Journal of Classification (2025) 42:31–53
https://doi.org/10.1007/s00357-024-09480-4

ORIGINAL RESEARCH

A New Look at the Dirichlet Distribution: Robustness, Clustering, and Both Together

Salvatore D. Tomarchio[1] · **Antonio Punzo**[1] · **Johannes T. Ferreira**[2] · **Andriette Bekker**[2,3]

Accepted: 8 June 2024 / Published online: 2 July 2024

Abstract
Compositional data have peculiar characteristics that pose significant challenges to traditional statistical methods and models. Within this framework, we use a convenient mode parametrized Dirichlet distribution across multiple fields of statistics. In particular, we propose finite mixtures of unimodal Dirichlet (UD) distributions for model-based clustering and classification. Then, we introduce the contaminated UD (CUD) distribution, a heavy-tailed generalization of the UD distribution that allows for a more flexible tail behavior in the presence of atypical observations. Thirdly, we propose finite mixtures of CUD distributions to jointly account for the presence of clusters and atypical points in the data. Parameter estimation is carried out by directly maximizing the maximum likelihood or by using an expectation-maximization (EM) algorithm. Two analyses are conducted on simulated data to illustrate the effects of atypical observations on parameter estimation and data classification, and how our proposals address both aspects. Furthermore, two real datasets are investigated and the results obtained via our models are discussed.

Keywords Compositional data · Dirichlet distribution · Mode · Model-based clustering · Robustness

1 Introduction

In many fields, data is typically represented as parts of a whole, such as vectors of proportions or ratios, which are subject to constraints that ensure that they sum to 1 (unit sum) and are non-negative. This special type of multivariate data is known in the literature as compositional

✉ Salvatore D. Tomarchio
daniele.tomarchio@unict.it

1 Department of Economics and Business, University of Catania, Catania, Italy

2 Department of Statistics, University of Pretoria, Pretoria, South Africa

3 Department of Geography, Geoinformatics and Meteorology, Centre for Environmental Studies, Pretoria, South Africa

data. The peculiar characteristics of compositional data, which lead to using the simplex as the sample space, pose significant challenges to traditional statistical methods and models. Particularly, they cannot be straightforwardly used because this would lead to biased results due to the constrained nature of the simplex (Filzmoser et al., 2018; Ongaro et al., 2020).

In a broad sense, it is possible to distinguish between two main approaches for modeling compositional data. The first relies on mapping the simplex to an (unconstrained) Euclidean space via specific log-ratio transformations and then employing standard statistical methodologies as usual. However, this approach has several drawbacks since each log-ratio transformation has its pros and cons and, more generally, it is difficult to interpret parameters estimated by the models with respect to the original variables (Pawlowsky-Glahn, V., and Buccianti, A., 2011; Filzmoser et al., 2018; Ongaro et al., 2020).

A second approach consists of defining statistical models directly on the simplex. Within this framework, the Dirichlet distribution is the most commonly adopted, and several generalizations of it have been proposed in the literature (see, e.g., Lochner 1975; Ng et al. 2011; Ongaro and Migliorati 2013; Botha et al. 2021). Our paper is concerned with this second approach. Specifically, we consider a mode-parameterization of the Dirichlet distribution, and we illustrate its use for robust analyses, model-based clustering, and both together.

As discussed by Chacón (2020), there is an increased interest in inspecting many statistical problems from a modal point of view. The mode, median, and mean are the three measures of central tendency that are commonly used in data analysis. Nevertheless, the mode may be a more informative and meaningful measure of central tendency for data that originates from distributions that are skewed and have heavy tails (Kruschke 2014; Chacón 2020). Additionally, the mode is easy to comprehend, remains unaffected by extreme values, and can be identified visually (see, e.g., Nolan 1998). Thus, we consider a convenient parametrization of the Dirichlet distribution based on the mode vector $\boldsymbol{\theta}$ and on a positive scalar parameter γ that is closely related to the distribution variability. We refer to the resulting distribution as unimodal Dirichlet (UD).

The adopted parametrization simplifies and motivates the use of the Dirichlet distribution in different areas of statistics. As a first example, we introduce the contaminated UD (CUD) distribution, a heavy-tailed generalization of the UD distribution that allows for a more flexible tail behavior in the presence of atypical observations. In addition to the parameters of the UD distribution, the CUD distribution has two additional parameters: β controlling the proportion of atypical observations, and η specifying the degree of contamination. The CUD distribution is advantageous over the UD distribution in that it automatically down-weights atypical observations in the estimation of $\boldsymbol{\theta}$ and γ, making it a more robust method for parameter estimation. Furthermore, once it is fitted to a dataset, atypical observations can be easily identified via maximum *a posteriori* probabilities. This detection tool differs from what is typically done for compositional data. In fact, a standard approach involves mapping the data using specific log-ratio transformations and then calculating the squared Mahalanobis distances between the observations (expressed in coordinates) and the respective location estimator. A certain quantile of the χ^2 distribution is then used as a cut-off value to identify atypical observations (Filzmoser & Hron, 2008; Filzmoser & Gregorich, 2020). Such a procedure has several drawbacks such as (i) the subjective choice of the log-ratio transformation (with its pros and cons), (ii) the arbitrary selection of the quantile of the χ^2 distribution, and (iii) the not-so-straightforward interpretation of the reasons for the atypicality of the observations (Filzmoser et al., 2018).

A similar procedure consists of mapping the data via log-ratio transformations and fitting classical heavy-tailed distributions (Van den Boogaart and Tolosana-Delgado 2013). The related (model-based) detection tools, whether "automatic" or relying on squared Mahalanobis distances computed by using the estimated distribution parameters (see, e.g., Peel and McLachlan 2000; Punzo and McNicholas 2016), can be employed. However, the same drawbacks previously discussed also arise with this strategy. Contrarily, our methodology allows for an automatic tool that works directly on the simplex, does not require subjective choices, and facilitates the interpretation of the atypicality in terms of probabilities.

A second use for the UD distribution is model-based clustering. Finite mixture models based on the Dirichlet distribution have been investigated in the literature (see, e.g., Bouguila et al. 2004; Calif et al. 2011; Pal, S., and Heumann, C. 2022). Nevertheless, none of them guarantees that each mixture component is unimodal, although this assumption is fairly standard when working with finite mixture models. From an interpretative point of view, the importance of using unimodal components is justified, and natural, if we consider that multimodal distributions usually reflect the existence of several subpopulations within the distribution, and this can be modeled through a mixture density (Chacón 2020). As emphasized by Titterington et al. (1985) and McLachlan, G. J., and Basford, K. E. (1988), many papers prefer discussing multimodal distributions instead of mixtures. For instance, Murphy (1964) and Brazier et al. (1983) refer to bimodality instead of mixtures. Similarly, McNicholas (2016) discusses that a cluster should comprise points that diffuse from a mode, and mixture components should be unimodal, because if this were not the case, then two scenarios frequently arise: either the wrong mixture distribution is fitted (for example, multiple symmetric components are being used to model a single skewed cluster) or not enough components are being used. Thus, in light of the considerations above, we propose and discuss finite mixtures of UD distributions for the clustering and classification of compositional data.

As a last example, we consider finite mixtures of CUD distributions. This model jointly accounts for the presence of clusters and atypical points in the data. To the best of our knowledge, this is the first attempt to address both aspects by working directly on the simplex, i.e., without any data transformation.

The paper is organized as follows. In Sect. 2, the Dirichlet distribution is recalled and details about the obtained UD distribution are discussed. In Sect. 3, we discuss the uses of the UD distribution for robust analyses, model-based clustering, and both together. Depending on the considered use, parameter estimation is then carried on by directly maximizing the maximum likelihood (ML) or by using an expectation-maximization (EM) algorithm (Dempster et al., 1977). Two analyses on simulated data are conducted in Sect. 4 to illustrate the effects of atypical observations on parameter estimation and data classification, and how our proposals address both aspects. Two real datasets are analyzed in Sect. 5, whereas some concluding remarks are drawn in Sect. 6.

2 Background

In this section, we first recall the Dirichlet distribution (Sect. 2.1), and then we give details about the subclass of unimodal Dirichlet distributions we will use throughout this paper (Sect. 2.2).

2.1 Dirichlet Distribution

A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{T}_d$ is said to have a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top = \left(\boldsymbol{\alpha}_{-d}^\top, \alpha_d\right)^\top$, where $\alpha_j > 0$, $j = 1, \ldots, d$, if the joint probability density function (pdf) of $\boldsymbol{X}_{-d} = (X_1, \ldots, X_{d-1})^\top \in \mathbb{V}_{d-1}$ is

$$f_{\mathrm{D}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\alpha}_{-d}, \alpha_d\right) = \frac{\Gamma\left(\alpha_+\right)}{\prod_{j=1}^{d} \Gamma\left(\alpha_j\right)} \left(1 - x_+\right)^{\alpha_d - 1} \prod_{j=1}^{d-1} x_j^{\alpha_j - 1}, \tag{1}$$

$\Gamma(\cdot)$ denotes the gamma function,

$$\mathbb{T}_d = \left\{ (x_1, \ldots, x_d)^\top : x_j > 0,\, j = 1, \ldots, d, \text{ and } \sum_{j=1}^{d} x_j = 1 \right\}$$

is the d-dimensional closed simplex in $\mathbb{R}^d$ and

$$\mathbb{V}_{d-1} = \left\{ (x_1, \ldots, x_{d-1})^\top : x_j > 0,\, j = 1, \ldots, d-1, \text{ and } \sum_{j=1}^{d-1} x_j < 1 \right\}$$

denotes the $(d-1)$-dimensional open simplex in $\mathbb{R}^{d-1}$, with $x_+ = \sum_{j=1}^{d-1} x_j$ and $\alpha_+ = \sum_{j=1}^{d} \alpha_j$. Compactly, we will write $\boldsymbol{X} \sim \mathcal{D}(\boldsymbol{\alpha})$ on $\mathbb{T}_d$ or $\boldsymbol{X}_{-d} \sim \mathcal{D}\left(\boldsymbol{\alpha}_{-d}, \alpha_d\right)$ on $\mathbb{V}_{d-1}$ accordingly.

If $\boldsymbol{X} \sim \mathcal{D}(\boldsymbol{\alpha})$, then the coordinates of the mode are

$$x_j = \frac{\alpha_j - 1}{\alpha_+ - d}, \quad j = 1, \ldots, d, \tag{2}$$

which exist only if $\alpha_j > 1$. The univariate marginal distributions from $\boldsymbol{X} \sim \mathcal{D}(\boldsymbol{\alpha})$ are given by

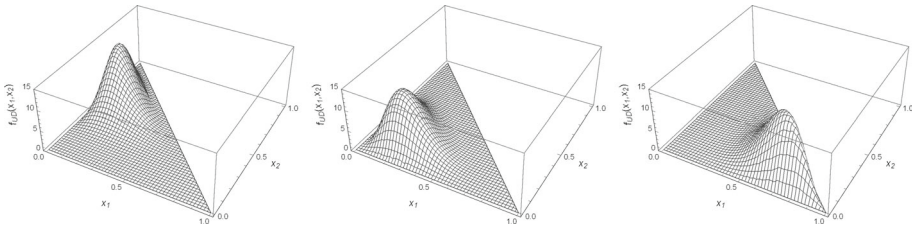
$$X_j \sim \mathcal{B}\left(\alpha_j, \alpha_+ - \alpha_j\right), \quad j = 1, \ldots, d, \tag{3}$$

where $\mathcal{B}(\alpha, \beta)$ denotes a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$.

2.2 Unimodal Dirichlet Distribution

The unimodal Dirichlet (UD) distribution we use was proposed by Aitchison and Lauder (1985) in the context of kernel density estimation, and studied more, always in the same context, by Ouimet and Tolosana-Delgado (2022) and Bertin et al. (2023). The UD distribution has pdf

$$f_{\mathrm{UD}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma\right) = \frac{\Gamma\left(d + \frac{1}{\gamma}\right)}{\Gamma\left(1 + \frac{1 - \theta_+}{\gamma}\right) \prod_{j=1}^{d-1} \Gamma\left(1 + \frac{\theta_j}{\gamma}\right)} \left(1 - x_+\right)^{\frac{1-\theta_+}{\gamma}} \prod_{j=1}^{d-1} x_j^{\frac{\theta_j}{\gamma}}, \tag{4}$$



(a) $\theta_1 = 0.1$ and $\theta_2 = 0.6$ (b) $\theta_1 = 0.2$ and $\theta_2 = 0.2$ (c) $\theta_1 = 0.7$ and $\theta_2 = 0.15$

Fig. 1 Plots of the pdfs of $(x_1, x_2)^\top \sim \mathcal{UD}(\boldsymbol{\theta}, \gamma)$ on $\mathbb{V}_2$, with $\gamma = 0.1$ and varying $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d-1})^\top \in \mathbb{V}_{d-1}$ and $\gamma > 0$, with $\theta_+ = \sum_{j=1}^{d-1} \theta_j < 1$. The link between the parametrizations in Eqs. 1 and 4 is

$$\begin{cases} \alpha_j = 1 + \dfrac{\theta_j}{\gamma}, \quad j = 1, \ldots, d-1 \\ \alpha_d = 1 + \dfrac{1-\theta_+}{\gamma} \end{cases} \Leftrightarrow \begin{cases} \theta_j = \dfrac{\alpha_j - 1}{\alpha_+ - d}, \quad j = 1, \ldots, d-1 \\ \gamma = \dfrac{1}{\alpha_+ - d} \end{cases}, \tag{5}$$

provided that $\alpha_j > 1$, $j = 1, \ldots, d$. Because of the constraint on the parameters α_j, the UD distributions are a subclass of the class of Dirichlet distributions in Eq. 1. If $\boldsymbol{X}_{-d}$ has the pdf in Eq. 5, then compactly we will write $\boldsymbol{X}_{-d} \sim \mathcal{UD}(\boldsymbol{\theta}, \gamma)$ on $\mathbb{V}_{d-1}$.

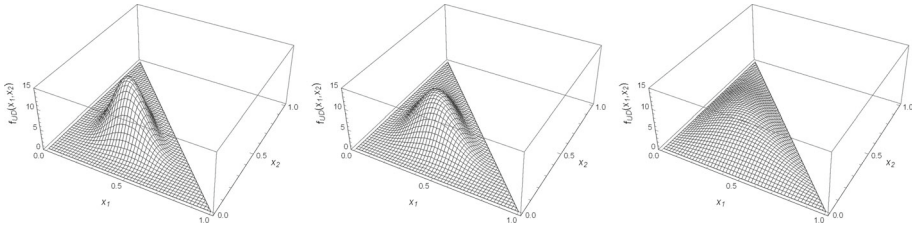
As for the interpretation of $\boldsymbol{\theta}$, first focus on the system on the right-hand side of Eq. 5. Here, recalling (2), the $d-1$ equations on the top guarantee that $\boldsymbol{\theta}$ is the mode of $\boldsymbol{X}_{-d}$ on $\mathbb{V}_{d-1}$; the effect of varying $\boldsymbol{\theta}$, with γ kept fixed, is illustrated in Fig. 1 in the case $d = 3$.

As for the interpretation of γ, the last equation of the system on the right-hand side of Eq. 5 is chosen so that γ is approximately related to the variability of the $d-1$ variables in $\boldsymbol{X}_{-d}$ on $\mathbb{V}_{d-1}$. The effect of varying γ in Eq. 4, the mode $\boldsymbol{\theta}$ kept fixed, is illustrated in Fig. 2 in the case $d = 3$.

Finally, starting from Eq. 3 and based on Eq. 5, it is straightforward to realize that the univariate marginal distributions from $\boldsymbol{X}_{-d} \sim \mathcal{UD}(\boldsymbol{\theta}, \gamma)$ are given by

$$X_j \sim \mathcal{B}\left(1 + \frac{\theta_j}{\gamma}, (d-1) + \frac{1-\theta_j}{\gamma}\right), \quad j = 1, \ldots, d-1, \tag{6}$$

where the first shape parameter is larger than one, while the other is larger than $d-1$. Therefore, when $d = 2$, we obtain all the unimodal beta (UB) distributions while, when



(a) $\gamma = 0.07$ (b) $\gamma = 0.1$ (c) $\gamma = 0.25$

Fig. 2 Plots of the pdfs of $(x_1, x_2)^\top \sim \mathcal{UD}(\boldsymbol{\theta}, \gamma)$ on $\mathbb{V}_2$, with $\boldsymbol{\theta} = (1/3, 1/3)^\top$ and increasing values of γ

$d > 2$, we obtain only a subclass of them. This is due to the constraints on the values in $\boldsymbol{X}_{-d}$: the larger the number of values in $\boldsymbol{X}_{-d}$, the closer to 0 the single value should be. In particular, in the case $d = 2$, $\theta_+ = \theta_1 = \theta$ and $X_1 = X \sim \mathcal{B}\left(1 + \frac{\theta}{\gamma}, 1 + \frac{1-\theta}{\gamma}\right)$, whose pdf is

$$f_{\text{UB}}(x; \theta, \gamma) = \frac{x^{\frac{\theta}{\gamma}}(1-x)^{\frac{1-\theta}{\gamma}}}{\text{B}\left(1 + \frac{\theta}{\gamma}, 1 + \frac{1-\theta}{\gamma}\right)}, \quad 0 < x < 1, \tag{7}$$

where $\text{B}(u, v)$ denotes the beta function. The pdf in Eq. 7 defines the subclass of unimodal beta (UB) distributions, with mode $\theta \in (0, 1)$ and dispersion (around the mode) parameter $\gamma > 0$ (the larger γ, the more concentrated the distribution about the mode θ_j), parametrized as in Chen (1999, 2000); see also Bagnato and Punzo (2013) and Tomarchio and Punzo (2019).

Maximum Likelihood Estimation

To find the estimates of the parameters for the UD distribution in Eq. 4, we consider the maximum likelihood (ML) approach. Given a random sample $\{\boldsymbol{x}_{i-d}\}_{i=1}^{n}$ of size n from the pdf in Eq. 4, the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}, \gamma) = \sum_{i=1}^{n} \log\left[f_{\text{UD}}\left(\boldsymbol{x}_{i-d}; \boldsymbol{\theta}, \gamma\right)\right]. \tag{8}$$

Since closed-form ML estimates are unavailable, we employ numerical methods to maximize the likelihood, adhering to any parameter constraints that may exist. From a computational point of view, we obtain maximization of Eq. 8, with respect to $\boldsymbol{\theta}$ and γ, by the general-purpose optimizer `optim()` for R (R Core Team , 2021), included in the **stats** package. The Nelder-Mead algorithm (Nelder & Mead, 1965), passed to `optim()` via the argument `method`, is used for maximization.

3 Methodological Examples

In this section, we show how the parameterization of the UD distribution in Eq. 4 allows/simplifies the use of the Dirichlet distribution in different areas of statistics for compositional data. We start by introducing finite mixtures of UD distributions for model-based clustering and classification (Sect. 3.1). Then, in Sect. 3.2, we define CUD distribution for robust estimation in the presence of atypical observations and how they are automatically detected. Additionally, we introduce finite mixtures of CUD distributions to jointly consider robust estimation as well as model-based clustering and classification. Even in this case, we can identify atypical observations once each of them is assigned to a mixture component.

3.1 Model-Based Clustering via the UD Distribution

As introduced in Sect. 1, it is important to consider unimodal components in a finite mixture model. Therefore, we present finite mixtures of UD distributions in Sect. 3.1.1. Parameter estimation via the EM algorithm is also illustrated.

3.1.1 Mixtures of UD Distributions

The pdf of a finite mixture of k UD distributions is

$$p\left(\boldsymbol{x}_{-d};\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{\gamma}\right)=\sum_{j=1}^{k}\pi_j f_{\text{UD}}\left(\boldsymbol{x}_{-d};\boldsymbol{\theta}_j,\gamma_j\right),\quad \boldsymbol{x}_{-d}\in\mathbb{V}_{d-1}, \tag{9}$$

where $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_k)^\top$ is the vector of mixture weights, $\boldsymbol{\Theta}=\left(\boldsymbol{\theta}_1^\top,\ldots,\boldsymbol{\theta}_k^\top\right)^\top$ is the vector of components' modes $\boldsymbol{\theta}_j$, and $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_k)^\top$ is the vector of components' dispersion parameters. Thus, there are $(k-1)+kd$ unknown parameters to be estimated. Notationally, we refer to model (9) as the UD-mixture (UD-M) model.

As outlined by Izenman (2008), Bagnato and Punzo (2013), and Punzo (2019), there is no assurance that (9) will generate a multimodal density with an identical number of modes as mixture components. The configuration of the mixture distribution is affected by both the spacing of the modes and the shapes of the component distributions. However, supported by Ray and Lindsay (2005), we retain that for well-separated components, the values of $\boldsymbol{\theta}_j$ can be used to approximate the location of the component modes.

EM Algorithm

Parameter estimation is carried out via the EM algorithm, which is a widely used approach for ML estimation when data are incomplete. Let $\left\{\boldsymbol{x}_{i-d}\right\}_{i=1}^{n}$ be a random sample of size n from the pdf in Eq. 9. Within the formulation of mixture models, data are viewed as incomplete because, for each observation, we do not know its component membership. To govern this source of incompleteness, we consider an indicator vector $\boldsymbol{z}_i=(z_{i1},\ldots,z_{ik})$, where $z_{ij}=1$ if $\boldsymbol{x}_{i-d}$ comes from component j and $z_{ij}=0$ otherwise. Therefore, we have the following complete-data log-likelihood

$$l_c\left(\boldsymbol{\Psi}\right)=l_{1c}\left(\boldsymbol{\pi}\right)+l_{2c}\left(\boldsymbol{\Theta},\boldsymbol{\gamma}\right), \tag{10}$$

where $\boldsymbol{\Psi}=\{\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{\gamma}\}$,

$$l_{1c}\left(\boldsymbol{\pi}\right)=\sum_{i=1}^{n}\sum_{j=1}^{k}z_{ij}\log\pi_j, \tag{11}$$

$$l_{2c}\left(\boldsymbol{\Theta},\boldsymbol{\gamma}\right)=\sum_{i=1}^{n}\sum_{j=1}^{k}z_{ij}\log\left[f_{\text{UD}}\left(\boldsymbol{x}_{i-d};\boldsymbol{\theta}_j,\gamma_j\right)\right]. \tag{12}$$

The EM algorithm alternates between two steps, an E-step and an M-step, until convergence. These steps are outlined below. It is worth noting that parameters denoted with one dot indicate updates from the previous iteration, while those marked with two dots represent updates from the current iteration.

E-step. The E-step requires the calculation of

$$Q\left(\boldsymbol{\Psi}|\dot{\boldsymbol{\Psi}}\right)=Q_1\left(\boldsymbol{\pi}|\dot{\boldsymbol{\Psi}}\right)+Q_2\left(\boldsymbol{\Theta},\boldsymbol{\gamma}|\dot{\boldsymbol{\Psi}}\right), \tag{13}$$

the conditional expectation of Eq. 10, given the observed data and using the current fit $\dot{\boldsymbol{\Psi}}$ for $\boldsymbol{\Psi}$. This implies the calculation of

$$\ddot{z}_{ij}=\frac{\dot{\pi}_j f_{\text{UD}}\left(\boldsymbol{x}_{i-d};\dot{\boldsymbol{\theta}}_j,\dot{\gamma}_j\right)}{p\left(\boldsymbol{x}_{i-d};\dot{\boldsymbol{\pi}},\dot{\boldsymbol{\Theta}},\dot{\boldsymbol{\gamma}}\right)},$$

which corresponds to the posterior probability that the unlabeled observation $\boldsymbol{x}_{i_{-d}}$ belongs to the jth component of the mixture.

M-step. The M-step requires the calculation of $\ddot{\boldsymbol{\Psi}}$ as the value that maximizes $Q\left(\boldsymbol{\Psi}|\dot{\boldsymbol{\Psi}}\right)$. Since the two terms on the right-hand side of Eq. 13 have zero cross-derivatives, they can be maximized separately. Maximizing $Q_1\left(\boldsymbol{\pi}|\dot{\boldsymbol{\Psi}}\right)$ with respect to $\boldsymbol{\pi}$ yields to

$$\ddot{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\ddot{z}_{ij}, \quad j = 1, \ldots, k. \tag{14}$$

Maximizing $Q_2\left(\boldsymbol{\Theta}, \boldsymbol{\gamma}|\dot{\boldsymbol{\Psi}}\right)$ with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\gamma}$ is equivalent to independently maximizing each of the k expressions

$$Q_{2j}\left(\boldsymbol{\theta}_j, \gamma_j\right) = \sum_{i=1}^{n}\ddot{z}_{ij}\log\left[f_{\text{UD}}\left(\boldsymbol{x}_{i_{-d}}; \boldsymbol{\theta}_j, \gamma_j\right)\right], \quad j = 1, \ldots, k. \tag{15}$$

From an operative standpoint, the maximization of Eq. 15 is numerically obtained in the fashion of what is already discussed in Sect. 2.2, but with each observation $\boldsymbol{x}_{i_{-d}}$ weighted according to $\ddot{z}_{ij}$.

3.2 Robustness Against Atypical Observations

Real data frequently contain atypical observations. As introduced in Sect. 1, the traditional ways of handling such observations, which consist of mapping compositional data via log-ratio transformations and then relying on traditional heavy-tailed distributions and/or squared Mahalanobis distances, have several drawbacks. An alternative approach is to define heavy-tailed distributions directly in the simplex. To achieve this, a target distribution (UD in our case) is embedded in a larger model with one or more additional parameters that represent the deviation from the target distribution due to atypical observations; for details about the concept of the target distribution, see Davies and Gather (1993); Tomarchio and Punzo (2020). Specifically, in Sect. 3.2.1, we introduce the CUD distribution using the UD as the target distribution. Then, we discuss the use of the CUD distribution within a finite mixture modeling setting in Sect. 3.2.2.

3.2.1 Contaminated UD Distribution

The CUD distribution comes in the form of a two-component mixture, with one component representing the typical observations (target distribution) and the other component, having the same mode and an inflated dispersion parameter, representing atypical observations (contaminant distribution). In detail, the pdf of the CUD distribution is

$$\begin{aligned} f_{\text{CUD}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma, \beta, \eta\right) =& \beta\, f_{\text{UD}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma\right) \\ &+ (1-\beta)\, f_{\text{UD}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \eta\gamma\right), \quad \boldsymbol{x}_{-d} \in \mathbb{V}_{d-1}. \end{aligned} \tag{16}$$

In Eq. 16:

- $\beta \in (0, 1)$ can be seen as the proportion of typical points. Note that, in robust statistics is usually assumed that at least half of the observations are typical (Punzo & McNicholas, 2016), and as such, we constrain β to be greater than 0.5.

- $\eta > 1$ represents the degree of contamination and, because of the assumption $\eta > 1$, it can be interpreted as the increase in variability due to the atypical observations compared to the target distribution $f_{\text{UD}}(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma)$. Therefore, it is an inflation parameter.

It should be remarked that $f_{\text{CUD}}(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma, \beta, \eta)$ produces a unimodal density, with mode $\boldsymbol{\theta}$, because both the target and contaminant pdfs have their maximum in $\boldsymbol{\theta}$. Thus, the parametrization of the UD distribution given in Eq. 4 is fundamental for guaranteeing the unimodality of the contaminated distribution. As a limiting case of Eq. 16, when $\beta \to 1^-$ and $\eta \to 1^+$, the target distribution $f_{\text{UD}}(\boldsymbol{x}_{-d}; \boldsymbol{\theta}, \gamma)$ is obtained.

Maximum Likelihood Estimation

To find the estimates of the parameters for CUD distribution in Eq. 16, we consider the ML approach. Given a random sample $\{\boldsymbol{x}_{i_{-d}}\}_{i=1}^{n}$ of size n from the pdf in Eq. 16, the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}, \gamma, \beta, \eta) = \sum_{i=1}^{n} \log\left[f_{\text{CUD}}\left(\boldsymbol{x}_{i_{-d}}; \boldsymbol{\theta}, \gamma, \beta, \eta\right)\right]. \tag{17}$$

Similarly to Sect. 2.2, closed-form ML estimates are not available; thus we numerically maximize (17), with respect to $\boldsymbol{\theta}$, γ, β, and η by the general-purpose optimizer `optim()`.

Automatic Detection of Atypical Observations

An advantageous feature of the CUD distribution is that, once its parameters are estimated (denoted with a 'hat' in the following), it becomes feasible to ascertain whether a given observation is typical or not through the *a posteriori* probability

$$P\left(\boldsymbol{x}_{-d} \text{ is typical}; \widehat{\boldsymbol{\theta}}, \widehat{\gamma}, \widehat{\beta}, \widehat{\eta}\right) = \frac{\widehat{\beta} f_{\text{UD}}\left(\boldsymbol{x}_{-d}; \widehat{\boldsymbol{\theta}}, \widehat{\gamma}\right)}{f_{\text{CUD}}\left(\boldsymbol{x}_{-d}; \widehat{\boldsymbol{\theta}}, \widehat{\gamma}, \widehat{\beta}, \widehat{\eta}\right)}. \tag{18}$$

Based on Eq. 18, $\boldsymbol{x}_{-d}$ is considered typical if $P\left(\boldsymbol{x}_{-d} \text{ is typical}; \widehat{\boldsymbol{\theta}}, \widehat{\gamma}, \widehat{\beta}, \widehat{\eta}\right) > 0.5$, while it will be considered atypical otherwise. Thus, we have an automatic tool for characterizing the degree of typicality of each data observation.

3.2.2 Mixtures of CUD Distributions

The pdf of a finite mixture of k CUD distributions is

$$g(\boldsymbol{x}_{-d}; \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j f_{\text{CUD}}\left(\boldsymbol{x}_{-d}; \boldsymbol{\theta}_j, \gamma_j, \beta_j, \eta_j\right), \quad \boldsymbol{x}_{-d} \in \mathbb{V}_{d-1}, \tag{19}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ is the vector of components' proportion of typical points, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^\top$ is the vector of components' degree of contamination. Therefore, there are $(k-1) + k(d+2)$ unknown parameters to be estimated. Notationally, we refer to model (19) as the CUD-mixture (CUD-M) model.

Since the mixture components have heavier tails than those of the UD distribution, model (19) can cope with clusters having potential atypical observations in a better way than model (9). This suggests an improved fitting of the data, circumventing the disruption of the true underlying grouping structure, and protects model (19) from possible misspecifications due to heavier-than-UD tails component distributions.

EM Algorithm

Similarly to Sect. 3.1.1, parameter estimation is carried out using the EM algorithm. Thus, consider the complete-data log-likelihood function according to the following specification

$$l_c\left(\boldsymbol{\Phi}\right) = l_{1c}\left(\boldsymbol{\pi}\right) + l_{2c}\left(\boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}\right), \tag{20}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}\}$, $l_{1c}\left(\boldsymbol{\pi}\right)$ is as in Eq. 11, while

$$l_{2c}\left(\boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log\left[f_{\text{CUD}}\left(\boldsymbol{x}_{i_{-d}}; \boldsymbol{\theta}_j, \gamma_j, \boldsymbol{\beta}_j, \boldsymbol{\eta}_j\right)\right]. \tag{21}$$

The EM algorithm proceeds as follows until convergence.

E-step. The E-step requires the calculation of

$$Q\left(\boldsymbol{\Phi}|\dot{\boldsymbol{\Phi}}\right) = Q_1\left(\boldsymbol{\pi}|\dot{\boldsymbol{\Phi}}\right) + Q_2\left(\boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\dot{\boldsymbol{\Phi}}\right), \tag{22}$$

the conditional expectation of Eq. 20, given the observed data and using the current fit $\dot{\boldsymbol{\Phi}}$ for $\boldsymbol{\Phi}$. This implies the calculation of

$$\ddot{z}_{ij} = \frac{\dot{\pi}_j f_{\text{CUD}}\left(\boldsymbol{x}_{i_{-d}}; \dot{\boldsymbol{\theta}}_j, \dot{\boldsymbol{\beta}}_j, \dot{\boldsymbol{\eta}}_j\dot{\gamma}_j\right)}{g\left(\boldsymbol{x}_{i_{-d}}; \dot{\boldsymbol{\pi}}, \dot{\boldsymbol{\Theta}}, \dot{\boldsymbol{\gamma}}, \dot{\boldsymbol{\beta}}, \dot{\boldsymbol{\eta}}\right)},$$

which corresponds to the posterior probability that the unlabeled observation $\boldsymbol{x}_{i_{-d}}$ belongs to the jth component of the mixture.

M-step. The M-step requires the calculation of $\ddot{\boldsymbol{\Phi}}$ as the value that maximizes $Q\left(\boldsymbol{\Phi}|\dot{\boldsymbol{\Phi}}\right)$. Since the two terms on the right-hand side of Eq. 22 have zero cross-derivatives, they can be maximized separately. Maximizing $Q_1\left(\boldsymbol{\pi}|\dot{\boldsymbol{\Phi}}\right)$ with respect to $\boldsymbol{\pi}$ yields the same update reported in Eq. 14. Maximizing $Q_2\left(\boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\eta}|\dot{\boldsymbol{\Phi}}\right)$ with respect to $\boldsymbol{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}$, and $\boldsymbol{\eta}$ is equivalent to independently maximizing each of the k expressions

$$Q_{2j}\left(\boldsymbol{\theta}_j, \gamma_j, \boldsymbol{\beta}_j, \boldsymbol{\eta}_j\right) = \sum_{i=1}^{n} \ddot{z}_{ij} \log\left[f_{\text{CUD}}\left(\boldsymbol{x}_{i_{-d}}; \boldsymbol{\theta}_j, \gamma_j, \boldsymbol{\beta}_j, \boldsymbol{\eta}_j\right)\right], \quad j = 1, \ldots, k. \tag{23}$$

Operationally, the weighted maximization of Eq. 23 is numerically obtained in the fashion of what is already discussed in Sect. 3.1.1.

Automatic Detection of Atypical Observations

To classify each observation, a two-step procedure is needed. First of all, we have to determine the cluster memberships. To this aim, we rely on the maximum *a posteriori* probabilities (MAP) operator, that is

$$\text{MAP}(\widehat{z}_{ij}) = \begin{cases} 1 & \text{if } \max_h\{\widehat{z}_{ih}\} \text{ occurs in group } h = j, \\ 0 & \text{if otherwise,} \end{cases}$$

where $\widehat{z}_{ij}$ is the value obtained at the convergence of the EM algorithm. Then, it is possible to determine whether a generic observation is typical or not via the *a posteriori* probability

$$P\left(\boldsymbol{x}_{-d} \text{ is typical}; \widehat{\boldsymbol{\theta}}_h, \widehat{\gamma}_h, \widehat{\beta}_h, \widehat{\eta}_h\right) = \frac{\widehat{\boldsymbol{\theta}}_h f_{\text{UD}}\left(\boldsymbol{x}_{-d}; \widehat{\boldsymbol{\theta}}_h, \widehat{\gamma}_h\right)}{f_{\text{CUD}}\left(\boldsymbol{x}_{-d}; \widehat{\boldsymbol{\theta}}_h, \widehat{\gamma}_h, \widehat{\beta}_h, \widehat{\eta}_h\right)}. \tag{24}$$

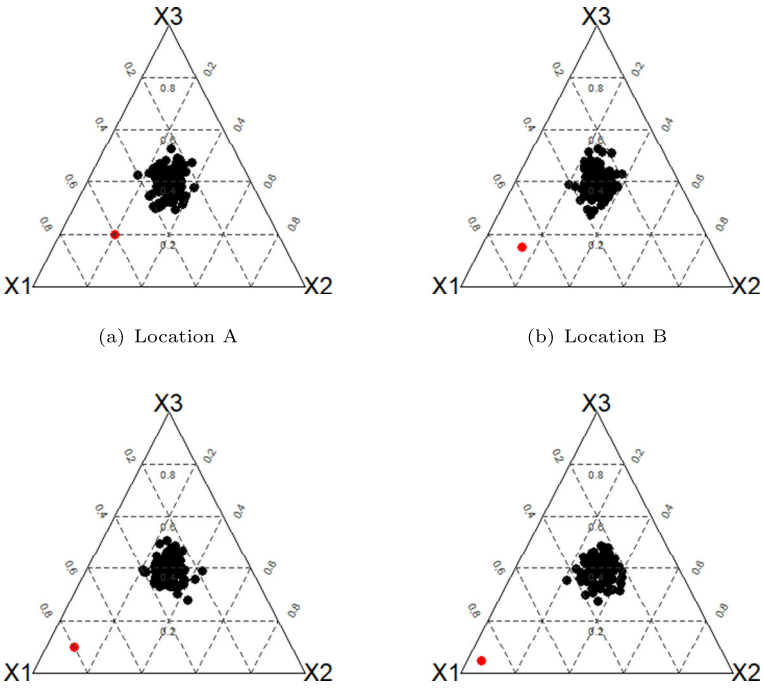
According to Eq. 24, $\boldsymbol{x}_{-d}$ is considered typical in cluster h if $P\left(\boldsymbol{x}_{-d} \text{ is typical}; \widehat{\boldsymbol{\theta}}_h, \widehat{\gamma}_h, \widehat{\beta}_h, \widehat{\eta}_h\right) > 0.5$, while it will be considered atypical otherwise.

4 Simulated Data Analyses

In this section, we investigate several aspects via simulated data. Specifically, in Sect. 4.1, we perform a sensitivity analysis that aims at evaluating the impact of a single atypical observation on the parameter estimates of the UD and CUD distributions. In Sect. 4.2, we implement a further sensitivity analysis to assess the impact of background noise on the data classification of UD-M and CUD-M models. Furthermore, we also evaluate the behavior of the CUD-M model in automatically detecting atypical observations. Finally, in Sect. 4.3, we evaluate the parameter recovery and mixture order selection for the CUD-M model under different data-generating settings.

4.1 Simulation Study 1

In this study, we simulate datasets of size $n = 100$ from the UD distribution in Eq. 4, after setting $d = 3$, $\boldsymbol{\theta} = (0.3, 0.3)$, and $\gamma = 0.01$. We consider three scenarios (labeled as Scenario 1, 2, and 3) determined by the direction in which a single atypical point is placed in



(a) Location A (b) Location B

(c) Location C (d) Location D

Fig. 3 Four simulated datasets from Scenario 1 with the four atypical point locations. The atypical point is in red

the simplex. Then, for a given scenario, the atypical point assumes four different locations (labeled as A, B, C, and D) such that it gradually moves away from the bulk of the data. Thus, we consider $3 \times 4 = 12$ data configurations. Examples of simulated datasets for Scenario 1 at the four locations are illustrated in Fig. 3 via ternary diagrams.

As we can see, when moving from location A to D, the atypical point goes further away from the rest of the points toward the lower left direction in the simplex. When we consider Scenarios 2 and 3, the atypical point moves along the lower right and upper direction in the simplex, respectively.

For each of the 12 data configurations, 500 datasets are generated, leading to a total of $12 \times 500 = 6000$ datasets. On each simulated dataset, we fit both the UD and CUD distributions and compute the absolute difference between the true parameter values and the estimated ones. Results are reported via the box plots of Fig. 4. Each subfigure refers to a specific scenario over the different locations. In turn, each box plot summarizes the behavior of the considered differences for the available 500 replications.

The first and immediate result is that the differences under evaluation remain essentially the same for the CUD distribution regardless of the considered data configuration. Conversely, the estimation performance for the UD distribution in each scenario gradually becomes worse as we pass from location A to location D, i.e., the atypical point moves further away from the bulk of the data. Solely at location A, the UD and CUD distributions exhibit similar results throughout the three distinct scenarios. Thus, the CUD distribution guarantees a more robust estimation of the parameters in the presence of atypical points, differently from the UD distribution.

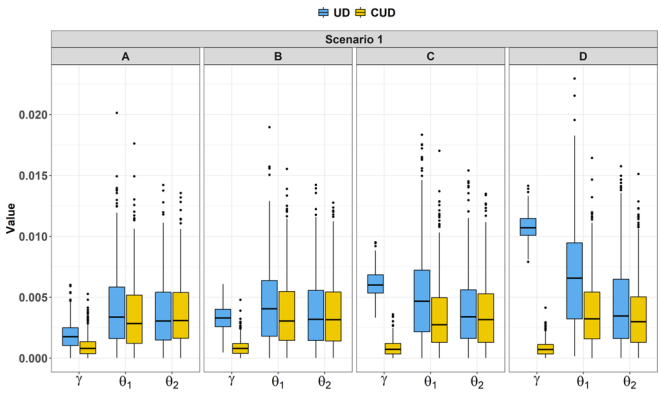
4.2 Simulation Study 2

In this study, we simulate 500 datasets of size $n = 250$ from the UD-M model in Eq. 9, after setting $d = 3$, $\boldsymbol{\theta}_1 = (0.4, 0.2)$, $\boldsymbol{\theta}_2 = (0.2, 0.4)$, $\gamma_1 = \gamma_2 = 0.01$, and $\pi_1 = \pi_2 = 0.5$. For each dataset, we randomly select 15% of the observations and replace their values with random numbers generated from a uniform distribution over the interval $[0, 1]$. Each noisy observation is then rescaled to guarantee the unit sum constraint.

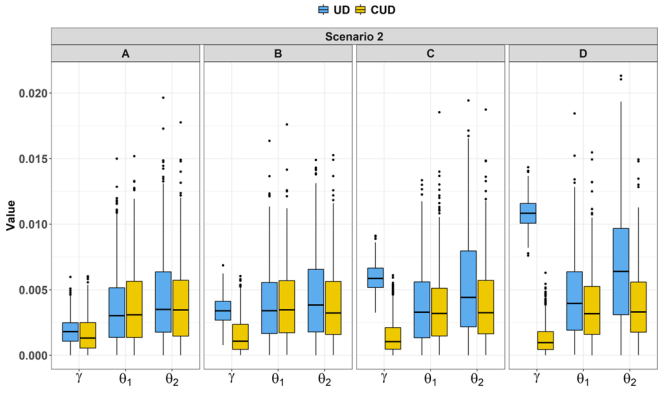
On each dataset, we fit UD-M and CUD-M models with $k = 2$. Then, we first evaluate the classification behavior by computing the adjusted Rand index (ARI; Hubert and Arabie 1985), which compares the pairwise agreement between the true classification and the one predicted by a model. Note that, in the fashion of Punzo and McNicholas (2016), the ARI is calculated only by considering the true typical observations, i.e., the noisy points are not taken into consideration. Secondly, we assess the behavior of the CUD-M model in detecting atypical observations using the true positive rate (TPR), which measures the proportion of atypical observations that are correctly identified as atypical, and the false positive rate (FPR), which corresponds to the proportion of typical points incorrectly classified as atypical.

Results are reported in Figs. 5 and 6 via box plots.

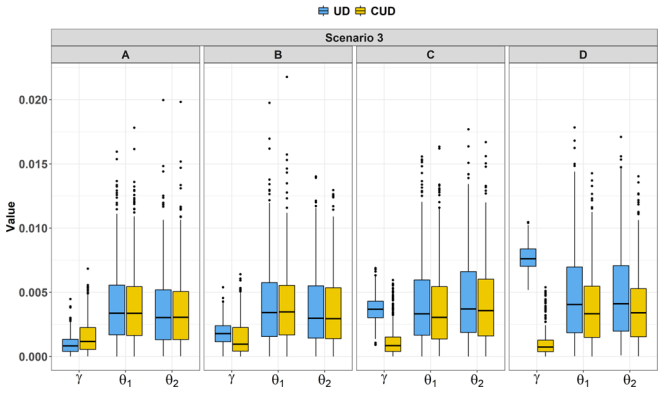
By starting with the analysis of Fig. 5, we see that the CUD-M model almost regularly provides an excellent data classification. On the contrary, the UD-M model shows far worse ARI values than those of the CUD-M model. This is due to the lower flexibility of the UD-M mixture components in the presence of noisy observations. In particular, we observed in approximately 48% of simulated datasets that the typical observations from both groups are aggregated into a single cluster, while the second component endeavors to model the background noise. An example of a dataset showing this issue is illustrated in Fig. 7, where the colors refer to the estimated classification by the UD-M model. As we can see, the true underlying group structure is disrupted.



(a)



(b)



(c)

Fig. 4 Box plots of the absolute difference between the true and the estimated parameter values, for the three scenarios (1 to 3) over the four locations (A to D)

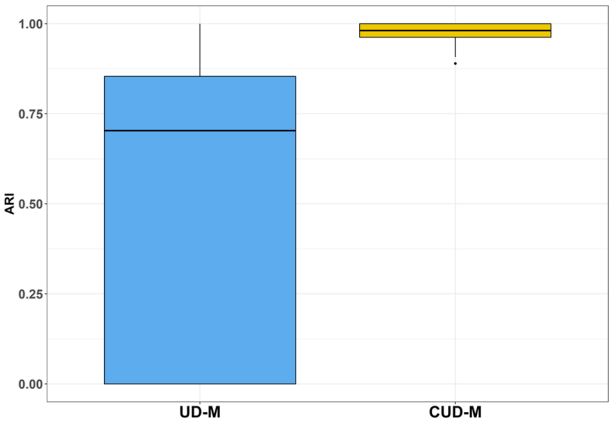


Fig. 5 Box plots of ARI values for the UD-M and CUD-M models over 500 simulated datasets

As concerns the results illustrated in Fig. 6, we notice that the detection rule for the CUD-M model provides almost optimal results in terms of FPR values since, with the exclusion of a few cases, they are regularly close to zero. Regarding the TPR values, we report that they have a median value of 0.66. The absence of convergence to one does not necessarily indicate an error: the manner in which atypical observations are incorporated into the data allows for the possibility that some of them will exhibit values akin to typical points, leading the CUD-M model to classify them as typical. For example, consider the dataset illustrated in Fig. 8 having the minimum TPR among the 500 datasets, i.e., 0.39. In detail, Fig. 8a is colored according to the true data classification, whereas Fig. 8b is colored according to the estimated classification by the CUD-M. The atypical points (true in Fig. 8a and estimated in Fig. 8b) are colored in red.

As we can see, the majority of random noisy points are located within the bulk of the data. Thus, it is reasonable that they are classified as typical observations by the CUD-M model.

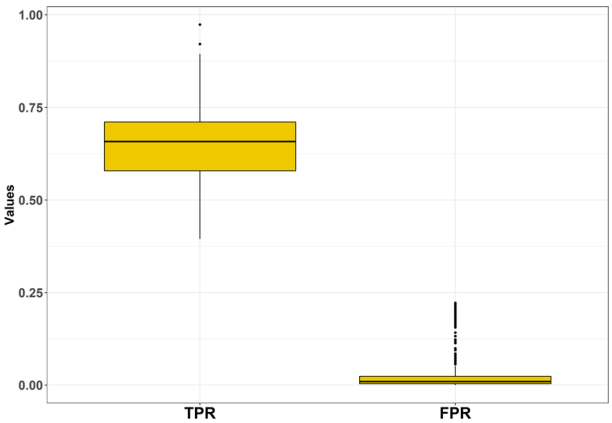


Fig. 6 Box plots of the TPR and FPR values for the CUD-M model over 500 simulated datasets

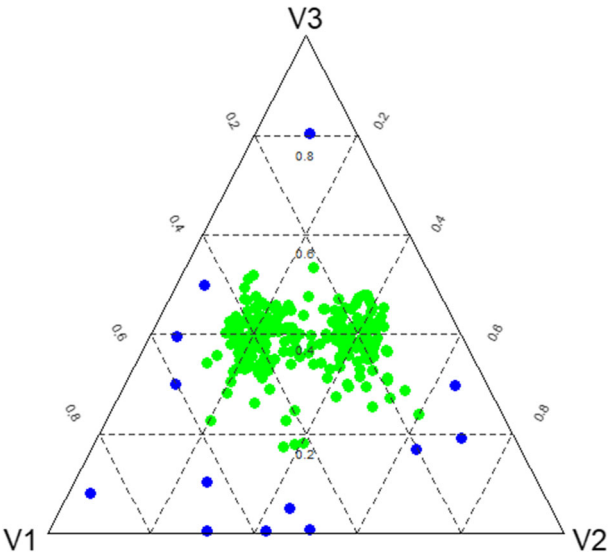


Fig. 7 Example of a simulated dataset colored according to the estimated classification by the UD-M model

4.3 Simulation Study 3

In this study, we simulate datasets from the CUD-M model in Eq. 19 after varying several factors in the data-generating process. In detail, we set $k = 2$ and vary the following factors:

1. the sample size ($n \in \{150, 300\}$),
2. the number of variables ($d \in \{3, 5\}$),
3. the mixing proportions ($\pi_1 = \pi_2$ and $\pi_1 \neq \pi_2$),
4. the degree of contamination ($\eta \in \{5, 10\}$).

Therefore, the result is a design with $2^4 = 16$ different data configurations. For each combination of these factors, we draw 100 datasets, resulting in a total of 1600 simulated datasets.

The parameters used for the simulation, for both values of d, are $\gamma_1 = \gamma_2 = 0.01$ and $\beta_1 = \beta_2 = 0.70$. As concerning the mixing proportions, when they are equal across the

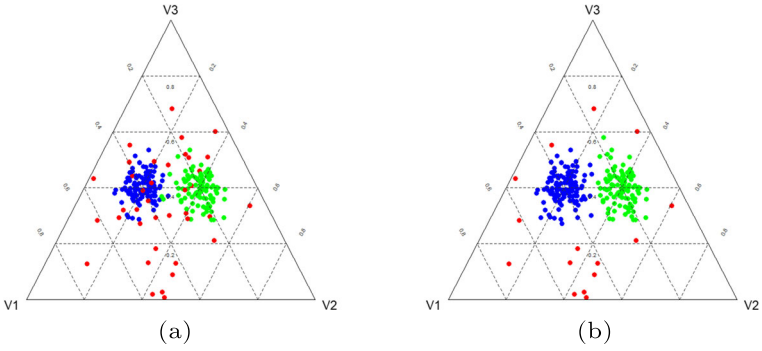


Fig. 8 Example of a simulated dataset colored according to the true data classification (**a**) and that estimated by the CUD-M (**b**). In red are the atypical points: true (**a**) and detected by the CUD-M (**b**)

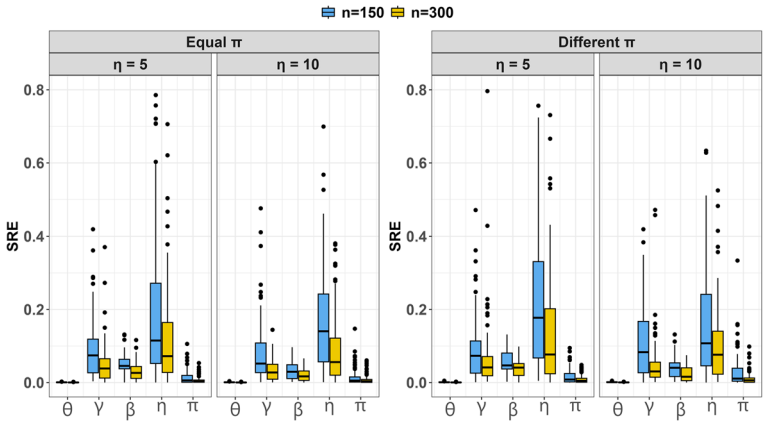


Fig. 9 Box plots of SRE values, over 100 simulated datasets for each data configuration, when $d = 3$

group, we set $\pi_1 = \pi_2 = 0.50$, while in the case they are different, we set $\pi_1 = 0.70$ and $\pi_2 = 0.30$. Lastly, when $d = 3$, we have $\boldsymbol{\theta}_1 = (0.3, 0.3)$ and $\boldsymbol{\theta}_2 = (0.2, 0.2)$, while when $d = 5$ we have $\boldsymbol{\theta}_1 = (0.30, 0.30, 0.10, 0.15)$ and $\boldsymbol{\theta}_2 = (0.20, 0.20, 0.20, 0.05)$.

For each simulated dataset, we initially fit the CUD-M model directly with $k = 2$ and assess parameter recovery using the square relative error (SRE). We recall that, for a generic parameter λ, $\text{SRE} = [(\hat{\lambda}_m - \lambda)/\lambda]^2$, where $\hat{\lambda}_m$ its estimate on the mth dataset. It is worth noting that, for simplicity in reporting results, we compute the average among the SREs of each estimated parameter across both groups, thereby summarizing the information for each parameter into a single value. Thus, for example, we will refer to $\boldsymbol{\theta}$ instead of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in this section.

Then, we fit the CUD-M model for $k \in \{1, \ldots, 3\}$, and use the Bayesian information criterion (BIC; Schwarz 1978) for assessing the capability of the BIC in detecting the correct

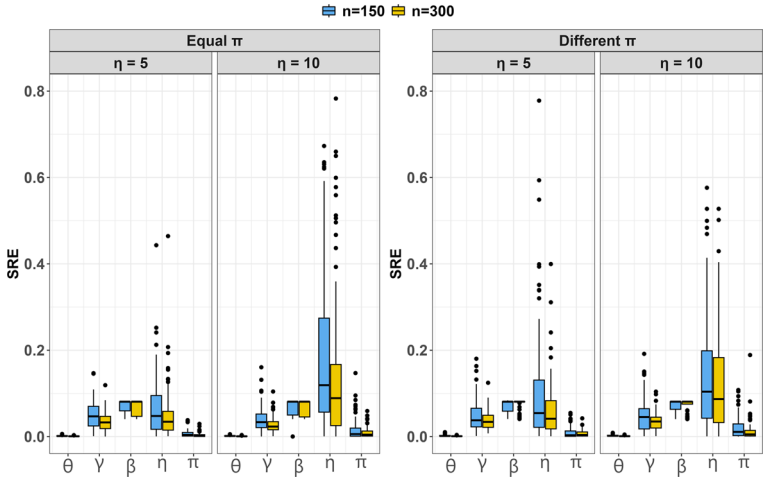


Fig. 10 Boxplots of SRE values, over 100 simulated datasets for each data configuration, when $d = 5$

mixture order. Note that, with the formulation of the BIC adopted herein, the higher the value, the better the fit.

Results for the parameter recovery are separately reported for each value of d in Figs. 9 and 10 via box plots, respectively. Each box plot summarizes the SRE values over the available 100 replications.

The estimation of $\boldsymbol{\theta}$ is nearly perfect across all data configurations, closely followed by the mixing proportions. After this, we note an increase in the SRE values for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Particularly, $\boldsymbol{\gamma}$ exhibits higher values and variability compared to $\boldsymbol{\beta}$ in configurations where $d = 3$, while the reverse is true for $d = 5$. This opposite trend might primarily stem from the diverse nature of data configurations created with the selected parameters, as achieving perfectly comparable scenarios across different factors poses significant challenges. Nevertheless, SRE values for both parameters appear relatively low. Conversely, estimating $\boldsymbol{\eta}$ seems more challenging, evident from its comparatively higher and more variable SRE values. However, there is a noticeable decline in SRE values with an increase in sample size, mirroring the pattern observed for other parameters.

Regarding the mixture order detection by the BIC, we report that the correct k has been regularly selected, with only three exceptions. In detail, $k = 1$ has been selected two times in the data configuration where $n = 150$, $d = 3$, $\pi_1 \neq \pi_2$ and $\eta = 10$, where as $k = 3$ has been selected one time when $n = 300$, $d = 5$, $\pi_1 = \pi_2$ and $\eta = 10$. Therefore, the BIC seems to correctly identify the true mixture order of the data-generating model.

5 Real Data Analyses

In this section, the models presented in Sect. 3 are fitted to two real datasets.

5.1 Aphyric Skye Lavas Data

The first dataset, called `skyeLavas`, is available within the **robCompositions** package (Templ et al., 2011). It contains atomic force microscopy compositions of $n = 23$ aphyric lavas of the Isle of Skye on the following $d = 3$ variables: (i) sodium-potassium, (ii) iron, and (iii) magnesium. All the variables are expressed as percentages.

This dataset has been introduced by Thompson et al. (1972), and subsequently analyzed in the literature by many authors (see, e.g., Aitchison, 1982; Barceló et al. 1996; Filzmoser and Hron, 2008; Fišerová and Hron, 2010). In the cited contributions, this dataset has been regularly investigated after the application of a log-ratio transformation. Barceló et al. (1996) and Filzmoser and Hron (2008) particularly focus on the outlier detection problem. Based on their results, Barceló et al. (1996) do not report any outliers, whereas Filzmoser and Hron (2008) identify two potential outliers. The methods employed to identify atypical points rely on Mahalanobis distances (refer to Sect. 1) and, as Filzmoser and Hron (2008) point out, the use of a different outlier cut-off value could result in these observations being within the bulk of the data.

Differently from the discussed contributions, we analyze this dataset directly in the simplex. Thus, we fit UD-M and CUD-M models for $k \in \{1, \ldots, 3\}$, and use the BIC to select the best-fitting model. Results are reported in Table 1.

The best-fitting model according to the BIC is the UD-M with $k = 3$. It is interesting to note that, differently from Barceló et al. (1996) and Filzmoser and Hron (2008), our findings

Table 1 `skyeLavas` dataset: log-likelihood, BIC, and number of parameters (# par) for each fitted model

Model	k	Log-likelihood	BIC	# par
UD-M	1	45.85	82.30	3
	2	66.22	110.48	7
	3	79.08	123.67	11
CUD-M	1	55.60	95.52	5
	2	69.33	104.17	11
	3	80.40	107.51	17

indicate an underlying group structure in the data. In detail, the detected groups are illustrated in Fig. 11.

The discussion of Thompson et al. (1972) seems to support the existence of multiple groups. Indeed, they argue that different kinds of basalts are present in the Isle of Skye, which differ by the proportions of major chemical elements. By using several experiments, they identify at least two main groups that, depending on the kind of experiment conducted, sometimes appear to be spaced out by intermediate lavas.

The estimated modes for the three groups are $\widehat{\boldsymbol{\theta}}_1 = (0.13, 0.52)$, $\widehat{\boldsymbol{\theta}}_2 = (0.46, 0.48)$, and $\widehat{\boldsymbol{\theta}}_3 = (0.22, 0.57)$. The first and the third groups (colored in blue and black in Fig. 11, respectively) have relatively close modes, whereas the second group appears to be more well-separated. Remarkably, the first group exhibits the highest magnesium mode value, the second group the highest sodium-potassium mode value, and the third group the highest iron mode value. This feature is also helpful in disclosing the different chemical compositions of each detected lava cluster. As concerns the dispersion parameter, we have $\widehat{\gamma}_1 = 0.002$, $\widehat{\gamma}_2 = 0.009$, and $\widehat{\gamma}_3 = 0.005$. Thus, the second group also stands out for having the greatest dispersion among the groups.

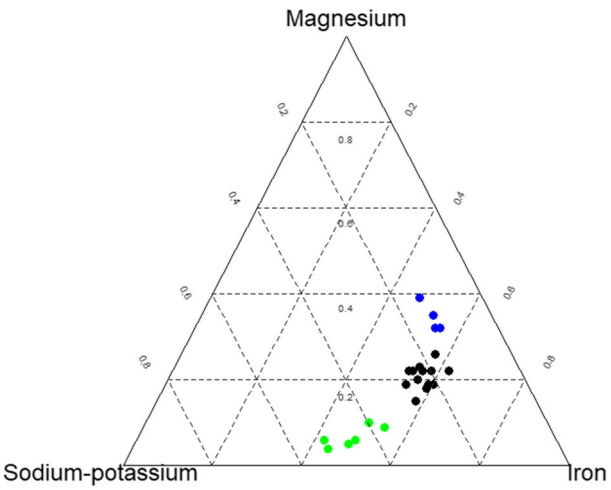


Fig. 11 Ternary diagram of the `skyeLavas` dataset, colored according to the classification estimated by the best-fitting model

Table 2 OECD dataset: log-likelihood, BIC, and number of parameters for each fitted model

Model	k	Log-likelihood	BIC	# par
UD-M	1	234.83	451.47	5
	2	305.03	570.05	11
	3	336.86	611.88	17
	4	346.42	609.18	23
CUD-M	1	271.54	517.62	7
	2	336.42	618.28	15
	3	342.69	601.72	23
	4	350.02	587.28	31

5.2 OECD Countries Data

The second dataset is freely available at https://stats.oecd.org/ and will be labeled herein as OECD. In particular, we analyze the labor force participation (LFP) for the $n = 38$ member countries of the Organization for Economic Co-operation and Development (OECD) in 2011. The LFP is available over the following $d = 5$ age classes: (i) $15-24$, (ii) $25-34$, (iii) $35-44$, (iv) $45-54$, and (v) $55-64$. The *closure* operation (Van den Boogaart and Tolosana-Delgado 2013) is implemented so that the LFP of each country sums up to 1.

We fit UD-M and CUD-M models for $k \in \{1, \ldots, 4\}$, and report the obtained results in Table 2.

The best-fitting model according to the BIC is the CUD-M with $k = 2$. Thus, differently from Sect. 5.1, there seems to be a group structure as well as atypical points in the data. In particular, we illustrate in Fig. 12 the scatter plot of the data, colored according to the classification estimated by the best-fitting model (in blue the first group and green the second group) and to the detected atypical points (in red).

As we note, there are two relatively separated clusters and several atypical points. In this regard, the estimated proportion of typical points and degree of contamination for the first group are $\widehat{\beta}_1 = 0.51$ and $\widehat{\eta}_1 = 26.60$, respectively, whereas, for the second group they are $\widehat{\beta}_2 = 0.99$ and $\widehat{\eta}_2 = 1.03$. Therefore, the atypical points are all accommodated by the first group, and their estimated a *posteriori* probabilities to be typical are reported in Table 3.

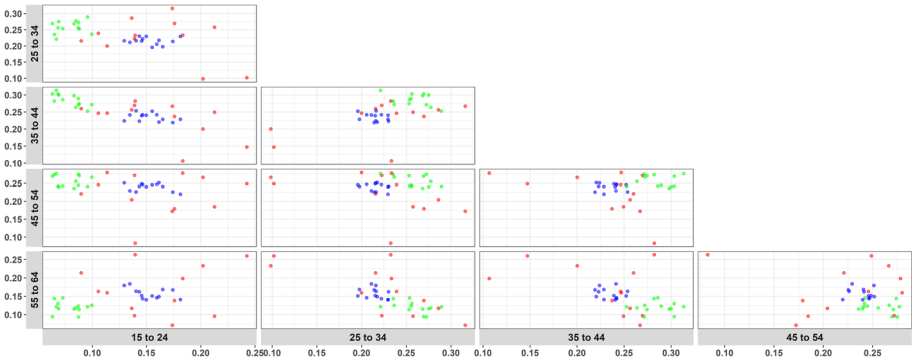


Fig. 12 Scatter plot of the OECD dataset, colored according to the classification estimated by the best-fitting model (in blue the first group and green the second group) and to the detected atypical points (in red)

Table 3 OECD dataset: countries detected as atypical using the CUD-M with $k = 2$, with corresponding a *posteriori* probability to be typical

Country	$P\left(\boldsymbol{x}_{-d} \text{ is typical}\right)$
Estonia	3.06×10^{-2}
Germany	6.25×10^{-2}
Israel	3.46×10^{-5}
Austria	3.25×10^{-6}
Ireland	3.99×10^{-6}
Japan	2.98×10^{-9}
Mexico	4.30×10^{-13}
Costa Rica	1.22×10^{-25}
Turkiye	7.94×10^{-27}
Colombia	5.19×10^{-29}
Latvia	9.00×10^{-44}
Lithuania	1.56×10^{-49}

We immediately notice that the probability values differ among the countries, with some of them being more extreme than others. For example, the two most atypical countries are Lithuania and Latvia, respectively. To visualize these countries, we can look at the subplots of Fig. 12, even though each of them provides only partial information about the whole data structure. As concerns Lithuania, it is the isolated point at the bottom of the subplots having the age class 45–54 as the y-axis. Additionally, it is one of the two points at the top of the subplots having the age class 55–64 as the y-axis. Regarding Latvia, it is the point in the bottom-right corner of the subplot comparing the 25–34 and 15–24 age classes, and one of the two points at the top of the subplots having the age class 55–64 as the y-axis. Similar conclusions can be drawn for the other countries listed in Table 3.

The estimated modes for the two groups are $\widehat{\boldsymbol{\theta}}_1 = (0.15, 0.22, 0.23, 0.24)$ and $\widehat{\boldsymbol{\theta}}_2 = (0.08, 0.26, 0.28, 0.26)$. Thus, the first group (colored in blue in Fig. 12) has relatively higher modes for the 15–24 and 55–64 age classes, i.e., the younger and older people, respectively, while the second group (colored in green in Fig. 12) has higher mode values in the remaining age classes. As concerns the dispersion parameter, we have $\widehat{\gamma}_1 = 0.001$ and $\widehat{\gamma}_2 = 0.002$. Therefore, the two groups seem to have a similar dispersion.

6 Conclusions

A unimodal parametrization of the Dirichlet distribution has been considered for robustness, model-based clustering, and both together. In particular, we first proposed the use of unimodal Dirichlet distribution for model-based clustering since a modal approach allows for a natural and convenient interpretation of the detected clusters. Secondly, we introduce the contaminated unimodal Dirichlet distribution that, because of its heavy tails, conveniently allows for the modeling of compositional data with atypical observations directly in the simplex, in an opposite manner to what is traditionally done in the related literature that relies on log-ratio transformations. A useful aspect of this distribution is its ability to automatically identify atypical observations. Thirdly, we proposed the use of the contaminated unimodal Dirichlet distribution for model-based clustering, thus jointly accounting for the presence of

clusters and atypical observations in the data. Similarly to before, we can identify atypical observations after their assignment to the considered clusters.

From a computational point of view, a direct ML approach (for the contaminated unimodal Dirichlet distribution) and two EM algorithms (for unimodal and contaminated unimodal Dirichlet mixtures) have been also illustrated. Unfortunately, closed-form expressions for estimating the parameters are not available, requiring the implementation of numerical optimization methods.

The usefulness of the proposed methodologies has been demonstrated both via simulation studies and real data applications. Specifically, the first simulation study showed the robustness of the contaminated Dirichlet distribution in the presence of atypical points. The second simulated analysis illustrated the benefits of using the contaminated Dirichlet mixture model, in terms of classification and atypical point detection, when noisy observations are present in the data. The third simulation study disclosed the adequacy of parameter and mixture order recovery.

Regarding real case studies, we considered two datasets. In the first one, we illustrated the presence of a clustering structure, modeled via mixtures of unimodal Dirichlet distributions. This result contrasts with what has been done in the literature, where this dataset has always been considered as having a single group with atypical points. From an interpretative point of view, the detected clusters consist of different kinds of basalts present in the Isle of Skye, which differ by the proportions of major chemical elements.

The second dataset provided an example of a grouping structure with atypical points. Indeed, the fitting of the contaminated Dirichlet mixture model was the best according to the BIC. By using the features of this model, it has been possible to automatically detect and evaluate the atypical points assigned to one of the two clusters.

There are different possibilities for further work, four of which are worth mentioning. Firstly, the CUD-M model we propose follows a "componentwise" approach to protect the UD-M against the presence of atypical observations. Under the alternative "additional component" approach, and borrowing the famous idea from Banfield and Raftery (1993), the UD-M model could be protected against atypical observations via an "additional component" approach where a uniform distribution across the simplex is used as an additional component. The uniform component would hopefully capture the contaminated observations. Secondly, covariates are often available along with compositional data. A straightforward extension of the proposed methodology would deal with the regression framework, where the response random vector is conditionally distributed according to the CUD(-M) model. Thirdly, parsimony is a fundamental aspect of statistical modeling. Parsimony could be achieved by forcing the parameters (excluding the modes) to be equal across mixture components. Finally, extensions to hidden Markov models can be considered.

Acknowledgements Punzo and Tomarchio acknowledge the support of MUR, grant no. 2022XRHT8R (CUP: E53D23005950006) - The SMILE project: Statistical Modelling and Inference to Live the Environment, funded by the European Union - Next Generation EU. Bekker and Ferreira acknowledge the support of the National Research Foundation (NRF) of South Africa (SA), ref. SRUG2204203865 (Bekker) and ref. RA201125576565 - no. 145681 (Ferreira), the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa (grant nr. 2022-047-STA and 2024-033-STA), and the Department of Research and Innovation at the University of Pretoria (SA). The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Data Availability The real datasets used in this manuscript are publicly available as described in the manuscript.

Declarations

Competing Interests The authors declare no competing interests.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44*(2), 139–160.

Aitchison, J., & Lauder, I. (1985). Kernel density estimation for compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 34*(2), 129–137.

Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the *k*-bumps algorithm. *Computational Statistics, 28*(4), 1571–1597.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803–821.

Barceló, C., Pawlowsky, V., & Grunsky, E. (1996). Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology, 28*, 501–518.

Bertin, K., Genest, C., Klutchnikoff, N., et al. (2023). Minimax properties of Dirichlet kernel density estimators. *Journal of Multivariate Analysis, 195*(105), 158.

Botha, T., Ferreira, J., & Bekker, A. (2021). Alternative Dirichlet priors for estimating entropy via a power sum functional. *Mathematics, 9*(13), 1493.

Bouguila, N., Ziou, D., & Vaillancourt, J. (2004). Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application. *IEEE Transactions on Image Processing, 13*(11), 1533–1543.

Brazier, S., Sparks, R. S. J., Carey, S. N., et al. (1983). Bimodal grain size distribution and secondary thickening in air-fall ash layers. *Nature, 301*, 115–119.

Calif, R., Emilion, R., & Soubdhan, T. (2011). Classification of wind speed distributions using a mixture of Dirichlet distributions. *Renewable Energy, 36*(11), 3091–3097.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review, 88*(1), 122–141.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis, 31*(2), 131–145.

Chen, S. X. (2000). Beta kernel smoothers for regression curves. *Statistica Sinica, 10*(1), 73–91.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association, 88*(423), 782–792.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: series B, 39*(1), 1–22.

Filzmoser, P., & Gregorich, M. (2020). Multivariate outlier detection in applied data analysis: Global, local, compositional and cellwise outliers. *Mathematical Geosciences, 52*(8), 1049–1066.

Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences, 40*, 233–248.

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer.

Fišerová, E., & Hron, K. (2010). Total least squares solution for compositional data using linear models. *Journal of Applied Statistics, 37*(7), 1137–1152.

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

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York: Springer.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lochner, R. H. (1975). A generalized Dirichlet distribution in Bayesian life testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 37*(1), 103–113.

McLachlan, G. J., Basford, K. E. (1988) Mixture models: Inference and applications to clustering. *Statistics: A Series of Textbooks and Monographs*, Marcel Dekker, New York
McNicholas, P. D. (2016). *Mixture model-based classification*. CRC Press.
Murphy, E. A. (1964). One cause? Many causes? The argument from the bimodal distribution. *Journal of Chronic Diseases, 17*(4), 301–324.
Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313.
Ng, K. W., Tian, G. L., & Tang, M. L. (2011). *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons.
Nolan, J. P. (1998). Parameterizations and modes of stable distributions. *Statistics & Probability Letters, 38*(2), 187–195.
Ongaro, A., & Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis, 114*, 412–426.
Ongaro, A., Migliorati, S., & Ascari, R. (2020). A new mixture model on the simplex. *Statistics and Computing, 30*, 749–770.
Ouimet, F., & Tolosana-Delgado, R. (2022). Asymptotic properties of Dirichlet kernel density estimators. *Journal of Multivariate Analysis, 187*(104), 832.
Pal, S., Heumann, C. (2022) Clustering compositional data using Dirichlet mixture model. Plos one *17*(5):e0268,438
Pawlowsky-Glahn, V., Buccianti, A. (2011) Compositional data analysis. *Wiley Online Library*
Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*, 339–348.
Punzo, A. (2019). A new look at the inverse Gaussian distribution with applications to insurance and economic data. *Journal of Applied Statistics, 46*(7), 1260–1287.
Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal, 58*(6), 1506–1537.
R Core Team (2021) R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria, https://www.R-project.org/
Ray, S., & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Annals of Statistics, 33*(5), 2042–2065.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
Templ, M., Hron, K., & Filzmoser, P. (2011). *robCompositions: An R-package for robust statistical analysis of compositional data*. John Wiley and Sons.
Thompson, R., Esson, J., & Dunham, A. (1972). Major element chemical variation in the Eocene lavas of the Isle of Skye. *Scotland. Journal of Petrology, 13*(2), 219–253.
Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: John Wiley & Sons.
Tomarchio, S. D., & Punzo, A. (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(4), 1247–1266.
Tomarchio, S. D., & Punzo, A. (2020). Dichotomous unimodal compound models: Application to the distribution of insurance losses. *Journal of Applied Statistics, 47*(13–15), 2328–2353.
Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R,* (Vol. 122). Springer.