Biometrics, 2025, 81(3), ujaf125
https://doi.org/10.1093/biomtc/ujaf125
Biometric Practice

OXFORD

A Bayesian semiparametric mixture model for clustering zero-inflated microbiome data

Suppapat Korsurat and Matthew D. Koslovsky *

Department of Statistics, Colorado State University, Fort Collins, CO, 80523, United States

*Corresponding author: Matthew D. Koslovsky, Department of Statistics, Colorado State University, 851 Oval Dr., Fort Collins, CO, 80523, USA (matt.koslovsky@colostate.edu).

ABSTRACT

Microbiome research has immense potential for unlocking insights into human health and disease. A common goal in human microbiome research is identifying subgroups of individuals with similar microbial composition that may be linked to specific health states or environmental exposures. However, existing clustering methods are often not equipped to accommodate the complex structure of microbiome data and typically make limiting assumptions regarding the number of clusters in the data which can bias inference. Designed for zero-inflated multivariate compositional count data collected in microbiome research, we propose a novel Bayesian semiparametric mixture modeling framework that simultaneously learns the number of clusters in the data while performing cluster allocation. In simulation, we demonstrate the clustering performance of our method compared to distance- and model-based alternatives and the importance of accommodating zero-inflation when present in the data. We then apply the model to identify clusters in microbiome data collected in a study designed to investigate the relation between gut microbial composition and enteric diarrheal disease.

KEYWORDS: compositional data; enterotypes; mixture models; multivariate count data.

1 INTRODUCTION

The human microbiome is a complex ecosystem of microorganisms residing within and on our bodies. Each individual possesses a unique microbiome pattern which is influenced by various external factors, such as environment, climate, diet, and medical conditions (Ursell et al., 2012; Rojo et al., 2017; Allaband et al., 2019; Ahn and Hayes, 2021). A common goal in human microbiome research is identifying subgroups of individuals with similar microbial composition, referred to as enterotypes, that may be linked to specific health states or environmental exposures (Arumugam et al., 2011; Holmes et al., 2012; Marcos-Zambrano et al., 2021). For example, this research was motivated by data collected in a study investigating the relation between intestinal microbial community composition and enteric infection (Singh et al., 2015). Given the critical role intestinal microbiota play in maintaining a healthy immune response, there is considerable interest in uncovering patterns in microbial composition to investigate the feasibility of microbiota-based diagnostics, therapies, or prevention of disease, potentially through personalized treatment strategies (Sekirov and Finlay, 2009; Costea et al., 2018).

Typically, microbiome data take the form of an $N \times J$-dimensional matrix of counts, where N represents the number of observations and J represents the number of unique microbial taxa. The conventional approach for obtaining taxa counts is to sequence the 16S rRNA gene, as it contains well-conserved and hypervariable regions to differentiate different species. Then, sequenced reads are clustered into operational taxonomic units, or OTUs, and classified to a reference database using various methods (Huson et al., 2007; Wang et al., 2007; Edgar, 2013; Allard et al., 2015; Bolyen et al., 2019). More recently, researchers have promoted the use of amplicon sequence variants instead of OTUs as the unit of analysis in microbiome research, which can distinguish sequence variants differing by as little as one nucleotide (Eren et al., 2013; Callahan et al., 2016, 2017; Amir et al., 2017). Regardless of the approach taken, these data are inherently challenging to analyze due to their high-dimensionality, overdispersion, compositional structure, and zero-inflation.

Zero reads in microbiome data occur when (1) the organism is not present in the sampling region, and therefore the probability of occurrence is 0 (i.e., structural zero), or (2) the organism is present but was not sampled (i.e., at-risk zero). To model zero-inflation, researchers typical construct a two-component mixture of a point mass at 0 and a sampling distribution for the counts where a latent at-risk indicator is introduced to differentiate between at-risk and structural zeros (Xu et al., 2015; Neelon, 2019; Zhang and Yi, 2020; Shuler et al., 2021; Jiang et al., 2023; Koslovsky, 2023; 2025). While numerous methods have been developed to accommodate zero-inflation in microbiome research, including network analysis (Ha et al., 2020), dimension reduction (Koslovsky, 2023; Zeng et al., 2021), regression modeling (Tang and Chen, 2019), longitudinal data analysis (Zhang et al., 2020; Zhang and Yi, 2020), association tests (Ling et al., 2021), and causal inference (Yang and Xu, 2023), among others, existing methods for cluster analysis often are not equipped to accommodate the complex structure of the data and

Received: September 23, 2024; Revised: June 5, 2025; Accepted: August 11, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of The International Biometric Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

may make limiting assumptions regarding the number of clusters in the data which can bias inference.

Two common approaches for identifying clusters of microbial samples are distance-based methods [e.g., K-means, partitioning around the median (PAM), and hierarchical clustering (Xu and Tian, 2015; Shi et al., 2022)] and model-based methods (Holmes et al., 2012; Subedi et al., 2020; Mao and Ma, 2022; Shi et al., 2023). Distance-based methods use the distance between 2 observations to determine cluster allocation without assuming a statistical distribution for the observed counts. A key challenge in applying distance-based clustering methods to microbiome data is choosing the proper distance metric to capture the similarity or dissimilarity of microbial communities, often referred to as β-diversity (Namkung, 2020). Several β-diversity metrics have been proposed, including Aitchison, Bray–Curtis, and UniFrac distances (Aitchison et al., 2000; Chen et al., 2012). See Plantinga and Wu (2021) for a detailed review of common β-diversity metrics used in microbiome data analysis and Shi et al. (2022) for an extensive simulation study investigating how clustering performance can depend on the chosen β-diversity metric and clustering algorithm.

Model-based clustering methods are a popular alternative to distance-based methods that assume the population consists of observations from distinct clusters, each with their own statistical distribution. Finite mixture models (FMMs), which belong to the class of model-based clustering methods, assume that the population consists of observations originating from a finite number of underlying clusters. In mixture modeling, the concept of determining the number of clusters in the data lies in the distinction between K, the number of components in the model (i.e., the number of potential clusters), and K_+, the number of clusters that are actually present in the data (i.e., non-empty components) (Miller and Harrison, 2018). Typically the number of clusters in FMMs is specified prior to analysis (i.e., $K_+ = K$), and model comparison techniques are used to select K (Yeung et al., 2001; Marin et al., 2005). Alternatively, mixture of finite mixtures (MFM) models place a prior on the number of mixture components to draw inference on the number of clusters in the data (Miller and Harrison, 2018). To relax assumptions on the number of clusters in the data, researchers commonly use infinite mixture models (IMMs), such as Dirichlet process mixture models (DPMMs), that allow $K = \infty$ (McAuliffe et al., 2006; Li et al., 2019). While IMMs do not require specifying the number of clusters a priori, they have been shown to overestimate the true number of clusters, often leading to the formation of numerous singleton clusters (Miller and Harrison, 2013; Li et al., 2019; Frühwirth-Schnatter et al., 2021). Notably, Ascolani et al. (2023) recently showed the posterior for the number of clusters in DPMMs is consistent under certain conditions for the prior on the concentration parameter.

Sparse finite mixture models (sFMMs) were recently introduced as a semiparametric alternative for model-based clustering that bridges standard FMMs and IMMs and is recommended when the number of clusters is not expected to increase with larger sample sizes (Malsiner-Walli et al., 2016; Frühwirth-Schnatter and Malsiner-Walli, 2019). This approach deliberately overspecifies the number of potential components in the model (i.e., $K > K_+$) and then places a sparsity-inducing prior on the mixture weights that shrinks them toward 0 to encourage empty components. One of the challenges of implementing sFMMs is that clustering performance has been shown to depend heavily on the chosen value of the mixture weights, or prior thereof (Celeux et al., 2019; Frühwirth-Schnatter and Malsiner-Walli, 2019), and none of the empty components' probabilities are set exactly to 0.

In this work, we develop a Bayesian model-based clustering method for zero-inflated microbiome data. Our approach belongs to the class of MFM models and can be thought of as a *discrete* alternative to existing sFMMs that similarly overspecifies the number of potential clusters in the data. However, instead of shrinking empty components' mixture weights toward 0, the proposed method places a point mass at 0 to remove empty components from the model. Commonly, the Dirichlet-multinomial (DM) distribution and its extensions are used to model microbial counts and their corresponding relative abundances as it inherently accommodates the compositional structure of the data and overdispersion (Holmes et al., 2012; Wadsworth et al., 2017; Harrison et al., 2020). In this work, we model the multivariate count data collected in the application study with a zero-inflated DM (ZIDM) model which further accounts for excess zeros typically found in microbiome data. As such, our modeling approach extends the work of Koslovsky (2023) to account for heterogeneity in zero-inflated multivariate count data, effectively using a ZIDM distribution to model zero-inflation in the count data as well as induce sparsity in the mixture weights. Together, our approach accommodates the complex structure of microbiome data observed in practice, simultaneously estimates cluster-specific taxa relative abundances while performing cluster allocation, and does not require specifying the number of clusters in the data a priori. For inference, we take a fully Bayesian approach and implement a telescoping sampler (Frühwirth-Schnatter et al., 2021) to help improve the mixing of the Markov chain Monte Carlo (MCMC) algorithm. We demonstrate how the proposed method outperforms alternative approaches for clustering microbiome data in a variety of simulation scenarios. In the application study, we identify 2 main clusters of microbial composition; one dominated by a combination of *Bacteroides* and *Cronobacter* and composed mostly of enteric diarrheal disease (EDD) patients, and another dominated by *Bacteroides* with a balance between healthy individuals and EDD patients.

2 METHODS

In this section, we first introduce standard notation and definitions used for model-based cluster analysis in the context of multivariate count data. Thereafter, we propose a novel discrete sparse Dirichlet-multinomial mixture model (DSDM^3) that we use to simultaneously accommodate uncertainty in the number of clusters in the data while performing cluster allocation. Let $\boldsymbol{z}_i$ represent the J-dimensional vector of taxa counts for the i^{th} individual, $i = 1, 2, \ldots, N$. We assume

$$\boldsymbol{z}_i \mid c_i = k, \boldsymbol{\Theta}_k \sim F\left(\boldsymbol{\Theta}_k\right), \tag{1}$$

where $c_i = k$ indicates the ith individual is assigned to the kth, $k = 1, 2, \ldots, K$, cluster, K represents the number of

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

components, $F(\cdot)$ is the assumed probability mass function for the multivariate count data, and $\boldsymbol{\Theta}_k$ is a multivariate set of parameters for the k^{th} component, both of which we describe in more detail below. Equivalently, we can formulate a mixture model as $f(\boldsymbol{z}_i) = \sum_{k=1}^{K} w_k F(\boldsymbol{\Theta}_k)$, where w_k are the mixture weights that sum to one with $\boldsymbol{w} = (w_1, \ldots, w_K)$. It is then common to assume $\boldsymbol{w} \mid K \sim \text{Dirichlet}(\psi_1, \ldots, \psi_K)$, where $\psi_k \equiv \psi$, for $k = 1, \ldots, K$. Unlike FMMs in which K is fixed a priori, MFM modeling frameworks place a prior distribution on the total number of cluster components. Common prior assumptions for K include the truncated Poisson and more recently a beta-negative-binomial, which generalizes the Poisson, negative-binomial, and geometric distributions (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021).

2.1 A semiparametric mixture model

We propose a DSDM^3 that deliberately overspecifies the number of potential components and induces sparsity in the number of components by allowing empty components' corresponding mixture weights to take on a 0 value. As such, our approach draws similarities to sFMMs that shrink mixture weights *toward* 0 and also belongs to the class of MFM models, which we show below. Specifically, we assume the ith, $i = 1, \ldots, N$, individual's cluster assignment

$$\begin{aligned} c_i \mid \boldsymbol{w} &\sim \text{Multinomial}(1, \boldsymbol{w}), \\ w_k &= \frac{\psi_k}{\sum_{k'=1}^{K_m} \psi_{k'}}, \\ \psi_k \mid \lambda_k &\sim \lambda_k \text{Gamma}(\theta, 1) + (1 - \lambda_k)\, \delta_0(\psi_k), \text{ and} \\ \lambda_k &\sim \text{Bernoulli}(\pi_\lambda), \end{aligned} \tag{2}$$

with the constraint that at least 1 $\lambda_k = 1, k = 1, \ldots, K_m$, where K_m is the maximum number of potential components, $\lambda_k \in \{0, 1\}$ indicates whether or not the kth component exists in the model, θ is a hyperparameter controlling the mixture weights for active components, $\delta_0(\cdot)$ is a Dirac delta function, or point mass, at 0, and the hyperparameter π_λ represents the probability that the kth component exists in the model. This formulation ensures that the probability of being assigned to the kth cluster is 0 when λ_k equals 0. Ideally, K_m should be set large enough to accommodate the true number of clusters in the data but not so large that it leads to excessive computations.

As mentioned previously, the DSDM^3 belongs to the class of MFM models. To make this connection, we first note that given the set of $\lambda_k = 1$, Equation 2 is equivalent to the standard Dirichlet-multinomial formulation used in mixture models with $K = \sum_{k'=1}^{K_m} \lambda_{k'}$ components. Then under the assumptions for λ_k described above, the implied distribution for K is a zero-truncated binomial distribution with $p(K = k) = \frac{\binom{K_m}{k} \pi_\lambda^k (1-\pi_\lambda)^{K_m - k}}{1 - (1-\pi_\lambda)^{K_m}}$. In the Supplementary Material, we provide more details of this connection, derivations for the induced distribution on K_+, the number of active clusters in the model (i.e., the number of non-empty components), and the corresponding exchangeable partition probability function.

2.2 Zero-inflated multivariate count data

In this work, we apply the proposed method to cluster zero-inflated multivariate compositional count data collected in human microbiome research. The DM distribution is commonly used to model microbial counts as it accommodates the compositional structure of the data and overdispersion (Wadsworth et al., 2017; Harrison et al., 2020; Koslovsky et al., 2020; Pedone et al., 2023; Shi et al., 2023). However, naively assuming a DM for microbial counts can bias parameter estimates as it is not inherently equipped to handle zero-inflation (Koslovsky, 2023). To accommodate potential zero-inflation, we assume a ZIDM model for the multivariate compositional counts. Specifically, we let

$$\begin{aligned} \boldsymbol{z}_i \mid \boldsymbol{\phi}_i &\sim \text{Multinomial}\left(\sum_{j=1}^{J} z_{ij}, \frac{\boldsymbol{\phi}_i}{\sum_{j'=1}^{J} \phi_{ij'}}\right), \\ \phi_{ij} \mid c_i = k, \gamma_{ij}, \xi_{kj} &\sim \gamma_{ij} \text{Gamma}\left(\exp(\xi_{kj}), 1\right) \\ &\quad + (1 - \gamma_{ij}) \delta_0(\phi_{ij}), \text{ and} \\ \gamma_{ij} &\sim \text{Bernoulli}(\pi_{\gamma_j}), \end{aligned} \tag{3}$$

where $\phi_{ij} / \sum_{j'=1}^{J} \phi_{ij'}$ is the relative abundance of the jth, $j = 1, \ldots, J$, taxon for the ith, $i = 1, \ldots, N$, observation, γ_{ij} represents an at-risk indicator, and $\pi_{\gamma_j} \sim \text{Beta}(\alpha_\gamma, \beta_\gamma)$ is the probability of an at-risk observation for the jth taxon. The hyperparameters α_γ and β_γ control the probability of an at-risk observation, where $\frac{\alpha_\gamma}{\alpha_\gamma + \beta_\gamma}$ is the expected probability a priori. For zero counts (i.e., $z_{ij} = 0$), $\gamma_{ij} = 1$ indicates an at-risk zero and $\gamma_{ij} = 0$ a structure zero. Note that $\gamma_{ij} = 1$ for $z_{ij} > 0$. The cluster-specific concentration parameters, ξ_{kj}, govern the corresponding relative abundances and are assumed to follow a Normal (μ_j, σ^2). Given the dimension of the parameter space found in microbiome applications, we recommend setting $\mu_j = \log(s * \bar{RA}_j)$, where s is a scaling parameter and $\bar{RA}_j$ is the average relative abundance for the jth taxon observed in the data. Setting s large and σ^2 small will result in the prior for the relative abundances concentrating around the observed mean. For interpretation, s can be thought of as the hypothetical total number of reads used to inform the prior.

2.3 Posterior sampling and inference

The full joint posterior distribution of the proposed DSDM^3-ZIDM is written as

$$\begin{aligned} p(\boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\pi}_\gamma, \boldsymbol{\psi}, \boldsymbol{\lambda} \mid \boldsymbol{z}) &\propto p(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\pi}_\gamma, \boldsymbol{\psi}, \boldsymbol{\lambda}) \\ &= p(\boldsymbol{z} \mid \boldsymbol{\phi})\, p(\boldsymbol{\phi} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\xi})\, p(\boldsymbol{\xi}) \\ &\quad \times p(\boldsymbol{\gamma} \mid \boldsymbol{\pi}_\gamma)\, p(\boldsymbol{\pi}_\gamma)\, p(\boldsymbol{c} \mid \boldsymbol{\psi}) \\ &\quad \times p(\boldsymbol{\psi} \mid \boldsymbol{\lambda})\, p(\boldsymbol{\lambda}), \end{aligned}$$

where $\boldsymbol{z}$ represents the $N \times J$ matrix of counts observed across individuals, $\boldsymbol{\phi}$ the $N \times J$ matrix of individual-specific relative abundances, $\boldsymbol{c}$ the N-dimensional vector of cluster assignments, $\boldsymbol{\gamma}$ the $N \times J$ matrix of at-risk indicators, $\boldsymbol{\xi}$ the $K_m \times J$ matrix of cluster-specific concentration parameters, $\boldsymbol{\psi}$ the K_m-dimensional vector of potential cluster weights, $\boldsymbol{\pi}_\gamma$ the J-dimensional vector of at-risk probabilities, and $\boldsymbol{\lambda}$ the

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

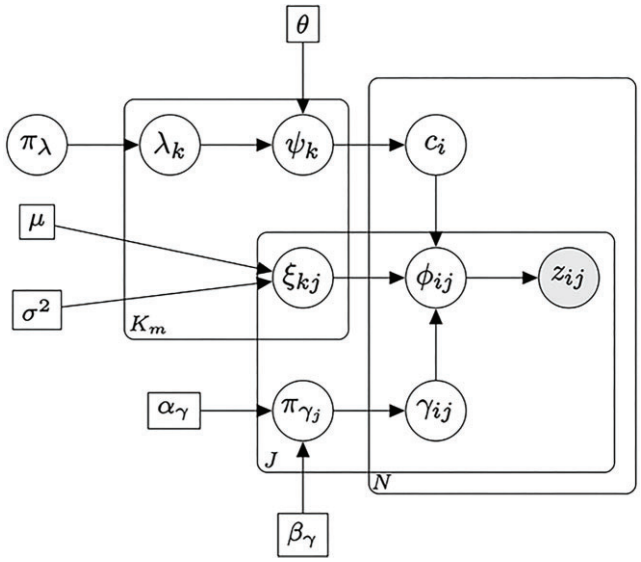

FIGURE 1 Graphical representation of the proposed clustering framework for zero-inflated multivariate compositional count data. The white (shaded) nodes represent the parameters (observed data). The rectangular (circular) shapes represent fixed (random) variables.

K_m-dimensional vector of indicators for the cluster components. Figure 1 presents a graphical representation of the proposed model in the context of zero-inflated multivariate count data.

To generate samples from the posterior distribution, we construct a Metropolis–Hastings within Gibbs algorithm. In practice, we are typically only interested in drawing inference on a subset of the parameters in the model. Therefore, we integrate out $\boldsymbol{\phi}$ and π_{γ_j} to reduce the computational cost of the resulting MCMC sampler. Following Miller and Harrison (2018), we could implement Algorithm 8 of Neal (2000) coupled with a Split-Merge sampler (Jain and Neal, 2007) to update the cluster allocations, $\boldsymbol{c}$. However, due to the dimension of the cluster-specific parameters, $\boldsymbol{\xi}_k$, we observed poor mixing for the number of active clusters. Instead, we implemented the telescoping sampler proposed by Frühwirth-Schnatter et al. (2021), in which the full conditional distribution of K, and thus the creation of a new component, does not depend on the cluster parameters. As a result, we observed improved mixing and convergence of the MCMC sampler overall. More technical details of the MCMC algorithm and relevant derivations are provided in the Supplementary Material. To determine the final cluster assignment for each observation, we applied the `salso` algorithm (Dahl et al., 2022), a randomized greedy search algorithm, to minimize the lower bound of the variation of information loss (Meilă, 2003).

3 SIMULATION STUDY

Before applying the model to data collected in the EDD study, we first performed a simulation study to evaluate the clustering performance of the proposed method and compare it to alternative distance- and model-based methods for clustering multivariate compositional count data in 6 different scenarios. Specifically, we compared the proposed DSDM^3 to a Dirichlet process (DP) mixture model with a Dirichlet-tree multinomial distribution for the counts (Mao and Ma, 2022) and a shrinkage-based sparse Dirichlet-multinomial mixture model (sSDM^3), similar to Saraiva et al. (2020), with a ZIDM for the counts (DSDM^3-ZIDM, DP-DTM, and sSDM^3-ZIDM, respectively). Additionally, we compared the model to a finite Dirichlet-multinomial mixture model (DM^3), similar to Holmes et al. (2012), implemented with a ZIDM for the counts (DM^3-ZIDM). To assess how ignoring zero-inflation may affect inference, we also evaluated the clustering performance of the DSDM^3 and sSDM^3 with a DM likelihood for the counts (DSDM^3-DM and sSDM^3-DM, respectively). Further, we compared the model to the distance-based clustering method PAM (Rousseeuw and Kaufman, 1990), using various distance metrics: Aitchison (PAM-AT) (Aitchison et al., 2000), Bray–Curtis (PAM-BC) (Bray and Curtis, 1957), unweighted UniFrac (PAM-UU), weighted UniFrac (PAM-WU) (Lozupone and Knight, 2005). Note that DP-DTM, PAM-UU, and PAM-WU require specifying a phylogenetic tree for the microbial count data. Lastly, we compared to a Gaussian mixture model (GMM) with an additive log-ratio transformation for the compositional count data (Egozcue et al., 2003; Scrucca et al., 2023).

For the Bayesian mixture models, we assumed similar prior assumptions when applicable (see the Supplementary Material for more details). Each model was run for 10 000 MCMC iterations, treating the first 5000 iterations as burn-in. Final cluster assignments for the Bayesian models were obtained using the `salso` algorithm. To determine the number of clusters using the finite mixture models (i.e., DM^3-ZIDM and GMM) and distance-based methods (i.e., PAM-AT, PAM-BC, PAM-UU, and PAM-WU), we fit each model with 2–10 different components and selected a final model for inference using the Bayesian information criterion [BIC; (Swartz et al., 2004)] and average silhouette width, respectively. Clustering performance of the methods was evaluated using the adjusted Rand index [ARI; (Hubert and Arabie, 1985)], where higher values imply better clustering.

In scenarios 1–5, we generated data to evaluate and compare the methods' clustering performance with varying percentages of 0 cell counts, numbers of taxa, and numbers of clusters using an approach similar to that taken in Shi et al. (2023). In each scenario, we first specify the number of observations, N, the number of taxa which are not differentiated by cluster assignment, J_{noise}, the number of taxa that are cluster-specific, J_{signal}, the total depth for the noise taxa, $\dot{z}_{\text{noise}}$, the total depth for the signal taxa, $\dot{z}_{\text{signal}}$, and the expected proportion of at-risk observations in each data set. In scenarios 1, 2, and 3, we generated 2 true clusters and 100 taxa with increasing percentages of 0 cell counts in the data (i.e., 28%, 51%, and 73%, respectively). The data in scenarios 4 and 5 were simulated similar to scenario 2; however, in scenario 4, we increased the number of taxa to 250 and in scenario 5 the number of true clusters to 6. For scenarios 1–5, the sequencing depth was set to 5000 reads per sample (4000 for noise taxa and 1000 for signal taxa). Additionally, we compared the models in a setting designed with a more complex correlation structure among the counts (scenario 6). Specifically, multivariate count data were generated using a Dirichlet-tree multinomial distribution with added zero-inflation. The phylogenetic structure used to simu-

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

TABLE 1 Summary of the simulation study scenarios.

Scenario	K	N $(N_1, \ldots, N_K)$	J $(J_{\text{noise}}, J_{\text{signal}})$	Average proportion of 0 counts
1	2	100 (50, 50)	100 (80, 20)	0.28
2	2	100 (50, 50)	100 (80, 20)	0.51
3	2	100 (50, 50)	100 (80, 20)	0.73
4	2	100 (50, 50)	250 (200, 50)	0.54
5	6	150 (30, 30, 20, 20, 25, 25)	100 (80, 20)	0.50
6	4	300 (75, 75, 75, 75)	79∗	0.58

Each scenario outlines different configurations for true number of clusters, K, sample size, N, with N_k representing the number of observations in each true cluster in parentheses, the total number of taxa, J, comprising J_{noise} taxa that were undifferentiated across clusters and J_{signal} taxa that differentiated the clusters, and the average proportion of 0 counts computed over 20 replicate datasets. ∗—see the Supplementary Material for more details of the data generation in this scenario.

TABLE 2 Simulation results: average adjusted Rand index (ARI) with standard deviations in parentheses for all methods across 20 replicate data sets.

	Scenario					
Model	**1**	**2**	**3**	**4**	**5**	**6**
DSDM^3-ZIDM	0.91 (0.09)	0.84 (0.15)	0.39 (0.40)	0.90 (0.11)	0.70 (0.13)	0.86 (0.04)
sSDM^3-ZIDM	0.90 (0.31)	0.70 (0.47)	0.14 (0.31)	0.90 (0.23)	0.34 (0.15)	0.86 (0.03)
DM^3-ZIDM	1.00 (0.00)	1.00 (0.02)	0.36 (0.41)	0.88 (0.31)	0.59 (0.08)	0.52 (0.26)
DSDM^3-DM	0.05 (0.22)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)
sSDM^3-DM	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.05)
DP-DTM	0.00 (0.02)	0.00 (0.02)	0.00 (0.01)	0.00 (0.01)	0.01 (0.02)	0.00 (0.00)
PAM-BC	0.76 (0.18)	0.42 (0.21)	0.26 (0.15)	0.38 (0.27)	0.43 (0.08)	0.06 (0.08)
PAM-AT	0.01 (0.02)	0.01 (0.02)	0.07 (0.09)	0.02 (0.05)	0.03 (0.02)	0.01 (0.01)
PAM-UU	0.01 (0.04)	0.01 (0.02)	0.02 (0.03)	0.00 (0.01)	0.03 (0.02)	0.04 (0.03)
PAM-WU	0.76 (0.18)	0.42 (0.21)	0.26 (0.15)	0.39 (0.26)	0.43 (0.08)	0.37 (0.06)
GMM	0.00 (0.03)	0.01 (0.03)	0.00 (0.02)	-0.01 (0.02)	0.01 (0.02)	0.01 (0.01)

Higher values of ARI represent better clustering performance.

late these data was obtained from the taxonomy observed in the application study. Additionally, we set the sequencing depth to match the median depth of the EDD dataset (i.e., 2500 reads per sample). See Table 1 and the Supplementary Material for more details of each scenario, including a plot of the phylogenetic tree used to simulate data in scenario 6.

Table 2 presents the clustering performance results of the methods in each simulation scenario. Overall, the proposed DSDM^3-ZIDM obtained similar or improved clustering performance in terms of the average ARI compared to the alternative distance- and model-based approaches in all 6 scenarios, with the exception of DM^3-ZIDM in scenarios 1 and 2. Recall that DM^3-ZIDM is essentially the same model as the proposed DSDM^3-ZIDM; however, the number of components K are fixed a priori. As the percentage of 0 cell counts increased across scenarios 1–3, the clustering performance of all methods declined. The proposed method's performance was relatively robust to the number of taxa, J. However, we observed a slight decrease in performance in the scenario with 6 true clusters (i.e., scenario 5). In scenario 6, which was designed to incorporate a more complex correlation structure among the counts, DSDM^3-ZIDM and sSDM^3-ZIDM outperformed all other methods, including PAM-UU, PAM-WU, and DP-DTM, which were the only methods that incorporated phylogenetic information when performing clustering. In all 6 scenarios, we observed that versions of the models that accommodated zero-inflation outperformed those that ignored zero-inflation, regardless of the clustering framework used. Among the distance-based and frequentist methods, PAM-WU and PAM-BC demonstrated the best clustering performance. Notably, PAM-WU and PAM-BC are equivalent in scenarios 1–5, as all simulated samples share the same sequencing depth and no phylogenetic structure is assumed among the OTUs (see the Supplementary Material for more details). In scenario 6, PAM-WU outperformed PAM-BC given the additional information contained in the phylogenetic tree. Lastly, we compared the computation time for each of the methods. DSDM^3-ZIDM took roughly 15 min in scenarios 1, 2, and 3; 130 min in scenario 4; and 45 min in scenarios 5 and 6 to run 10 000 MCMC iterations on a MacBook Pro with an Apple M1 Pro chip (8-core CPU: 6 performance and 2 efficiency cores), 16 GB RAM, and macOS Sequoia 15.4 (Table 3). The alternative Bayesian methods were typically able to run 10 000 iterations quicker than the proposed method in scenarios 1–5, especially for those that ignored potential zero-inflation. However, DM^3-ZIDM, which relies on model comparison techniques to determine the number of clusters in the data, obtained a cumulative computation time nearly 10 times that of the proposed method. While DP-DTM was the quickest Bayesian method in scenarios 1–4, it slowed considerably as the complexity of the data increased in scenarios 5 and 6. The frequentist and distance-based approaches were much faster than the Bayesian methods, as expected.

4 ZERO-INFLATED MICROBIOME DATA APPLICATION

In this section, we apply our proposed clustering method to microbiome data collected in a study investigating the relation between gut microbial composition and EDD (Singh et al., 2015).

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

TABLE 3 Simulation Results: Average computation time in minutes with standard deviations in parentheses for all methods across 20 replicate data sets.

	Scenario					
Model	**1**	**2**	**3**	**4**	**5**	**6**
DSDM^3-ZIDM	16.52 (1.17)	12.61 (0.71)	14.54 (0.62)	130.85 (9.79)	44.79 (3.04)	44.20 (2.19)
sSDM^3-ZIDM	16.51 (1.12)	13.19 (1.05)	32.65 (2.93)	132.67 (9.53)	44.47 (3.34)	84.14 (5.58)
DM^3-ZIDM	153.80 (1.79)	116.55 (1.07)	134.63 (1.48)	1217.04 (19.81)	358.83 (11.69)	368.57 (1.87)
DSDM^3-DM	13.07 (0.91)	8.90 (0.61)	8.67 (0.38)	89.16 (5.51)	32.67 (1.68)	25.98 (1.87)
sSDM^3-DM	13.51 (0.97)	9.11 (0.71)	9.01 (0.37)	89.68 (6.25)	31.44 (1.37)	51.50 (3.82)
DP-DTM	10.21 (2.83)	7.07 (1.40)	6.16 (1.22)	11.86 (3.69)	65.14 (9.45)	252.58 (25.48)
PAM-BC	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
PAM-AT	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
PAM-UU	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.12 (0.02)	0.25 (0.01)	0.83 (0.09)
PAM-WU	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.12 (0.02)	0.27 (0.06)	0.82 (0.09)
GMM	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.01)	0.04 (0.01)	0.07 (0.01)

For models that rely on model comparison techniques to determine the number of clusters, we report the cumulative computation time for each model fit.

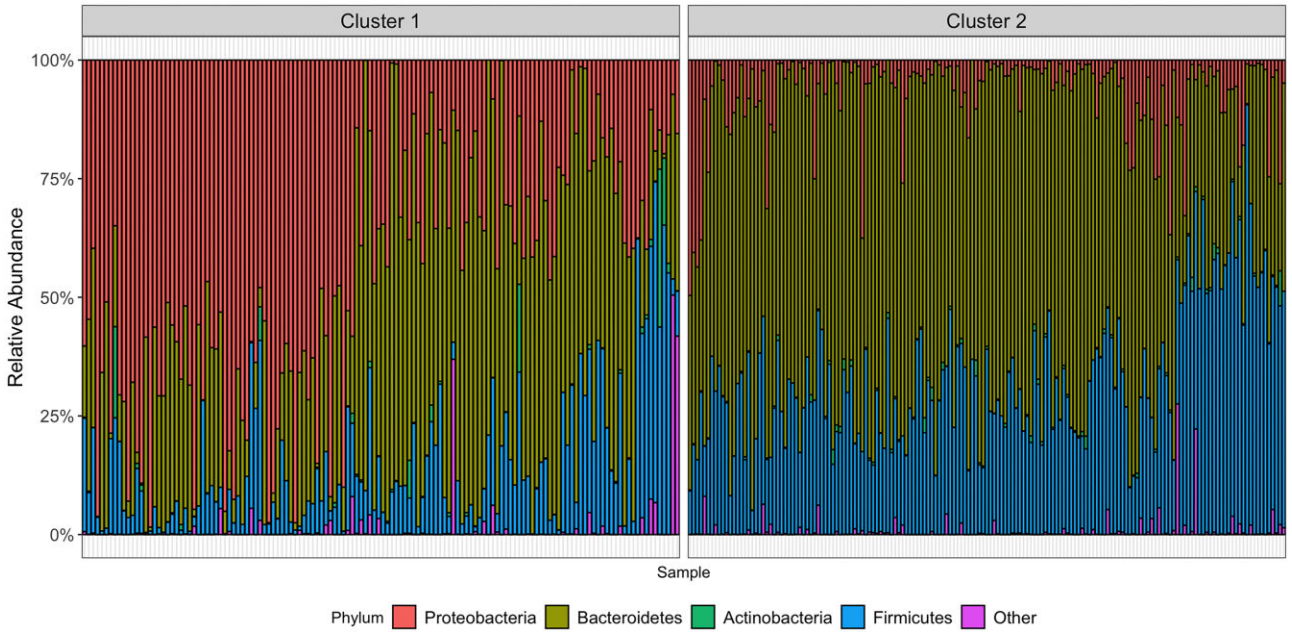

FIGURE 2 EDD Application Results: Observed relative abundances of each patient in the 2 main clusters obtained with DSDM^3-ZIDM. Relative abundances are aggregated at the phylum level for ease of presentation.

These data were made available by Duvallet et al. (2017) as part of a meta-analysis for gut microbiome studies. Following the data processing pipeline described in Duvallet et al. (2017), the data investigated in this analysis consisted of $N = 303$ observations (221 EDD patients and 82 healthy controls) of $J = 79$ taxa at the genus level with 58% zero counts.

For the proposed DSDM^3-ZIDM, we set $K_m = 10$, $\theta = 0.1$, $\alpha_\gamma = 1$, $\beta_\gamma = 1$, $\pi_\lambda = 0.5$, $s = 200$, $\sigma^2 = 10$, and $\sigma^2_{\text{MH}} = 1$. Values of σ^2_{MH} and s were chosen to ensure adequate mixing of the MCMC sampler. The MCMC sampler was initialized at a singleton cluster with $\xi_{1j} = \log(200 * \bar{RA}_j)$ and run for 15 000 iterations. We treated the first 5000 iterations as burn-in and used the remaining 10 000 samples for inference. Convergence was then assessed visually using trace plots for the mixture weights, $\boldsymbol{w}$, and cluster-specific concentration parameters for the relative abundances, $\boldsymbol{\xi}_k$. See the Supplementary Material for more details. Cluster allocation was then determined using the `salso` method to minimize the lower bound of the variation of information loss.

In a recent study assessing gut microbial community composition conducted on 3 large metagenomic datasets, the authors identified 3 enterotypes: one characterized by *Bacteroides*, another by *Prevotella*, and third by Firmicutes with *Ruminococcus* typically the most abundant (Costea et al., 2018). Our analysis identified 2 main clusters in the data; one with genera balanced between phyla Firmicutes (0.40), Proteobacteria (0.28), and Bacteroidetes (0.27) (cluster 1) and the other by Bacteroidetes (0.47) and Firmicutes (0.42) (cluster 2). Posterior estimates of the relative abundances obtained with DSDM^3-ZIDM are in parentheses for reference. At the genus level, cluster 1 was dominated by a combination of *Bacteroides* (0.15) and *Cronobacter* (0.13), while cluster 2 was dominated by *Bacteroides* (0.33).

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

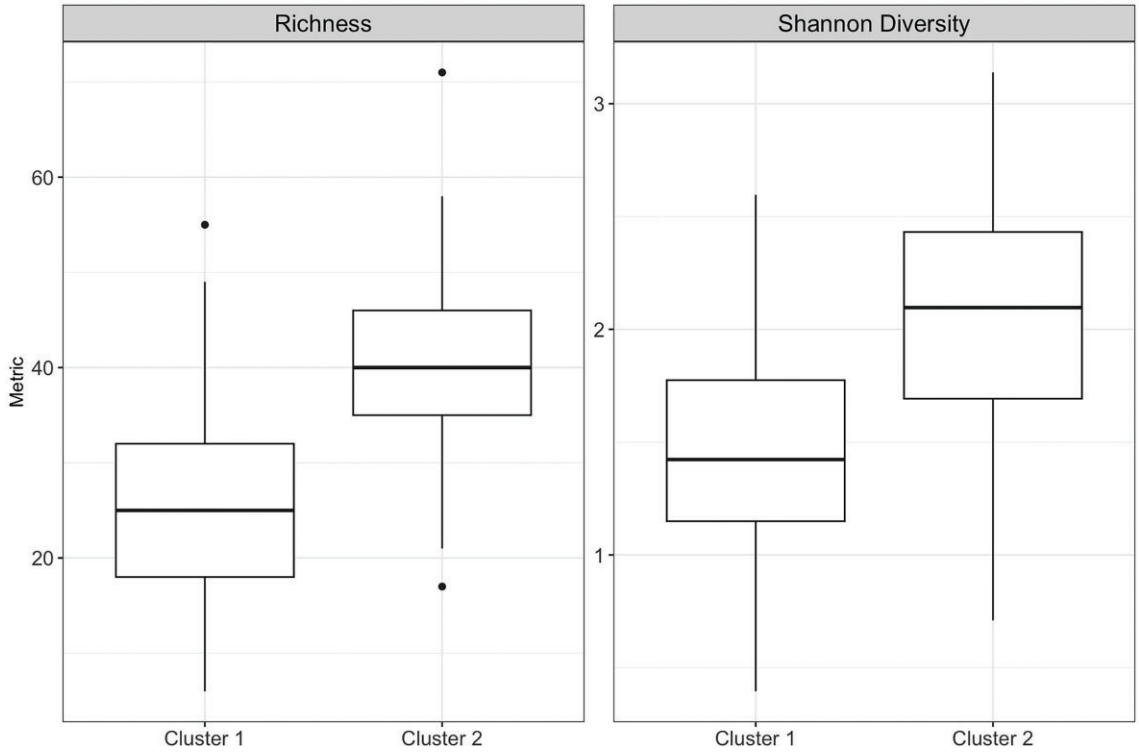

FIGURE 3 EDD Application Results: Boxplots of the richness and diversity of microbial communities across the 2 main clusters obtained with DSDM^3-ZIDM. The left panel depicts the richness (number of different species) for each cluster, while the right panel illustrates the Shannon diversity index (a measure of species diversity accounting for both abundance and evenness).

TABLE 4 Sensitivity Analysis for EDD Application Results: Sensitivity of the results obtained with DSDM^3-ZIDM to hyperparameter specification.

	$\theta = 0.01$	$\theta = 1$	$\sigma^2_{MH} = 0.1$	$\sigma^2_{MH} = 10$	$s = 100$
Clusters	3	1	4	6	4
ARI	0.98	0.00	0.96	0.95	0.65
	$s = 300$	$\beta_\gamma = 4$	$\beta_\gamma = 9$	$\pi_\lambda = 0.2$	$\pi_\lambda = 0.1$
Clusters	5	9	5	4	1
ARI	0.95	0.48	0.44	0.97	0.00

ARI, adjusted Rand index. Clusters, the number of clusters estimated by the model.

Figure 2 presents the observed relative abundances aggregated at the phyla level for ease of presentation for the 2 main clusters. In the Supplementary Material, we additionally present the abundances aggregated at the family level for reference. Of the 136 individuals assigned to cluster 1, 132 were EDD patients and 4 were healthy controls. Whereas, almost half of the 163 individuals assigned to cluster 2 were healthy controls (78/163). We observed that richness and Shannon diversity indices were lower for the cluster dominated by EDD patients (Figure 3), similar to the results presented in Singh et al. (2015). Additionally, the model identified 2 singleton clusters and another cluster of only 2 individuals. Interestingly, the individuals assigned to these clusters were all EDD patients with microbial compositions dominated by *Cronobacter* (> 0.86 in each). Recall that we also observed cluster 1, which was mostly EDD patients, having higher levels of genus *Cronobacter*. *Cronobacter* is a Gram-negative bacterium associated with foodborne diseases and infections (Forsythe, 2018). Previously, researchers have found *Cronobacter* levels increased for patients with colorectal polyps, colorectal cancer, irritable bowel syndrome, autism spectrum disorder, and metabolic syndrome (Wu et al., 2025).

4.1 Sensitivity analyses

In this section, we perform a thorough sensitivity analysis of the application results to hyperparameter specification for DSDM^3-ZIDM. Thereafter, we compare the results obtained in the application study to those obtained with alternative distance- and model-based methods. To assess the model's sensitivity to prior specification, we set each of the hyperparameters to default values (i.e., those used in application study) and then evaluated the effect of manipulating each term on inference. We observed that the clustering results were relatively sensitive to θ, which can be interpreted as the concentration parameter of the Dirichlet distribution on the mixture weights (see Table 4). While the model was robust to smaller values of θ, larger values resulted in the MCMC chain getting stuck at $K = 1$ mixture components.

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

These results were not surprising as previous studies have shown that clustering performance is sensitive to the prior specification of the mixture weights (Celeux et al., 2019; Frühwirth-Schnatter and Malsiner-Walli, 2019). We found that inference was relatively robust to the variance of the proposal distribution for $\boldsymbol{\xi}$. However, the model was moderately sensitive to s, the scaling factor for the prior mean of $\boldsymbol{\xi}$. Specifically, we found that the model had difficulty finding favorable new cluster parameters given their high dimensionality, which resulted in poorer mixing of the MCMC sampler. We observed that reducing the probability of an at-risk observation resulted in an increase in the number of smaller clusters. Lastly, as the prior probability for the number of active components decreased, so did the number of clusters, as expected.

To assess the sensitivity of the application results to model selection and specification, we first reanalyzed the data with alternative MFM models that place different priors on the number of components in the model. Specifically, we fit a MFM model assuming K followed a truncated Poisson (MFM-P), geometric (MFM-G), and beta-negative-binomial (MFM-BNB) distribution with a ZIDM for the count data. The hyperparameters for each model were specified so that the prior mean for the number of components matched DSDM^3-ZIDM. All other hyperparameters were set similar to the application study. Posterior samples were obtained using the telescoping sampler of Frühwirth-Schnatter et al. (2021), and post-hoc inference was performed with the `salso` algorithm. The alternative MFM models all identified 2 main clusters with comparable numbers of healthy individuals and EDD patients and similar microbial compositions, resulting in all methods obtaining > 0.95 ARI compared to the cluster allocation found with the proposed method. Additionally, the MFM-P, MFM-G, and MFM-BNB models identified 1, 2, and 3 singleton clusters of EDD patients with high levels of *Cronobacter*, respectively. Similar cluster allocations were found using the DM^3-ZIDM model. The DP-DTM model was not able identify any clusters in the data. sSDM^3-DM and sSDM^3-ZIDM suggested 8 and 9 clusters with 0.29 and 0.39 ARI compared to the cluster allocation found with the proposed method, respectively. For comparison, we also fit each of the PAM methods using the distance metrics evaluated in the simulation study with 2–10 clusters and then selected the model with higher average silhouette width for inference. While each of the distance-based methods identified 2 clusters in the data (i.e., one dominated by *Cronobacter* and *Bacteroides* and another by *Bacteroides*), the results differed from those found with the proposed method (i.e., ARIs ranging from 0.34 to 0.53). Lastly, we evaluated the sensitivity of the results to likelihood assumptions by applying the DSDM^3 with a DM distribution for the multivariate count data (DSDM^3-DM), which ignores potential zero-inflation. All else equal, DSDM^3-DM was unable to identify any clusters in the data, highlighting the importance of accommodating zero-inflation when clustering microbiome data.

5 DISCUSSION

In this work, we performed a cluster analysis on microbial composition data collected in a study designed to investigate the relation between gut microbial composition and EDD using a novel Bayesian semiparametric clustering method for multivariate compositional count data with zero-inflation. Through sensitivity analysis, we show how inference can depend heavily on decisions regarding the clustering method and its specification, underscoring some of the challenges of clustering microbiome data that are commonly faced by the field (Costea et al., 2018). Our approach uses a discrete sparse mixture modeling framework to determine cluster allocation, which allows the mixture weights of empty cluster components to take on 0 values. As such, it can be seen as an extension to the work of Koslovsky (2023) that additionally accounts for heterogeneity in zero-inflated multivariate count data, effectively using a ZIDM model to accommodate zero-inflation in the count data as well as sparsity in the mixture weights. Additionally, we show how the proposed clustering method belongs to the class of MFM models. This connection opens up a suite of sampling algorithms for posterior inference, as demonstrated in Miller and Harrison (2018) and Frühwirth-Schnatter et al. (2021). In the Supplementary Material, we provide derivations for the MCMC algorithm using the telescoping sampler and those needed to implement our approach with the algorithms described in Miller and Harrison (2018) and Neal (2000). In the simulation study, we find that our model is able to obtain similar or improved clustering performance compared to alternative distance- and model-based clustering methods while properly accounting for zero-inflation in the data.

In the EDD application study, we cluster the data at the genus level. However, in practice, the model is agnostic to the level in which the multivariate count data are aggregated prior to analysis. While designed for zero-inflated multivariate compositional count data collected in human microbiome research settings, the semiparametric clustering framework is flexible to other data structures by adjusting the likelihood function accordingly. Another future extension of the proposed model we aim to explore is how different prior specifications for the active component indicators, λ_k, or hyperpriors for θ, which control the mixing weights, may induce more desirable clustering behavior and/or accommodate available information to improve clustering performance and inference. For example, we may allow the mixture weights to depend on covariate information so that individuals with similar covariate values are more likely to cluster together. Further, the notion of placing a point mass at 0 for mixture weights was also recently proposed in infinite settings via the atom-skipping process and the plaid atoms model for multiple groups (Bi and Ji, 2023). Future work could explore the use of the proposed DSDM^3 in grouped settings as a semiparametric alternative.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 4 are available with this paper at the Biometrics website on Oxford Academic. The case study data, code for generating the simulated data, the `R` package implementing the proposed method, and package vignette, are available at https://github.com/skorsu/DSDM3/tree/main.

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

FUNDING

S.K. and M.D.K. gratefully acknowledge the support of National Science Foundation (NSF) grant DMS-2245492. The opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the NSF.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The case study data from Singh et al. (2015) are available in the MicrobiomeHD database (Duvallet et al., 2017)(https://zenodo.org/records/569601).

REFERENCES

Ahn, J. and Hayes, R. B. (2021). Environmental influences on the human microbiome and implications for noncommunicable disease. *Annual Review of Public Health*, 42, 277–292.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geosciences*, 32, 271–275.

Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A. et al. (2019). Microbiome 101: studying, analyzing, and interpreting gut microbiome data for clinicians. *Clinical Gastroenterology and Hepatology*, 17, 218–230.

Allard, G., Ryan, F. J., Jeffery, I. B. and Claesson, M. J. (2015). SPINGO: A rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, 16, 1–8.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z. et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2, 10–1128.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R. et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473, 174–180.

Ascolani, F., Lijoi, A., Rebaudo, G. and Zanella, G. (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika*, 110, 551–558.

Bi, D. and Ji, Y. (2023). A class of dependent random distributions based on atom skipping. arXiv, arXiv:2304.14954, preprint: not peer reviewed.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A. et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37, 852–857.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27, 326–349.

Callahan, B. J., McMurdie, P. J. and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, 11, 2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.

Celeux, G., Fruewirth-Schnatter, S. and Robert, C. P. (2019). Model selection for mixture models—perspectives and strategies. *Handbook of Mixture Analysis*. 117–154 . New York: Chapman and Hall/CRC.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D. et al. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28, 2106–2113.

Costea, P. I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M. J., Bushman, F. D. et al. (2018). Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3, 8–16.

Dahl, D. B., Johnson, D. J. and Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31, 1189–1201.

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8, 1784.

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10, 996.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geosciences*, 35, 279–300.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G. et al. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4, 1111–1119.

Forsythe, S. J. (2018). Updates on the *Cronobacter* genus. *Annual Review of Food Science and Technology*, 9, 23–44.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13, 33–64.

Frühwirth-Schnatter, S., Malsiner-Walli, G. and Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16, 1279–1307.

Ha, M. J., Kim, J., Galloway-Peña, J., Do, K.-A. and Peterson, C. B. (2020). Compositional zero-inflated network estimation for microbiome data. *BMC Bioinformatics*, 21, 1–20.

Harrison, J. G., Calder, W. J., Shastry, V. and Buerkle, C. A. (2020). Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources*, 20, 481–497.

Holmes, I., Harris, K. and Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One*, 7, 1–15.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17, 377–386.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2, 445–472.

Jiang, R., Zhan, X. and Wang, T. (2023). A flexible zero-inflated Poisson–Gamma model with application to microbiome sequence count data. *Journal of the American Statistical Association*, 118, 792–804.

Koslovsky, M. D. (2023). A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. *Biometrics*, 79, 3239–3251.

Koslovsky, M. D. (2025). Analyzing microbiome data with taxonomic misclassification using a zero-inflated Dirichlet-multinomial model. *BMC Bioinf.*, 26, 69.

Koslovsky, M. D., Hoffman, K. L., Daniel, C. R. and Vannucci, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *The Annals of Applied Statistics*, 14, 1471–1492.

Li, Y., Schofield, E. and Gönen, M. (2019). A tutorial on Dirichlet Process mixture modeling. *Journal of Mathematical Psychology*, 91, 128–144.

Ling, W., Zhao, N., Plantinga, A. M., Launer, L. J., Fodor, A. A., Meyer, K. A. et al. (2021). Powerful and robust non-parametric association testing for microbiome data via a zero-inflated quantile approach (ZINQ). *Microbiome*, 9, 1–19.

Lozupone, C. and Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71, 8228–8235.

Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26, 303–324.

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026

Mao, J. and Ma, L. (2022). Dirichlet-tree multinomial mixtures for clustering microbiome compositions. *The Annals of Applied Statistics*, 16, 1476.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O. et al. (2021). Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*, 12, 634511.

Marin, J.-M., Mengersen, K. and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25, 459–507.

McAuliffe, J. D., Blei, D. M. and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16, 5–14.

Meilă, M. (2003). Comparing clusterings by the variation of information. In: *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, 173–187. Springer.

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *Advances in Neural Information Processing Systems*, 26:1–8.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113, 340–356.

Namkung, J. (2020). Machine learning methods for microbiome studies. *Journal of Microbiology*, 58, 206–216.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249–265.

Neelon, B. (2019). Bayesian zero-inflated negative binomial regression based on Pólya-gamma mixtures. *Bayesian Analysis*, 14, 829.

Pedone, M., Amedei, A. and Stingo, F. C. (2023). Subject-specific Dirichlet-multinomial regression for multi-district microbiota data analysis. *The Annals of Applied Statistics*, 17, 539–559.

Plantinga, A. M. and Wu, M. C. (2021). *Beta Diversity and Distance-Based Analysis of Microbiome Data*, 101–127. Cham: Springer International Publishing.

Rojo, D., Méndez-García, C., Raczkowska, B. A., Bargiela, R., Moya, A., Ferrer, M. et al. (2017). Exploring the human microbiome from multiple perspectives: Factors altering its composition and function. *FEMS Microbiology Reviews*, 41, 453–478.

Rousseeuw, P. J. and Kaufman, L. (1990). Partitioning Around Medoids. In: *Finding Groups in Data*, 68–125. Hoboken, NJ: John Wiley & Sons, Inc.

Saraiva, E. F., Suzuki, A. K. and Milan, L. A. (2020). A Bayesian sparse finite mixture model for clustering data from a heterogeneous population. *Brazilian Journal of Probability and Statistics*, 34, 323–344.

Scrucca, L., Fraley, C., Murphy, T. B. and Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. New York: Chapman and Hall/CRC.

Sekirov, I. and Finlay, B. B. (2009). The role of the intestinal microbiota in enteric infection. *Journal of Physiology* , 587, 4159–4167.

Shi, Y., Zhang, L., Do, K.-A., Jenq, R. and Peterson, C. B. (2023). Sparse tree-based clustering of microbiome data to characterize microbiome heterogeneity in pancreatic cancer. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72, 20–36.

Shi, Y., Zhang, L., Peterson, C. B., Do, K.-A. and Jenq, R. R. (2022). Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome*, 10, 1–12.

Shuler, K., Verbanic, S., Chen, I. A. and Lee, J. (2021). A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70, 961–979.

Singh, P., Teal, T. K., Marsh, T. L., Tiedje, J. M., Mosci, R., Jernigan, K. et al. (2015). Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome*, 3, 1–12.

Subedi, S., Neish, D., Bak, S. and Feng, Z. (2020). Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69, 1163–1187.

Swartz, T. B., Haitovsky, Y., Vexler, A. and Yang, T. Y. (2004). Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics*, 32, 285–302.

Tang, Z.-Z. and Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20, 698–713.

Ursell, L. K., Metcalf, J. L., Parfrey, L. W. and Knight, R. (2012). Defining the human microbiome. *Nutrition Reviews*, 70, S38–S44.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A. and Vannucci, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18, 94.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73, 5261–5267.

Wu, X.-R., He, X.-H. and Xie, Y.-F. (2025). Characteristics of gut microbiota dysbiosis in patients with colorectal polyps. *World Journal of Gastrointestinal Oncology*, 17, 98872.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2, 165–193.

Xu, L., Paterson, A. D., Turpin, W. and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS One*, 10, e0129606.

Yang, D. and Xu, W. (2023). Estimation of mediation effect on zero-inflated microbiome mediators. *Mathematics*, 11, 2830.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17, 977–987.

Zeng, Y., Zhao, H. and Wang, T. (2021). Model-based microbiome data ordination: A variational approximation approach. *Journal of Computational and Graphical Statistics*, 30, 1036–1048.

Zhang, X., Guo, B. and Yi, N. (2020). Zero-Inflated Gaussian mixed models for analyzing longitudinal microbiome data. *Plos One*, 15, e0242073.

Zhang, X. and Yi, N. (2020). NBZIMM: Negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*, 21, 1–19.

Received: September 23, 2024; Revised: June 5, 2025; Accepted: August 11, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of The International Biometric Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/biometrics/article/81/3/ujaf125/8262358 by guest on 03 June 2026