microbiota, each well represented by a cluster. An alternative explanation is that the averaged sample represents an intermediate, potentially temporary state of the microbial composition, between the more stable ones represented by the clusters themselves.

Technically, we formalize this model as a generalization of k-means. We derive a simple algorithm to infer such a structure and validate its benchmarks on simulated data.

Applying our algorithm to real data from the Human Vaginal Microbiome Project provides empirical support to the 2-way model. We showed that while most of the samples lie in six clusters: four well-defined clusters and two subclusters. Furthermore, while previously, a sizable fraction of samples in between clusters was ignored, the 2-way model characterized the entire distribution. Using 2-way k-means, we can tell that a large portion of the previously unclustered samples, which lie in between two clusters, contains shared properties. In addition, we see that certain clusters have mixed relationships, while others have almost no interaction.

5. Further Research

In addition, this paper leaves several open questions and opportunities for further research:

(i) How can we efficiently characterize a 2-way distribution with nonspherical covariance matrices?

(ii) How can we efficiently characterize a k-way distribution?

(iii) How can we efficiently characterize a 2-way distribution with nonlinear paths between cluster representatives?

Addressing these questions will further help us understand the composition of microbial populations.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Science Foundation under CISE EAGER Grant no. 1547120.

References

[1] Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[2] J. Qin, R. Li, J. Raes et al., "A human gut microbial gene catalog established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[3] A. Ramette and P. L. Buttigieg, "A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses," *FEMS Microbiology Ecology*, vol. 90, no. 3, pp. 543–550, 2014.

[4] Human Microbiome Project Consortium, "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, pp. 215–221, 2012.

[5] J. M. Fettweis, J. P. Brooks, M. G. Serrano et al., "Differences in vaginal microbiome in African American women versus women of European ancestry," *Microbiology*, vol. 160, Part 10, pp. 2272–2282, 2014.

[6] R. Plenge, M. Weinblatt, N. Shadick, A. Price, N. Patterson, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[7] J. Novembre, D. H. Alexander, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," *Genome Research*, vol. 19, no. 9, pp. 1655–1664, 2009.

[8] J. Novembre, T. Johnson, K. Bryc et al., "Genes mirror geography within Europe," *Nature*, vol. 456, no. 7219, p. 274, 2008.

[9] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[10] National Center for Biotechnology Information, 2014, https://www.ncbi.nlm.nih.gov/sra/?term=SRP002462.

[11] P. D. Schloss, S. L. Westcott, T. Ryabin et al., "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009, http://aem.asm.org/content/75/23/7537.short?rss=1&ssource=mfc.

[12] P. D. Schloss, D. Gevers, and S. L. Westcott, "Reducing the effects of PCR amplification and sequencing artifacts on 16s rRNA-based studies," *PLoS One*, vol. 6, no. 12, article e27310, 2011.

[13] W. J. Jackson, "2-way cluster assignment," 2016, https://github.com/westonjackson/2-Way-Cluster-Assignment.