Clustering of compositional data: with an application to integrome data

Adeline Denis, Agathe Guilloux, Emmanuelle Six

Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

université évry val-d'essonne
université PARIS-SACLAY

imagine
INSTITUT DES MALADIES GÉNÉTIQUES

Presentation of the data and the problem at hand

Clustering algorithms based on a transformation of profile data

Model based clustering

- Dirichlet multinomial mixture
- Generalized Dirichlet multinomial mixture

Simulations

Presentation of the data and the problem at hand

The count data

An initial dataset composed of IS characterized by their abundance in the 5 cell types (quantified using sonicAbundance measure). In this dataset:

x_i^l is the number of cells of type l ($l \in \{G, M, B, K, T\}$) containing IS i.

We can measure IS abundance by

$$x_i^{\cdot} = \sum_{l \in \{G,M,B,K,T\}} x_i^l$$

IS	Gene	G	M	B	K	T	IS Abund.
chr12+54311322	NFE2	109	19	57	19	23	227
chr3-52103146	POC1A	49	10	0	0	0	59
chr9-136685086	AGPAT2	0	0	2	0	22	24
chr1-8688776	RERE	19	0	0	0	0	19
chr17+7199073	DLG4	1	0	0	0	18	19
other IS		9534	2231	3520	671	7526	23482

Table 1: Example of IS data

The compositional data

We can next define the profile $p_i = (p_i^G, p_i^M, p_i^B, p_i^K, p_i^T)$ of each IS i as

$$p_i^l = \frac{x_i^l}{x_i^\cdot}.$$

IS	Gene	G	M	B	K	T
chr12+54311322	NFE2	0.480	0.084	0.251	0.084	0.101
chr3-52103146	POC1A	0.831	0.170	0.000	0.000	0.00
chr9-136685086	AGPAT2	0.000	0.000	0.083	0.00	0.917
chr1-8688776	RERE	1.000	0.000	0.000	0.000	0.000
chr17+7199073	DLG4	0.053	0.000	0.000	0.000	0.947
...	...	...	...	...	...	...

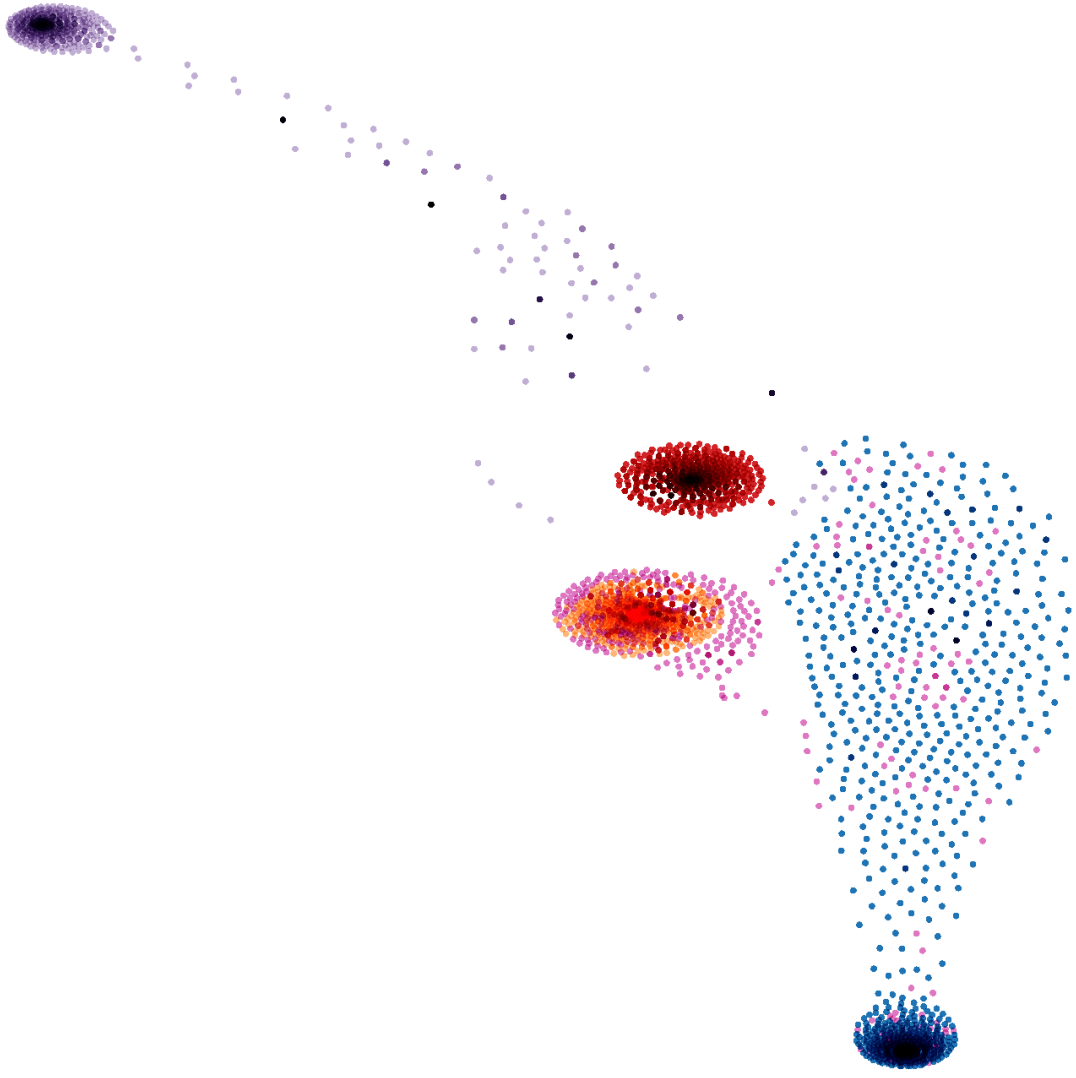
Table 2: Profiles from IS data in Table 1

Now each row sums up to 1 ! In other words $p_i = (p_i^G, p_i^M, p_i^B, p_i^K, p_i^T)$ belongs to the 5-simplex, where the d-simplex is defined as

$$\mathcal{S}^d = \left\{ p = (p^1, \ldots, p^d) \in \mathbb{R}^d \text{ s.t. } \forall l = 1, \ldots, d \; p^l > 0 \text{ and } \sum_{l=1}^{d} p^l = 1 \right\}.$$

Does the data exhibit clusters ?

We applied the t-SNE algorithm to our data.



Clustering algorithms based on a transformation of profile data

Kmeans with euclidean distance

We now describe clustering algorithms for compositional data based on transformations of profile data

$$p_i = (p_i^1, \ldots, p_i^d) \text{ for } i = 1, \ldots, N.$$

A first solution is to performed a Kmeans with the euclidean distance on raw profile data $\rightarrow$ very poor performances

The data has to be transformed via

- centered log ratio (CLT) [Ait82]
- log centered log ratio (logCLT) [GMR19]

Centered log ratio (CLT)

The CLT transformation is defined as

$$\text{CLR} : p \in \mathcal{S}^d \mapsto \text{CLR}(p) = \left(\log\left(\frac{p^1}{g(p)}\right), \ldots, \log\left(\frac{p^d}{g(p)}\right) \right)$$

where

$$g(p) = \left(\prod_{l=1}^{d} p^l\right)^{1/d}$$

is the geometric mean.

Problem: when the profile data exhibits 0 **!!**
a solution would be to first redefine profiles as

$$p_i^l = \frac{x_i^l + 1}{x_i^{\cdot} + d}.$$

Log centered log ratio (logCLR)

To be less sensitive to small fluctuations around near zero proportions, we can consider

$$\text{logCLR} : p \in \mathcal{S}^d \mapsto \text{logCLR}(p) = \Big(\text{logCLR}(p^1), \ldots, \text{logCLR}(p^d)\Big)$$

where

$$\text{logCLR}(p^l) = \begin{cases} -\Big(\log\big(1 - \log(p^l/g(p))\big)\Big)^2, & \text{if } p^l/g(p) \leq 1 \\ \Big(\log(p^l/g(p))\Big)^2, & \text{otherwise.} \end{cases}$$

Others transformations include the ALR [Ait82] and the ILR [Ego+03].

Model based clustering

Model based clustering: introduction

Mixture models have mainly been consider in problems

- from text mining, see e.g. [Nig+00]; [MH01]; [Bou08]
- with microbial data, see e.g. [HHQ12]; [Mor16].

We now consider the count data with value in $\mathbb{N}^d$, as shown in Table 1

$$x_i = (x_i^1, \ldots, x_i^d) \text{ for } i = 1, \ldots, N.$$

The observations $(x_i)_{i=1,\ldots,N}$ are assumed to be i.i.d realizations with a mixture of distributions.

Mixture of multinomial distributions

Generative process for the mixture of multinomials

Parameters:

- K number of clusters
- $\alpha = (\alpha_1, \ldots, \alpha_K)$ mixing proportions
- $(\pi_k^1, \ldots, \pi_k^d)$ for $k = 1, \ldots, K$ proportions in the multinomial distributions
- N number of trials in X

1. Choose $z \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_K)$
2. Choose $X \sim \mathcal{M}(N, \pi_z^1, \ldots, \pi_z^d)$.

Dirichlet multinomial mixture (1)

The Bayesian setup described in the remark above has been extended by several authors, for a simple generalization with a Dirichlet prior.Keeping the notation Z for the latent variable indicating the component of the mixture, the distribution of $X|Z = k$ is given by

$$\mathbb{P}\big(X = (x^1, \dots, x^d)|Z = k\big) = \frac{\Gamma(\sum_{l=1}^d x^l + 1)\Gamma(\sum_{l=1}^d \beta_k^l)}{\Gamma(\sum_{l=1}^d \beta_k^l + x^l)} \prod_{l=1}^d \frac{\Gamma(\beta_k^l + x^l)}{\Gamma(\beta_k^l)\Gamma(x^l + 1)}.$$

As the mixture of multinomial distributions, this model has $K + Kd$ parameters, but the parameters $\beta_k^1, \dots, \beta_k^d$ are not constrained to the d-simplex, to the contrary of the parameters $\alpha_k^1, \dots, \alpha_k^d$ of the previous model.

See e.g. [Nig+00]; [RCY07]; [YW14].

Dirichlet multinomial mixture (2)

Generative process for the mixture of Dirichlet-multinomials

Parameters:

- K number of clusters
- $\alpha = (\alpha_1, \ldots, \alpha_K)$ mixing proportions
- $(\beta_k^1, \ldots, \beta_k^d)$ for $k = 1, \ldots, K$ parameters of the Dirichlet distributions
- N number of trials in X

1. Choose $z \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_K)$
2. Choose $(\pi_z^1, \ldots, \pi_z^d) \sim Dir(\beta_z^1, \ldots, \beta_z^d)$
3. Choose $X \sim \mathcal{M}(N, \pi_z^1, \ldots, \pi_z^d)$.

Dirichlet multinomial mixture (3)

To regularize the log-evidence, [HHQ12] proposed to add Gamma priors (with common parameters) on the β_k^l's parameters have been proposed (see also the R package `DirichletMultinomial` [Mor16].

Generative process for the mixture of Dirichlet-multinomials with Gamma priors [HHQ12]

Parameters:

- K number of clusters
- $\alpha = (\alpha_1, \ldots, \alpha_K)$ mixing proportions
- (η, ζ) parameters of the Gamma distributions
- N number of trials in X

1. For all $k = 1, \ldots, K$, for all $l = 1, \ldots, d$ choose $\beta_k^l \sim \gamma(\eta, \zeta)$
2. Choose $z \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_K)$
3. Choose $(\pi_z^1, \ldots, \pi_z^d) \sim Dir(\beta_z^1, \ldots, \beta_z^d)$
4. Choose $X \sim \mathcal{M}(N, \pi_z^1, \ldots, \pi_z^d)$.

Dirichlet multinomial mixture (4)

One major drawback of this model is that the correlation between two proportions π_k^l and $\pi_k^{l'}$ in a vector π_k are always negative:

$$\text{cov}(\pi_k^l, \pi_k^{l'}) = -\frac{\beta_k^l \beta_k^{l'}}{(\sum_{l=1}^d \beta_k^l)(1 + \sum_{l=1}^d \beta_k^l)}.$$

Generalized Dirichlet multinomial mixture (1)

The generalized Dirichlet multinomial mixture was introduced in [Bou08]. The greater number of parameters permits in particular a more general form for the covariance between two random variables in π_k

$$\mathrm{cov}(\pi_k^l, \pi_k^{l'}) = \frac{\beta_k^{l'}}{\beta_k^{l'} + \gamma_k^{l'}} \prod_{j=1}^{l'-1} \frac{\gamma_k^j}{\beta_k^j + \gamma_k^j}$$
$$\left(\frac{\beta_k^l}{\beta_k^l + \gamma_k^l + 1} \prod_{j=1}^{l'-1} \frac{\gamma_k^j + 1}{\beta_k^j + \gamma_k^j + 1} - \frac{\beta_k^l}{\beta_k^l + \gamma_k^l} \prod_{j=1}^{l-1} \frac{\gamma_k^j}{\beta_k^j + \gamma_k^j} \right).$$

Generalized Dirichlet multinomial mixture (2)

The large number of parameters of the mixture model $K + K(2(d-1))$ might lead to instability in the estimation. We propose to add Gamma priors (with common parameters) for the parameters of the generalized Dirichlet distribution.

Generative process for the mixture of generalized Dirichlet-multinomials with Gamma priors

Parameters:

- K number of clusters
- $\alpha = (\alpha_1, \ldots, \alpha_K)$ mixing proportions
- (η, ζ) parameters of the Gamma distributions
- N number of trials in X

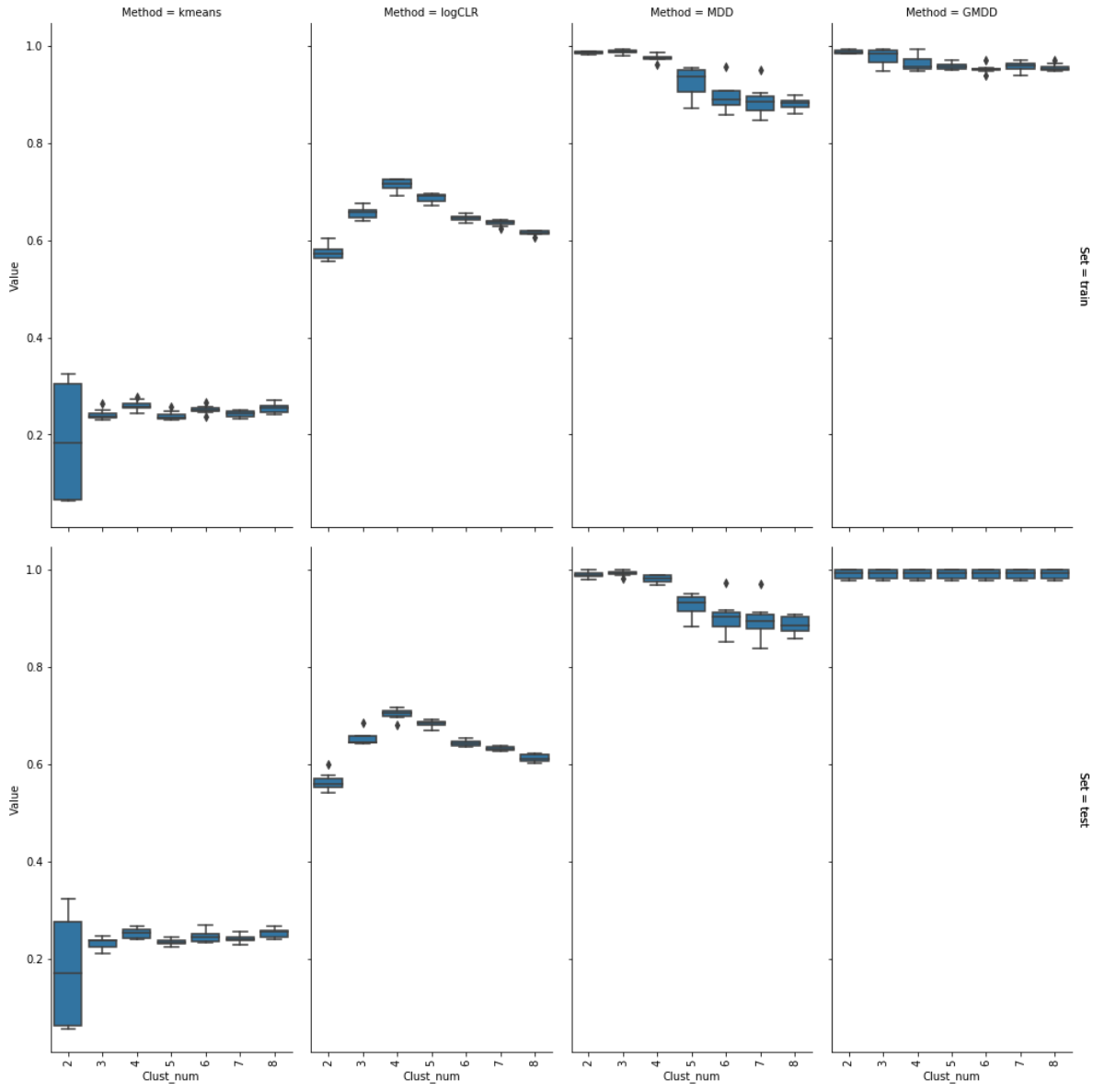
1. For all $k = 1, \ldots, K$, for all $l = 1, \ldots, d-1$ choose β_k^l and γ_k^l from a $\gamma(\eta, \zeta)$ distribution
2. Choose $z \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_K)$
3. Choose $(\pi_z^1, \ldots, \pi_z^d) \sim \mathit{GenDir}(\beta_k^1, \gamma_k^1, \ldots, \beta_k^{d-1}, \gamma_k^{d-1})$
4. Choose $X \sim \mathcal{M}(N, \pi_z^1, \ldots, \pi_z^d)$.

Further research

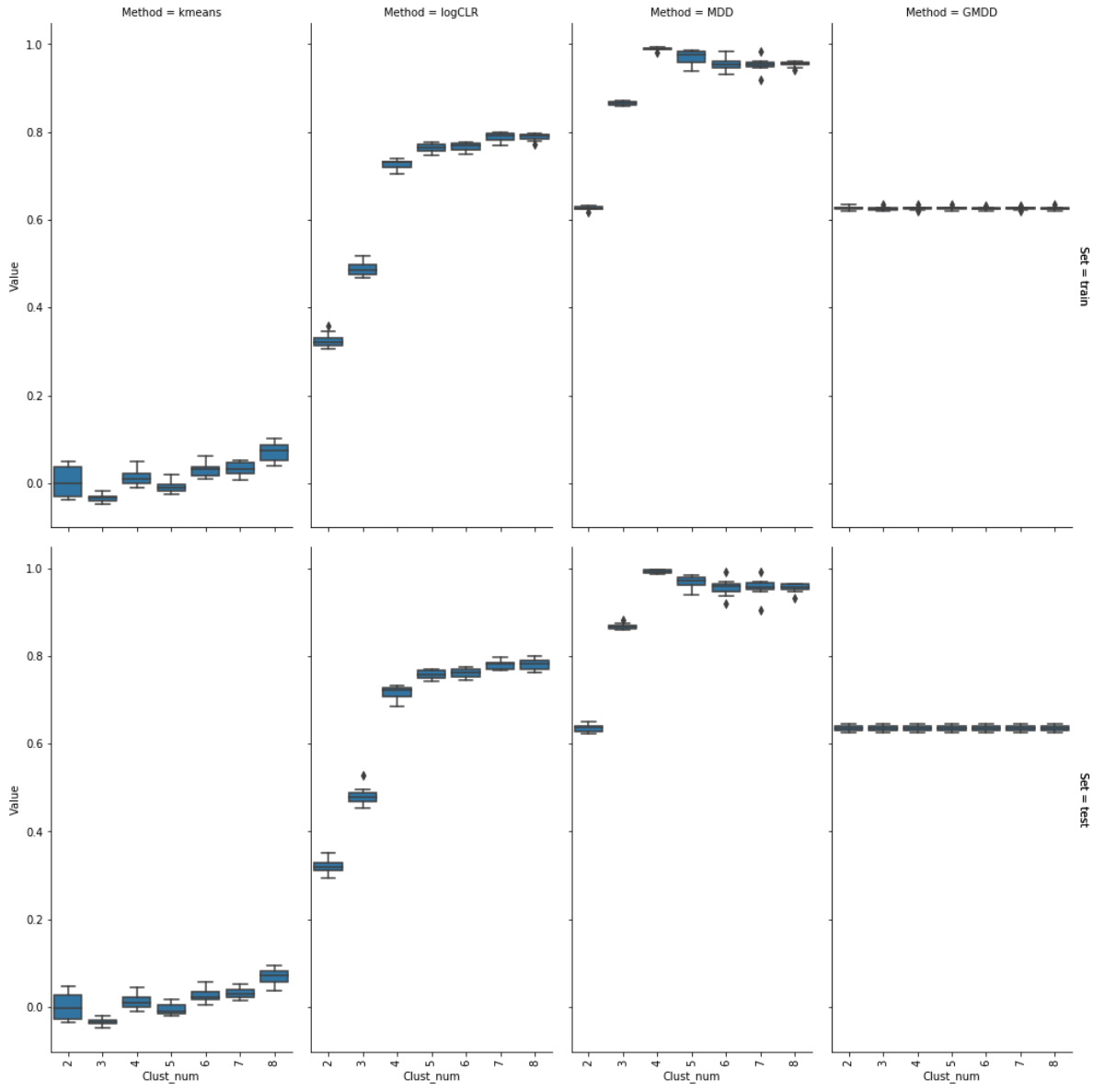
- Implement a choice of K the number of clusters (ICL ? [BCG00])
- Handle zero proportions: all the models presented do not allow for degenerate distributions for the π_k. Authors have proposed solutions
  - in the context of regression [TC18],
  - via imputation [MBP03],
  - or via Bayesian model selection [Tuyar].
- Extend the models to let the distribution of N (number of trials) depend on the clusters (Poisson or a Negative Binomial distributions)

Simulations

Homogeneity



Adjusted rand index



References I

John Aitchison. "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1982), pp. 139–177.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7 (2000), pp. 719–725.

Nizar Bouguila. "Clustering of count data using generalized Dirichlet multinomial distributions". In: *IEEE Transactions on Knowledge and Data Engineering* 20.4 (2008), pp. 462–474.

Juan José Egozcue et al. "Isometric logratio transformations for compositional data analysis". In: *Mathematical Geology* 35.3 (2003), pp. 279–300.

Antoine Godichon-Baggioni, Cathy Maugis-Rabusseau, and Andrea Rau. "Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data". In: *Journal of Applied Statistics* 46.1 (2019), pp. 47–65.

References II

Ian Holmes, Keith Harris, and Christopher Quince. "Dirichlet multinomial mixtures: generative models for microbial metagenomics". In: *PloS one* 7.2 (2012), e30126.

Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. "Dealing with zeros and missing values in compositional data sets using nonparametric imputation". In: *Mathematical Geology* 35.3 (2003), pp. 253–278.

Marina Meilă and David Heckerman. "An experimental comparison of model-based clustering methods". In: *Machine learning* 42.1-2 (2001), pp. 9–29.

Martin Morgan. *DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data*. R package version 1.16.0. 2016.

Kamal Nigam et al. "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2-3 (2000), pp. 103–134.

References III

Loïs Rigouste, Olivier Cappé, and François Yvon. "Inference and evaluation of the multinomial mixture model for text clustering". In: *Information processing & management* 43.5 (2007), pp. 1260–1280.

Zheng-Zheng Tang and Guanhua Chen. "Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis". In: *Biostatistics* (2018).

Franck Tuyl. "A Method to Handle Zero Counts in the Multinomial Model". In: *The American Statistician* (2019 (to appear)).

Jianhua Yin and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 233–242.