A structured Dirichlet mixture model for compositional data: inferential and applicative issues

Sonia Migliorati[1,2] · Andrea Ongaro[1,2] · Gianna S. Monti[1,2]

Received: 7 September 2015 / Accepted: 29 April 2016 / Published online: 19 May 2016

Abstract The flexible Dirichlet (FD) distribution (Ongaro and Migliorati in J. Multvar. Anal. 114: 412–426, 2013) makes it possible to preserve many theoretical properties of the Dirichlet one, without inheriting its lack of flexibility in modeling the various independence concepts appropriate for compositional data, i.e. data representing vectors of proportions. In this paper we tackle the potential of the FD from an inferential and applicative viewpoint. In this regard, the key feature appears to be the special structure defining its Dirichlet mixture representation. This structure determines a simple and clearly interpretable differentiation among mixture components which can capture the main features of a large variety of data sets. Furthermore, it allows a substantially greater flexibility than the Dirichlet, including both unimodality and a varying number of modes. Very importantly, this increased flexibility is obtained without sharing many of the inferential difficulties typical of general mixtures. Indeed, the FD displays the identifiability and likelihood behavior proper to common (non-mixture) models. Moreover, thanks to a novel non random initialization based on the special FD mixture structure, an efficient and sound estimation procedure can be devised which suitably combines EM-types algorithms. Reliable complete-data likelihood-based estimators for standard errors can be provided as well.

Keywords Simplex distribution · Dirichlet mixture · Identifiability · Multimodality · EM type algorithms

✉ Sonia Migliorati
sonia.migliorati@unimib.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

[2] NeuroMi - Milan Center for Neuroscience, Milan, Italy

1 Introduction

Compositional data consist of vectors with strictly positive elements subject to a unit-sum constraint (i.e. proportions of some whole), and they are prevalent in many disciplines (e.g. geology, medicine, economics, psychology, environmetrics etc.). Modeling compositional data requires the choice of a distribution defined on the proper bounded domain: the simplex. One of the most well-known distributions on the simplex is the Dirichlet, which shows a remarkable mathematical tractability as well as easiness of parameter interpretation. Though, it is often unsatisfactory for dealing with compositional data, due to its poor parametrization as well as to its implied extreme forms of independence. Indeed, the Dirichlet distribution possesses all common independence concepts developed for compositional data, as discussed in Aitchison (2003), according to whom the Dirichlet "has the ultimate independence structure for compositions".

Also in light of such inadequacies, Aitchison (2003) proposed a powerful and far-reaching approach, which is based on mapping the variables from the simplex to an (unconstrained) Euclidean space by means of various logratio transformations (see also Pawlowsky-Glahn et al. 2015). The transformed variables are then modeled as multinormal. This approach exhibits many elegant mathematical properties, and it largely enriches the set of tools for compositional data treatment, by resorting to multivariate techniques available on the transformed unconstrained space. Nevertheless, it does not produce parametric families of distributions containing the Dirichlet one, i.e. with extreme independence properties. Moreover, it encounters difficulties in modeling some relevant forms of simplicial independence, e.g. neutrality (see Connor and Mosimann 1969), subcompositional invariance and partition independence (see Aitchison

2003). In addition, it does not allow to deal with an important compositional operation, namely amalgamation, i.e. sums of composition elements aimed at grouping homogeneous parts of the whole.

Alternatively, one can define models directly on the simplex, looking for appropriate generalizations of the Dirichlet distribution. This paper is concerned with this second approach. In this direction, various proposals are present in the literature (see Connor and Mosimann 1969; Barndorff-Nielsen and Jørgensen 1991; Favaro et al. 2011; Gupta and Richards 1987, 1991, 1992, 1995, 1997, 2001a, b; Rayens and Srinivasan 1994; Smith and Rayens 2002).

Recently, a new generalization of the Dirichlet has been proposed: the flexible Dirichlet (FD) (Ongaro and Migliorati 2013). This distribution enables considerable flexibility in modeling dependence as well as various independence concepts appropriate for compositional data. In particular, through different parameter configurations, it is able to discriminate among subcompositional independence, left and right neutrality, subcompositional invariance as well as complete and high order partition independences (for a definition of such independences see Aitchison 2003). At the same time, the FD displays several probabilistic and compositional properties, thus guaranteeing a remarkable tractability from a theoretical viewpoint. In particular, we recall its representation as a special finite mixture of a varying number of Dirichlet distributions, closure under marginalization, conditioning, subcomposition, amalgamation and permutation, and explicit expressions of joint and conditional moments.

Compared to other distributions proposed on the simplex, the FD displays a dependence structure substantially richer than Connor and Mosimann (1969), Barndorff-Nielsen and Jørgensen (1991), Favaro et al. (2011), Gupta and Richards (1987, 1991, 1992, 1995, 1997, 2001a, b), and it shows a greater tractability than the generalized Liouville distribution (see Rayens and Srinivasan 1994; Smith and Rayens 2002), though with a similar capability of modeling dependence/independence. For further details on the FD and a thorough discussion of the literature see Ongaro and Migliorati (2013).

In this paper we shall focus on understanding the real potential of the FD from an inferential and applicative perspective. To this end, we shall analyze its peculiar mixture structure, study theoretical and specially computational issues deriving from its estimation, and carry out an extensive evaluation of its performance through simulations and applications to a multifaceted real data set.

The first fundamental aspect we shall address is the nature of the FD mixture structure. Generally speaking, a unimodal parametric model (the Dirichlet in our case) is often too simplistic to describe real data patterns. On the other hand, the general mixture of such unimodal parametric models (i.e. with arbitrary component parameters) may be unnecessarily involved, leading to unwanted inferential complications. Thus, it seems appealing to look for models in between such extremes, obtained by imposing suitable links among the component parameters (“structured” mixture models). The FD is one such model in the context of compositional data. The particular very strict links defining the FD will be shown to entail a number of distinguishing inferential features. More precisely, they determine a simple and clearly interpretable differentiation among mixture components. Despite their tightness, they allow the FD to display substantially greater flexibility than the Dirichlet, in particular in terms of density shapes and dependence structure. Furthermore, this increased flexibility is obtained without inheriting many of the inferential difficulties typical of general mixtures.

All such aspects are tackled in the remainder of the paper, as follows. After briefly recalling the definition and some properties of the FD (Sect. 2), we show that the FD entails a very clear and simple geometric interpretation of the cluster means and of the classification rule assigning observations to clusters (Sect. 3). From an inferential point of view, the special mixture structure of the FD guarantees identifiability in the classic strong sense, boundedness of the likelihood and existence of a global maximum (Sect. 4), properties which, typically, are not shared by general mixtures. In Sect. 5 we consider the EM algorithm and some of its variants to improve the performance of the estimation procedure. A complete-data likelihood-based algorithm is proposed in Sect. 6 to reliably estimate the (asymptotic) variance covariance matrix. In Sect. 7 great effort is devoted to the choice of the initial values of the EM algorithm, a crucial aspect for mixture estimation. We propose novel non random initialization strategies, which take advantage of the special FD mixture structure. The final suggested algorithm is chosen by means of an extensive simulation study. Moreover, the behavior of the MLE (as computed with the devised algorithm) and of the standard error estimators are tackled via simulation. Section 8 analyzes the applicative capacity of the FD by fitting a rich real data set showing a large variety of patterns. Finally, we make some concluding remarks in Sect. 9. Proofs of the lengthiest results are set out in the Appendix. All statistical computations were performed using the R software R Development Core Team (2015).

2 The flexible Dirichlet distribution

The FD distribution derives from the normalization of a basis of $D (D > 1)$ positive dependent random variables obtained by starting from the usual basis of independent equally scaled gamma random variables (i.e. the Dirichlet basis) and randomly allocating to the i-th element a further independent gamma random variable.

Its density function can be expressed as

$$f_{FD}(\mathbf{x};\boldsymbol{\theta}) = \frac{\Gamma\left(\alpha^{+}+\tau\right)}{\prod_{h=1}^{D}\Gamma(\alpha_h)}\left(\prod_{h=1}^{D}x_h^{\alpha_h-1}\right) \times \sum_{i=1}^{D} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau)}x_i^{\tau} \tag{1}$$

with $\mathbf{x} \in \mathcal{S}^D$, i.e. the simplex

$$\mathcal{S}^D = \left\{\mathbf{x} \in \mathbb{R}^D : x_i > 0\,, i = 1,\ldots,D,\ \sum_{i=1}^{D} x_i = 1\right\}, \tag{2}$$

$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{p}, \tau)$ is the complete set of parameters where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_D)$, $\mathbf{p} = (p_1,\ldots,p_D)$ and $\alpha^{+} = \sum_{i=1}^{D}\alpha_i$. The parameter space is $\boldsymbol{\Theta} = \{(\boldsymbol{\alpha},\mathbf{p},\tau) : \alpha_i > 0, 0 \le p_i < 1, i = 1,\ldots,D, \sum_{i=1}^{D} p_i = 1, \tau > 0\}$. The FD distribution will be denoted by $\mathcal{FD}(\boldsymbol{\theta})$.

Note that both (1) and (2) are defined symmetrically with respect to all D variables forming the composition. This makes it possible to treat all the variables alike. In this way the simplex $\mathcal{S}^D$ is a $D-1$-dimensional object embedded in $\mathbb{R}^D$. Therefore, formally the function (1) is not a density with respect to the Lebesgue measure on $\mathbb{R}^D$, but with respect to a suitable uniform measure on $\mathcal{S}^D$. It can be transformed into a standard Lebesgue density on $\mathbb{R}^{D-1}$ by replacing, for example, the last element x_D with $1 - x_1 - \cdots - x_{D-1}$.

The FD contains the Dirichlet distribution as the inner point $\tau = 1$ and $p_i = \alpha_i/\alpha^{+}, \forall i = 1,\ldots,D$.

A key feature of the FD is that it can be written as a finite mixture of Dirichlet distributions:

$$f_{FD}(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{D} p_i f_D(\mathbf{x};\boldsymbol{\alpha}_i) \tag{3}$$

where

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha} + \tau\mathbf{e}_i \tag{4}$$

and $\mathbf{e}_i$ is a canonical vector whose elements are all equal to 0 except for the i-th one which is 1. Here

$$f_D(\mathbf{x};\boldsymbol{\alpha}) = \frac{\Gamma\left(\alpha^{+}\right)}{\prod_{h=1}^{D}\Gamma(\alpha_h)}\left(\prod_{h=1}^{D}x_h^{\alpha_h-1}\right) \tag{5}$$

is the density of a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ (hereafter denoted by $\mathcal{D}(\boldsymbol{\alpha})$). It is noteworthy that the FD parameter space makes it possible to have a varying (up to D) number of components. Moreover, this mixture representation, among other aspects, allows a variety of different shapes to be assumed by the density, including both unimodality and multimodality.

Many relevant properties of the FD can be found in Ongaro and Migliorati (2013) as mentioned in Sect. 1. For future purposes, here we recall that, although the FD allows for more general dependence relationships than the Dirichlet, it shares negative correlations with the latter. Note that this limitation is less critical for compositional data, because the unit-sum constraint tends to induce negative dependence.

3 Mixture structure and clustering within the FD model

In this section we shall fully explore the special cluster structure implied by the finite mixture representation (3) of the FD model, assuming the usual correspondence between clusters and components of the mixture (hereafter "component" will be reserved to indicate an element of a mixture). To this end, we preliminarily describe (in the following subsection) a useful symmetric representation of the simplex, which will play a key role throughout the paper, being used to visualize data as well as to initialize the estimation algorithm.

3.1 Regular simplicial representation

To deal with compositional data, and particularly to graphically display them, it is advisable to work with symmetric representations, which ensure that all D variables forming the composition are treated alike. Symmetry means invariance with respect to element permutations, i.e. whatever the permutation of the variables, any subset of the simplex is mapped onto sets which are congruent (i.e. coincide after rigid motions), so that distances among points are preserved.

Note that this property is not possessed by common simplex representations. The problem is that for visual inspection (as well as for all problems where the effective dimensionality is a fundamental issue) it is obviously more convenient to represent the original D variables in a $D-1$-dimensional space. This typically generates asymmetry, due to the choice of the "left-out" variable. For example, the standard representation, obtained by choosing $D-1$ arbitrarily selected variables of the composition as coordinates of the space, is asymmetric unless $D = 2$.

Our aim is therefore to find a symmetric $D-1$-dimensional representation. The simplex $\mathcal{S}^D$, as defined in (2), is a regular polytope (Coxeter 1973) lying on the $D-1$-dimensional affine subspace of $\mathbb{R}^D$ determined by its vertexes $\mathbf{e}_1, \ldots, \mathbf{e}_D$ (canonical basis). In general, the convex hull of $k+1$ vertexes $\mathbf{v}_0,\ldots,\mathbf{v}_k \in \mathbb{R}^D$ $(1 \le k \le D)$ all with the same distance between each other, will be called k-

dimensional regular simplicial polytope (embedded in $\mathbb{R}^D$). Hence we need to determine a $D-1$-dimensional regular simplicial polytope (RSP^{D-1}) in $\mathbb{R}^{D-1}$. In particular, RSP^1 is a line segment in $\mathbb{R}^1$, RSP^2 is an equilateral triangle in $\mathbb{R}^2$ and RSP^3 is a regular tetrahedron in $\mathbb{R}^3$. Note that RSP^2 coincides with the so called ternary diagram, commonly used in applications to display 3-part compositions.

The vertexes $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{D-1} \in \mathbb{R}^{D-1}$ of RSP^{D-1} with common distance - say 1 - can be derived recursively as proved in Appendix 1. Choosing the line segment with vertexes $\mathbf{v}_0 = 0$ and $\mathbf{v}_1 = 1$ as a one dimensional simplex RSP^1, we have that $\mathbf{v}_0$ is the origin and $(\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ is an upper triangle matrix with i-th diagonal element equal to c_i and all other non null elements of the i-th row equal to a_i. Here $c_i = \sqrt{(1+i)/2i}$ and $a_i = 1/\sqrt{2i(1+i)}$, $(i = 1, \ldots, D-1)$. It also follows that $\mathbf{y} = [\mathbf{v}_0 \ \ldots \ \mathbf{v}_{D-1}]\mathbf{x}$ is a linear transformation mapping $\mathbf{x} \in \mathcal{S}^D$ onto $\mathbf{y} \in RSP^{D-1}$.

3.2 FD mixture structure

To better appreciate the mixture structure implied by the FD, we shall use a different parametrization in terms of mean and precision parameters for its Dirichlet components. More specifically, the Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$ defined in (5) will be reparametrized as $\mathcal{D}'(\overline{\boldsymbol{\alpha}}, \alpha^+)$ where $\overline{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$ is the mean vector and α^+ is a common precision parameter. Indeed, if $\mathbf{X} \sim \mathcal{D}(\boldsymbol{\alpha})$ then

$$\mathrm{Var}(X_i) = \frac{\overline{\alpha}_i(1-\overline{\alpha}_i)}{1+\alpha^+} \quad i = 1, \ldots, D.$$

Furthermore, the maximum variance of a random variable on [0, 1], given its mean μ, is $\mu(1-\mu)$. It follows that the factor $1/(1+\alpha^+)$ represents the common normalized variance of all X_i's for given means $\overline{\alpha}_i$'s.

In the new parametrization the FD distribution can be written in terms of the parameters $w = \tau/(\alpha^+ + \tau) \in (0, 1)$, $\gamma = \alpha^+ + \tau > 0$ and $\overline{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+ \in \mathcal{S}^D$ as

$$\sum_{i=1}^{D} p_i \mathcal{D}'(\boldsymbol{\mu}_i, \gamma)$$

where

$$\boldsymbol{\mu}_i = (1-w)\,\overline{\boldsymbol{\alpha}} + w\,\mathbf{e}_i. \tag{6}$$

Thus, the FD is a mixture of Dirichlet distributions with common precision parameter γ (determining the normalized variances) and different mean vectors $\boldsymbol{\mu}_i$. Apart from its component mean structure (which we shall shortly consider), the FD can be interpreted, in many respects, as the counterpart on the simplex to the basic mixture of multivariate normals with independent elements sharing the same overall marginal variance. Indeed, the Dirichlet is typically unimodal (when all parameters are bigger than 1) and it entails the most extreme types of independence for compositional data.

Let us now focus on the structure (6) of the means $\boldsymbol{\mu}_i$ of the Dirichlet components. They are linear convex combinations of a common "barycenter" $\overline{\boldsymbol{\alpha}}$ and the i-th vertex $\mathbf{e}_i$. Thus, the vector mean of the generic i-th component is characterized by a value of the i-th element higher than the corresponding element of the other component means. This introduces a very simple and reasonable form of differentiation among components, which is able to capture a broad range of cluster dissimilarities.

Moreover, the parameter w dictates how far the component means are from each other and from the barycenter $\overline{\boldsymbol{\alpha}}$ in direction of $\mathbf{e}_i$. It is the only parameter that the FD devotes to regulating all the relationships among components, which highlights the extremely structured nature of the FD mixture.

The following result provides a full geometric characterization of the mean vectors structure.

Proposition 1 (Mean structure of mixture components) *The D mean vectors $\boldsymbol{\mu}_i$ given by (6) are the vertexes of a $D-1$-dimensional regular simplicial polytope $\mathcal{S}^D_w$ strictly contained in the original $\mathcal{S}^D$, with edges parallel and proportional by a factor w to the $\mathcal{S}^D$ ones. Conversely, given any such regular simplicial polytope $\mathcal{S}^D_w$, there exists a unique value of the parameter $(\overline{\boldsymbol{\alpha}}, w)$ such that the vertexes of $\mathcal{S}^D_w$ coincide with the means $\boldsymbol{\mu}_i$.*

Proof Any representation of the simplex $\mathcal{S}^D$ can be used to prove the result, because the above characterization is a geometrical one. It is convenient, for symmetry reasons, to consider $\mathcal{S}^D$ as a $D-1$-dimensional object of $\mathbb{R}^D$. Specifically, let $\mathbf{e}_1, \ldots, \mathbf{e}_D$ (canonical basis of $\mathbb{R}^D$) be the vertexes of $\mathcal{S}^D$. Then, by (6), we have that

$$\boldsymbol{\mu}_i - \boldsymbol{\mu}_j = w(\mathbf{e}_i - \mathbf{e}_j),$$

$\forall i \neq j$, i.e. all edges of the polytope $\mathcal{S}^D_w$ with vertexes $\boldsymbol{\mu}_i$ are parallel to the $\mathcal{S}^D$ ones and have length w.

Let us now show that $\mathcal{S}^D_w$ is strictly contained in $\mathcal{S}^D$. Formally, this means that $\mathcal{S}^D_w$ is contained in the relative interior of $\mathcal{S}^D$ (relative since $\mathcal{S}^D$ does not have full dimension D). To this end it is enough to show that the vertexes $\boldsymbol{\mu}_i$ of $\mathcal{S}^D_w$ belong to the relative interior of $\mathcal{S}^D$. This is true because all the elements of $\boldsymbol{\mu}_i$ $(i = 1, \ldots, D)$ are strictly positive.

Conversely, suppose that $\mathcal{S}^D_w$ is a regular simplicial polytope contained in the relative interior of $\mathcal{S}^D$. Let $\boldsymbol{\nu}_i$ denote its vertexes and suppose that its edges are parallel to the $\mathcal{S}^D$ ones. Then, the $\mathcal{S}^D$ edges all have lengths proportional to the corresponding $\mathcal{S}^D_w$ ones, i.e. $\forall i \neq j$ the equality $\boldsymbol{\nu}_i - \boldsymbol{\nu}_j = c\,(\mathbf{e}_i - \mathbf{e}_j)$ must hold for some $0 < c < 1$. This implies $\boldsymbol{\nu}_i - c\,\mathbf{e}_i = \boldsymbol{\nu}_j - c\,\mathbf{e}_j = \boldsymbol{\nu}^*$, and, therefore,

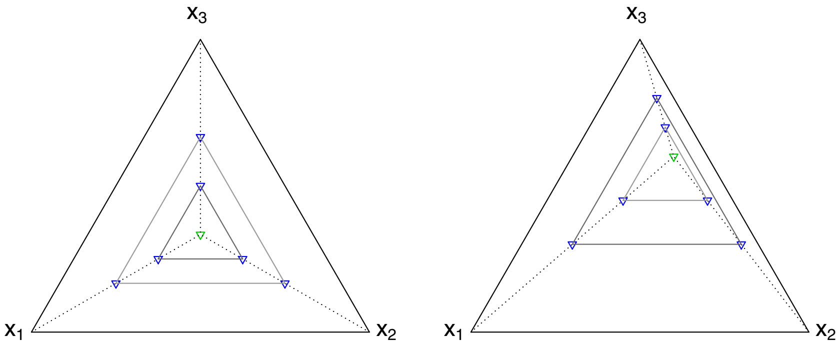

Fig. 1 Component mean structure. *Left panel* $\overline{\boldsymbol{\alpha}} = (1/3, 1/3, 1/3)$, $w = 0.25$ and $w = 0.5$. *Right panel* $\overline{\boldsymbol{\alpha}} = (0.1, 0.3, 0.6)$, $w = 0.25$ and $w = 0.5$

$\boldsymbol{\nu}_i = \boldsymbol{\nu}^* + c\,\mathbf{e}_i$ $\forall i$. Since $\boldsymbol{\nu}_i$ must belong to the relative interior of $\mathcal{S}^D$, its elements ν_{ij} must satisfy $\sum_j \nu_{ij} = 1$ and $\nu_{ij} > 0$, $j = 1, \ldots, D$. Therefore, $\boldsymbol{\nu}_i$ can be written uniquely as $\boldsymbol{\nu}_i = (1-c)\overline{\boldsymbol{\nu}}^* + c\,\mathbf{e}_i$ where the elements $\overline{\nu}_j^*$ of $\overline{\boldsymbol{\nu}}^*$ are such that $\sum_j \overline{\nu}_j^* = 1$ and $\overline{\nu}_j^* > 0$. Thus, such vertexes have the same structure as the means $\boldsymbol{\mu}_i$ given by (6). □

For example, when $D = 3$ the three component means can generate any equilateral triangle inscribed in the equilateral triangle RSP^2 with parallel edges. In Fig. 1 the component means are displayed for a symmetric case with equal $\overline{\alpha}_i$'s (left panel) and an asymmetric one (right panel). In both cases two different values of w have been considered to show how the means move along the segments (dashed line) connecting the barycenter $\overline{\boldsymbol{\alpha}}$ with the vertexes.

3.3 Clusters overlapping

In cluster analysis it is often of interest to determine the degree of overlap among the mixture components representing the clusters. From Sect. 3.2, the larger w, i.e. the larger τ with respect to α^+, the more distinct the component means. A precise evaluation of the overlapping can be obtained by means of a suitable distance between densities. A frequent choice is the symmetrized Kullback–Leibler divergence:

$$d_{SKL}(f_1, f_2) = d_{KL}(f_1, f_2) + d_{KL}(f_2, f_1)$$

where

$$d_{KL}(f_1, f_2) = \int f_1(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x}$$

and f_1 and f_2 are two arbitrary positive densities. This distance between two Dirichlet components of the FD can be explicitly computed as:

$$d_{SKL}(f_D(\boldsymbol{\alpha}_i), f_D(\boldsymbol{\alpha}_h)) = k(\alpha_i, \tau) + k(\alpha_h, \tau) \qquad (7)$$

where $\boldsymbol{\alpha}_i$ is given by (4),

$$k(\alpha, \tau) = \tau[\psi(\alpha + \tau) - \psi(\alpha)] \qquad (8)$$

and $\psi(\cdot)$ denotes the digamma function. In practice, a graphical investigation shows that values of (7) roughly larger than 10 already produce two appreciably separated components.

The behavior of the distance d_{SKL} can be understood by studying the function $k(\alpha, \tau)$. By applying properties of the digamma and trigamma (i.e. digamma derivative) functions, one can see that $k(\alpha, \tau)$ is increasing in τ ranging from 0 to ∞. Moreover, it is decreasing in α ranging from 0 to ∞.

3.4 FD classification rule

The assignment of a generic observation $\mathbf{x}_j$ $(j = 1, \ldots, n)$ to a cluster can be based on the posterior probabilities:

$$\begin{aligned} p_i(\mathbf{x}_j; \boldsymbol{\theta}) &= \frac{p_i f_D(\mathbf{x}_j; \boldsymbol{\alpha}_i)}{\sum_{h=1}^D p_h f_D(\mathbf{x}_j; \boldsymbol{\alpha}_h)} \\ &= p_i \frac{\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau)} x_{ji}^\tau}{\sum_{h=1}^D p_h \frac{\Gamma(\alpha_h)}{\Gamma(\alpha_h+\tau)} x_{jh}^\tau}, \quad i = 1, \ldots, D \end{aligned} \qquad (9)$$

representing the probability that observation $\mathbf{x}_j$ belongs to component i for a given value of $\boldsymbol{\theta}$. Consequently, $\mathbf{x}_j$ is assigned to cluster i $(j = 1, \ldots, n; i = 1, \ldots, D)$ if

$$\max_{h=1,\ldots,D} p_h(\mathbf{x}_j; \boldsymbol{\theta}) = p_i(\mathbf{x}_j; \boldsymbol{\theta})$$

which can be equivalently written as

$$\frac{x_{ji}}{x_{jh}} \geq \left[\frac{p_h q_i}{p_i q_h}\right]^{1/\tau}, \quad \forall h = 1, \ldots, D, h \neq i \qquad (10)$$

where $q_i = \Gamma(\alpha_i + \tau)/\Gamma(\alpha_i)$.

It is noteworthy that the above criterion has a simple graphical representation. Indeed, the coordinates

$$\frac{(q_i/p_i)^{1/\tau}}{\sum_{h=1}^D (q_h/p_h)^{1/\tau}} \quad i = 1, \ldots, D \qquad (11)$$

identify a pseudo-barycenter. By connecting it to the simplex edges along the line joining the pseudo-barycenter and the

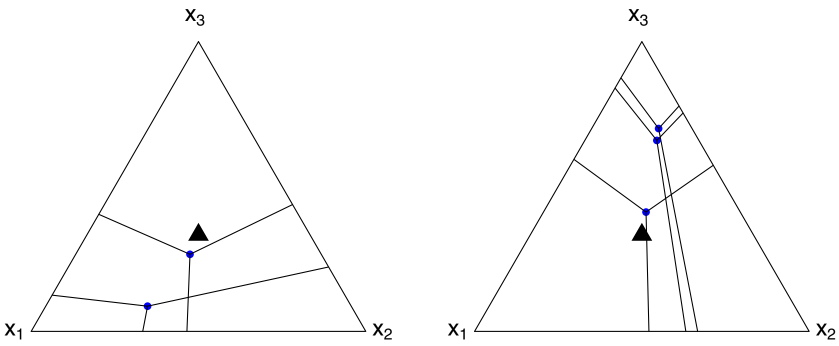

Fig. 2 Cluster regions. *Left panel* $\boldsymbol{\alpha} = (30, 30, 30)$, $\mathbf{p} = (0.1, 0.2, 0.7)$, $\tau = 1$ and $\tau = 5$. *Right panel* $\boldsymbol{\alpha} = (10, 20, 70)$, $\mathbf{p} = (1/3, 1/3, 1/3)$, $\tau = 1$, $\tau = 10$ and $\tau = 400$. *Filled triangle* indicates the point $(1/3, 1/3, 1/3)$

vertexes, D regions, corresponding to the clusters, are identified. The following Fig. 2 shows such regions in the case $D = 3$ for different parameter values.

Note that, when τ gets larger, the impact of differences in the p_i's or in the α_i's decreases, making the cluster regions more symmetric.

4 Identifiability and likelihood maximum existence of the FD

The finite mixture structure (3) allows the FD distribution to display high flexibility in terms of shape, enabling it to capture many specific data features. At the same time, it is well known that general mixture models entail inferential difficulties, for example, in the context of estimation. In particular, unlike "standard models", mixture models typically display non identifiability and often unboundedness of likelihood and non existence of a (finite) likelihood maximum. In this section we shall show that the FD does not share such difficulties, thanks to its special structure.

Consider first the identifiability issue, which has strong implications for estimation problems. For a generic parametric family of distributions, the classical notion of identifiability requires that two elements of the family are equal if and only if the corresponding parameters are identical. General mixture models, including mixtures with arbitrary Dirichlet components, do not meet this definition. A possible solution is to introduce appropriate constraints on the parameter space (see for example Frühwirth-Schnatter 2006). Otherwise, one may resort to weaker identifiability notions (see Frühwirth-Schnatter 2006), among which local identifiability, i.e. for any point $\boldsymbol{\theta}$ of the parameter space, there is a neighborhood containing no other point corresponding to the same distribution implied by $\boldsymbol{\theta}$ (see for example Rothenberg 1971).

A typical source of non-identifiability is due to invariance under permutations of the component labels of the mixture. Unlike general mixture models, in the FD the components are intrinsically distinct, being characterized by different parameter vectors $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha} + \tau \mathbf{e}_i)$, implying different component means as discussed in Sect. 3.2. This makes it possible to assign them a unique label, without having to introduce ad hoc constraints on the parameter space.

Another general kind of non-identifiability derives from potential overfitting due to allowing a variable number of components in the model. This option is very important in practice, because the number of components is often unknown. However, in general mixture models, it induces non-identifiable parameter subsets corresponding to some null mixing weights.

Formally, this kind of non-identifiability is usually solved by assuming positivity constraints on the weights. On the contrary, our model more properly allows for some null weights by entailing the presence of up to D components. Nevertheless, thanks again to its special mixture structure, this can be achieved without losing identifiability in the classic strong sense, as proved below.

Proposition 2 (Identifiability of the FD) *Let* $\mathbf{X} \sim \mathcal{FD}(\boldsymbol{\theta})$ *and* $\mathbf{X}' \sim \mathcal{FD}(\boldsymbol{\theta}')$, *where* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{p}, \tau) \in \boldsymbol{\Theta}$ *and* $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \mathbf{p}', \tau') \in \boldsymbol{\Theta}$. *Then* $\mathbf{X} \sim \mathbf{X}'$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

Proof We need to show that if $\mathbf{X} \sim \mathbf{X}'$ then $\boldsymbol{\theta} = \boldsymbol{\theta}'$, the converse being obvious. Clearly, if $\mathbf{X} \sim \mathbf{X}'$ then $X_i \sim X_i'$, $(i = 1, \ldots, D)$. By closure under marginalization of the FD model (see Ongaro and Migliorati 2013), X_i has density

$$g_i(x; \boldsymbol{\theta}) = x^{\alpha_i - 1}(1-x)^{\alpha^+ - \alpha_i - 1} \cdot \\ \cdot \left(p_i c_i(\boldsymbol{\alpha}, \tau) x^{\tau} + (1 - p_i) d_i(\boldsymbol{\alpha}, \tau)(1-x)^{\tau}\right)$$

for $x \in (0, 1)$, where

$$c_i(\boldsymbol{\alpha}, \tau) = \frac{\Gamma(\alpha^+ + \tau)}{\Gamma(\alpha^+ - \alpha_i)\Gamma(\alpha_i + \tau)},$$
$$d_i(\boldsymbol{\alpha}, \tau) = \frac{\Gamma(\alpha^+ + \tau)}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau)}.$$

It follows that $g_i(x; \boldsymbol{\theta}) = g_i(x; \boldsymbol{\theta}')$ a.s., $(i = 1, \ldots, D)$. As the two densities are continuous on $(0, 1)$, equality must hold identically for any $x \in (0, 1)$.

For any $i = 1, \ldots, D$, we have that $g_i(x; \boldsymbol{\theta}) x^{1-\alpha_i}$ tends to $(1 - p_i) d_i(\boldsymbol{\alpha}, \tau)$ when $x \to 0^+$.

Therefore $g_i(x;\boldsymbol{\theta}')x^{1-\alpha_i}$ must tend to the same quantity for $x \to 0^+$. As $p_i < 1$, it is easy to check that this can happen only if $\alpha_i = \alpha_i'$. Furthermore, it must be that $(1 - p_i)d_i(\boldsymbol{\alpha}, \tau) = (1 - p_i')d_i(\boldsymbol{\alpha}', \tau')$.

If we introduce these constraints, holding for $i = 1, \ldots, D$, in the equality $g_i(x;\boldsymbol{\theta}) = g_i(x;\boldsymbol{\theta}')$, for any $x \in (0, 1)$ and $i = 1, \ldots, D$ we obtain:

$$p_i c_i(\boldsymbol{\alpha}, \tau)x^{\tau} + (1 - p_i)d_i(\boldsymbol{\alpha}, \tau)(1 - x)^{\tau}$$
$$= p_i' c_i(\boldsymbol{\alpha}, \tau')x^{\tau'} + (1 - p_i)d_i(\boldsymbol{\alpha}, \tau)(1 - x)^{\tau'}$$

By taking the limit as $x \to 1^-$ on both sides of the above equation we have $p_i c_i(\boldsymbol{\alpha}, \tau) = p_i' c_i(\boldsymbol{\alpha}, \tau')$, $i = 1, \ldots, D$. Then, by computing both sides in $x = 0.5$ the implication $\tau = \tau'$ is obtained, which directly leads to equality of p_i and p_i' for all i. □

Let us now focus on likelihood features. Very often mixture model likelihoods are unbounded, which generates theoretical as well as computational difficulties. For example, in the normal mixture model the problem arises when the variances of the mixture components are allowed to vary freely (see Hathaway 1985). The same phenomenon appears in general Dirichlet mixture models. In these cases unboundedness arises because a component of the mixture is entirely dedicated to one observation. Such unbounded likelihood points are typically of no inferential interest, being caused by an overfitting problem, and they require care in the maximization procedure.

By contrast, under very weak conditions, not only is the FD likelihood bounded, but it also admits a finite global maximum on the assumed parameter space, as proved in the following proposition.

Proposition 3 (Boundedness and maximum of FD likelihood) *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an i.i.d. sample from a $\mathcal{FD}(\mathbf{x};\boldsymbol{\theta})$ and suppose that $n \geq D + 1$. Then, a.s. the log-likelihood $l(\boldsymbol{\theta})$ is bounded from above and it admits a finite global maximum, i.e. there exists $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ such that $l(\hat{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta})$ $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

The proof is reported in the Appendix 3 and is based on divergence properties of the Dirichlet log-likelihood derived in Lemma 1 of Appendix 2.

Statistically meaningful MLE can be defined even in the cases of unbounded likelihood or non-identifiable parametrization (see Redner 1981; Hathaway 1985; Peters and Walker 1978; Feng and McCulloch 1996). However, their treatment is more involved from both a theoretical and an applied perspective. For example, one can define the MLE as any of the local maximizers of the likelihood function if the latter is unbounded. It is then possible to prove that, under regularity conditions, there exists a (unique) sequence of likelihood equation roots which is consistent (for more details see Peters and Walker 1978). But then, a major difficulty is in determining the correct one (see Hathaway 1985; Lehmann and Casella 1998; McLachlan and Peel 2000).

Our results ensure that, for the FD, the search for a MLE leads to the solution of a well posed optimization problem having a global solution, thus avoiding the above mentioned difficulties. In this case, very general conditions for the global MLE to be consistent and asymptotically efficient are well known (see Wald 1949; Kiefer and Wolfowitz 1956; Hathaway 1985; Lehmann and Casella 1998).

5 Maximum likelihood estimation

In this section we shall consider some suitable EM-type algorithms and propose estimation strategies based on (combinations of) them. The selection of the most reliable and efficient estimation procedure(s) will be performed via simulation (Sect. 7) and real data analysis (Sect. 8).

5.1 EM-type algorithms

Standard methods for likelihood maximization fail to give a solution in the present setup. However, the finite mixture structure of the model makes it possible to treat the estimation issue as an incomplete data problem. Thus the EM algorithm (see Dempster et al. 1977; McLachlan and Peel 2000) can be suitably adapted. In particular, given n independent observations $\mathbf{x}_j$, $j = 1, \ldots, n$, from (1), the complete-data vector $\mathbf{x}_c$ is given by:

$$\mathbf{x}_c = (\mathbf{x}, \mathbf{z}) = (\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n)$$

where the component-label D-dimensional vectors $\mathbf{z}_j = (z_{j1}, \ldots, z_{jD})$ $(j = 1, \ldots, n)$ represent the missing data, z_{ji} being equal to 1 if the j-th observation has arisen from the i-th component of the mixture model and 0 otherwise $(j = 1, \ldots, n;\ i = 1, \ldots, D)$.

The true log-likelihood can be thought of as originating from the following complete-data log-likelihood:

$$\log L_c(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{D} z_{ji}\left\{\log p_i + \log f_D(\mathbf{x}_j;\boldsymbol{\alpha_i})\right\} \qquad (12)$$

where $f_D(\mathbf{x}_j;\boldsymbol{\alpha_i})$ is the density of a Dirichlet (5) with parameter $\boldsymbol{\alpha_i} = (\boldsymbol{\alpha} + \tau\mathbf{e}_i)$.

The $k + 1$ step of the EM algorithm can be described as follows.

E-step: given the current parameter estimates $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\alpha}^{(k)}, \mathbf{p}^{(k)}, \tau^{(k)})$, calculate the conditional expectation of the complete-data log-likelihood (12) given $\mathbf{x} = (\mathbf{x}_1, \ \ldots, \mathbf{x}_n)$ as:

$$\sum_{i=1}^{D}\sum_{j=1}^{n} p_i\left(\mathbf{x}_j;\boldsymbol{\theta}^{(k)}\right) \times \left\{\log p_i^{(k)} + \log f_D\left(\mathbf{x}_j;\boldsymbol{\alpha}^{(k)}+\tau^{(k)}\mathbf{e}_i\right)\right\} \quad (13)$$

where $p_i\left(\mathbf{x}_j;\boldsymbol{\theta}^{(k)}\right)$ represents the "posterior" probability (9) that $\mathbf{x}_j$ belongs to the i-th component of the mixture given $\boldsymbol{\theta}^{(k)}$.

M-step: maximize (13) to obtain the maximum likelihood estimates of the parameters. In particular, we have $\hat{p}_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} p_i\left(\mathbf{x}_j;\boldsymbol{\theta}^{(k)}\right)$, $(i = 1,\ldots,D-1)$, whereas $\hat{\boldsymbol{\alpha}}^{(k+1)}$ and $\hat{\tau}^{(k+1)}$ can be computed by implementing a Newton-Raphson method.

Iterations cycle between the two steps until a "sufficiently small" change in the observed log-likelihood (or in the parameter estimates) is reached.

Although the EM algorithm produces an increase in the likelihood at each step, our simulation study shows that for some parametric configurations it leads to solutions heavily dependent on the initial values, thus resulting in poor approximations of the real maximum.

In order to weaken this dependence, we also consider some adaptations of the EM, namely the Classification EM (CEM) and the Stochastic EM (SEM) (Celeux and Govaert 1992; Banfield and Raftery 1993; Celeux et al. 1996; Biernacki et al. 2003). Both algorithms have an EM structure enriched by incorporating a further step between the E and M steps.

On the $(k+1)$ iteration, CEM, after the E step, replaces the component-labels z_{ji} by one or zero according to whether or not

$$p_i^{(k)} f_D(\mathbf{x}_j;\boldsymbol{\alpha}_i^{(k)}) \geq p_h^{(k)} f_D(\mathbf{x}_j;\boldsymbol{\alpha}_h^{(k)}) \quad h = 1,\ldots,D;\, h \neq i$$

holds $(j = 1,\ldots,n)$. Therefore, it creates a partition of data by assigning each observation to the cluster maximizing the current estimate of the conditional probability $p_i(\mathbf{x}_j;\boldsymbol{\theta})$ given by (9). This rule for the FD model takes the form (10). The M step of CEM then consists in setting the updated mixing proportions equal to the relative number of observations of each cluster. The other parameters are updated by maximizing the so-called classified likelihood, i.e. the likelihood computed by assuming knowledge of which mixture component each observation comes from (see Sect. 5.2). CEM stops when the partition remains unchanged in two subsequent cycles, so that the algorithm converges in a finite number of iterations.

With SEM, after the E step, a partition is designed by conducting a single draw from the current conditional distribution of $\mathbf{z}_j$ $(j = 1,\ldots,n)$ given the observed data, i.e. by drawing from the multinomial distribution with parameter equal to the current estimates of the conditional probabilities (9). Hence, SEM generates a partition as well as CEM, but based on a random rule instead of a deterministic one. The M-step is then analogous to the CEM one. The final estimate provided by SEM is represented either by the mean of the sequence of estimates (after a burn-in period), or by the point of the sequence leading to the highest (true) likelihood.

CEM converges very rapidly, but it yields inconsistent estimates of the parameters and it performs poorly for clusters not widely separated or in disparate proportions (McLachlan and Peel 2000).

SEM typically requires more iterations. It generates a Markov chain which is expected to visit the whole parameter space, it thus gives the iterative process a chance to escape from a current path of convergence to a local maximizer.

We decide to test and compare via simulation EM, CEM and SEM separately, as well as two combinations of them obtained by refining CEM and SEM with EM. The idea behind the latter two combinations is to first use CEM (or SEM) to better span the parametric space, and then apply EM, which is very precise in finding local maxima close to the starting point.

5.2 Maximization of classified likelihood

As we have seen in Sect. 5.1 and we shall see in Sect. 7.1, the implementation of CEM and SEM algorithms (M step) as well as the determination of initial parameter values require maximization of the so-called classified likelihood. Suppose the sample has been partitioned into D groups $\{\mathcal{G}_1,\ldots,\mathcal{G}_D\}$. Then the classified likelihood is given by

$$\prod_{j\in A_i} f_D(\mathbf{x}_j;\boldsymbol{\alpha}_i) \quad (14)$$

$(i = 1,\ldots,D)$, where f_D is the Dirichlet density given by (5) and $A_i = \{j : \mathbf{x}_j \in \mathcal{G}_i\}$. Maximization of (14) can not be achieved by maximizing separately the Dirichlet likelihood relative to each group due to constraints on the cluster parameters $\boldsymbol{\alpha}_i$ $(i = 1,\ldots,D)$. Though, a direct maximization of (14) is (numerically) feasible because the corresponding likelihood equations take a simple form.

Note that an initial value for the parameters to be optimized over is needed. To this end, the method of moments can be fruitfully adapted. By recalling that if $\mathbf{X} \sim \mathcal{D}(\boldsymbol{\alpha})$, then $\mathrm{E}(X_i) = \alpha_i/\alpha^+$, one has that the ratios $\alpha_i/(\alpha^+ + \tau)$ can be initialized as the weighted mean of all sample group means $\bar{x}_{hi}$ $(\forall h \neq i)$:

$$\frac{\sum_{h=1,h\neq i}^{D} \bar{x}_{hi}\,\hat{p}_h}{1-\hat{p}_i}$$

and that the ratio $\tau/(\alpha^+ + \tau)$ can be initialized as the mean of the D estimates:

$$\bar{x}_{ii} - \frac{\sum_{h=1,h\neq i}^{D} \bar{x}_{hi}\, \hat{p}_h}{1-\hat{p}_i},\ i=1,\ldots,D.$$

Finally, by exploiting the Dirichlet variances, the total $(\alpha^+ + \tau)$ can be estimated as the weighted mean of the D estimates

$$\frac{1-\sum_{i=1}^{D}\bar{x}_{hi}^2}{\sum_{i=1}^{D} s_{hi}^2} - 1,\ h=1,\ldots,D$$

computed on each h-th group, where s_{hi}^2 are the sample group variances. These estimates are derived by summing these variances within each group so as to get stable estimates whenever the sample group variances are close to zero.

6 Estimation of MLE variance-covariance matrix

The asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be approximated by the inverse of the observed information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}}; \boldsymbol{x})$. However, the direct evaluation of the second-order derivatives of the (incomplete) log-likelihood derived from (1) is excessively cumbersome. Some methods for approximating the observed information matrix, such as the empirical information-based method (Meilijson 1989) or the supplemented EM algorithm proposed by (Meng and Rubin 1991), are feasible but will often underestimate the standard errors (Efron 1994).

A further approximation of the variance-covariance matrix of the MLE can be achieved via parametric bootstrap. Once the MLE $\hat{\boldsymbol{\theta}}$ is obtained, B' bootstrap samples are drawn from $\mathcal{FD}(\hat{\boldsymbol{\theta}})$ and the proposed estimation algorithm is applied to each sample. Some simulations have been performed showing that the parametric bootstrap produces satisfactory results. However, its implementation is extremely demanding in terms of computational burden because each replication requires an estimate computation.

An interesting indirect but exact evaluation of $\mathbf{I}(\hat{\boldsymbol{\theta}}; \boldsymbol{x})$ Louis (1982) can be obtained via decomposition of complete data into observed and missing ones (see Sect. 5.1), so that the following equality holds:

$$\mathbf{I}\left(\hat{\boldsymbol{\theta}};\mathbf{x}\right) = \left[\mathrm{E}_{\boldsymbol{\theta}}\left\{\mathbf{I}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)|\mathbf{x}\right\}\right]_{\theta=\hat{\theta}} - \left[\mathrm{E}_{\boldsymbol{\theta}}\left\{\mathbf{S}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)\mathbf{S}_c^T\left(\boldsymbol{\theta};\mathbf{X}_c\right)|\mathbf{x}\right\}\right]_{\theta=\hat{\theta}} \quad (15)$$

Here $\mathbf{S}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ denotes the complete-data score statistics, $\mathbf{I}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ is the negative of the Hessian of the complete-data log-likelihood (12) and $\mathrm{E}_{\boldsymbol{\theta}}[\cdot|\boldsymbol{x}]$ denotes the conditional expectation, given the observed data $\mathbf{x}$, using the parameter vector $\boldsymbol{\theta}$.

The elements of the score statistic $\mathbf{S}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ and of the $2D \times 2D$ matrix $\mathbf{I}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ can be easily derived from (12) and are reported in the Appendix 4.

The calculation of the conditional expectations required by (15) is feasible but very tedious. Here we shall propose and study a simpler but very accurate evaluation of the conditional expectation in (15) based on conditional bootstrap (Diebolt and Ip 1996). This is obtained by observing that, conditionally on $\mathbf{x}$, the component-label random vectors $\mathbf{Z}_j$ $(j=1,\ldots,n)$ are independently distributed as multinomials with parameter $\mathbf{p}_j^* = (p_{j1}^*,\ldots,p_{jD}^*)$ where $p_{ji}^* = p_i(\mathbf{x}_j;\boldsymbol{\theta})$ $(i=1,\ldots,D)$ is given by (9) with $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$. Therefore, the conditional expectations (15) can be approximated by averaging over B independent bootstrap samples $\mathbf{z}_{jb}, (j=1,\ldots,n; b=1,\ldots,B)$ from $\mathbf{Z}_j$, for a sufficiently high value of B.

7 Initialization of EM-type algorithms and simulation studies

This section will develop ad hoc initialization strategies of EM-type algorithms. Moreover, it includes extensive simulations aimed at: (1) comparing these initialization strategies as well as the different estimation procedures described in Sect. 5; (2) evaluating the MLE performances and their standard errors derived in Sect. 6.

7.1 Choice of the starting values for EM-type algorithms

It is common knowledge that optimizing initial values for the EM algorithm is crucial because they can heavily influence the speed of convergence of the algorithm, as well as, even more importantly, its ability to locate the global maximum. Many different strategies, often involving multiple random starts, have been developed in the literature to face this critical issue, especially within the normal mixture framework (Biernacki et al. 2003; O'Hagan et al. 2012). Typically, a first (random) partition of the data into groups is obtained.

In order to get starting values of higher quality, we shall propose two new initial non random partitions of the data based on clustering algorithms which are specifically devised for the problem at hand. This problem exhibits at least two peculiarities. Firstly, the observations belong to the simplex, so that clustering methods based on standard metrics on $\mathcal{R}^{\mathcal{D}}$ as well as on sum of square decompositions may not be suitable. Secondly, as discussed in Sect. 3, the parameter structure of the model involves specific cluster patterns.

More precisely, the first strategy is expressly based on the FD cluster structure. It constructs an initial partition which assigns observation $\mathbf{x}_j$ to group i if

$$\frac{x_{ji}}{x_{jh}} > \frac{B_i}{B_h}, \quad \forall h = 1, \ldots, D, h \neq i$$

where $\mathbf{B} = (B_1, \ldots, B_D)$ is a barycenter based on data, e.g. the mean or the median. This rule, hereafter called barycenter method, is coherent with the classification rule (10) prescribed by the FD model, provided that the pseudo-barycenter (11) is well approximated by a sample barycenter. It can be proved that this is the case if the α_i's are large (which is often true in applications) and the mixing weights are similar. In fact, for large α_i $\Gamma(\alpha_i + \tau)/\Gamma(\alpha_i) \approx \alpha_i^\tau$ so that $(q_i/q_h)^{1/\tau} \approx \alpha_i/\alpha_h \approx \frac{\alpha_i+p_i\tau}{\alpha_h+p_h\tau}$, the latter being the ratio between the i-th and h-th variable means of the composition.

Moreover, to take into account the compositional nature of data, initial partitions of a second type are obtained by considering the k-means algorithm applied to the transformations most commonly used for simplex data, i.e.:

1. The standard representation, i.e. select $D - 1$ of the D elements of the composition;
2. Additive logratio (alr), i.e. divide the elements by one arbitrarily chosen among them and take logarithms;
3. Centered logratio (clr), i.e. divide the elements by their geometric mean and take logarithms (Aitchison 2003);
4. Isometric logratio (ilr), i.e. endow the simplex with an inner product compatible with perturbation and powering (Aitchison 2003) and then express the elements with respect to an orthonormal basis in the simplex (Pawlowsky-Glahn et al. 2015);
5. Regular simplicial transformation as described in Sect. 3.1.

Note that the application of transformations 1., 2. and 3. requires the choice of a left out variable. In this case the results may heavily depend on this choice because such transformations are not symmetric, i.e. invariant under permutation of the variables. More precisely, the clr is a symmetric transformation, but it maps the $D - 1$-dimensional simplex to a linear $D - 1$-dimensional subspace of $\mathcal{R}^D$. To obtain a map to a full $D - 1$-dimensional space $\mathcal{R}^{D-1}$ so as to apply a clustering algorithm, one has to drop one of the D elements of such a map, thus making the transformation not symmetric. On the other hand, transformations 4. and 5., being symmetric, do not share this difficulty. Even more importantly, by treating all variables on an equal footing, they are more coherent with the FD model, which is invariant under permutation. For an application of the ilr transformation to a non-parametric clustering problem see Palarea-Albaladejo et al. (2012).

Moreover, in the FD all group means have the same distance between each other (see Sect. 3.2) in the symmetric D-dimensional space as well as in regular simplicial transformations. By contrast, non-symmetric transformations may produce a distortion in the construction of the clusters because they alter distances among points according to which variable is left out.

Note also that transformations based on the logarithm of suitable ratios of the original variables appear to be less compatible with the FD model than linear transformations. Indeed, the FD entails convex clusters (generated by Dirichlet density level curves) and partitions formed by straight segments (as shown in Sect. 3.4). These are perfectly compatible with the output produced by standard clustering algorithms when applied on linear transformations of the variables. On the contrary, logratio-type transformations typically map convex sets (as the ones generated by the FD) into "banana"- or "boomerang"-shaped configurations, which are more difficult to be captured by ordinary clustering algorithms.

The above remarks induce us to prefer symmetric linear transformations, i.e. regular simplicial ones. It is important to underline that this choice has been confirmed by the simulation studies and the real data analysis presented in Sects. 7.2 and 8. As a matter of fact, though not reported for space constraints, we also implemented the other transformations, never reaching higher maximum likelihood values.

Once an initial partition is obtained, group labeling needs to be established. Indeed, any clustering algorithm assigns the group labels randomly. On the other hand, the FD cluster structure entails a precise labeling scheme (see Sect. 3.2). A FD coherent labeling can be constructed by assigning label i to group with the largest sample mean of element i. If a single group displays two or more elements with maximum sample means, we consider all label permutations which are compatible with the largest sample mean positions and choose the one that maximizes the likelihood.

Finally, given a partition with properly labeled clusters, initial parameter estimates must be determined. The mixing proportions p_i's are obtained as the relative numbers of observations pertaining to each cluster. The other parameters are found by maximizing the corresponding classified likelihood, as shown in Sect. 5.2.

7.2 Assessment of estimation strategies via simulation

In order to compare the performances of the different EM-type algorithms and initializations illustrated in Sects. 5.1 and 7.1, we simulated samples of size 100 from 20 different FD parameter configurations with $D = 3$. More precisely, the following values were chosen: for the p_i's 0.01, 0.1, 0.3, 1/3, 0.495, 0.6, 0.98, for the α_i's 5, 10, 50, 100, 500, 1000 and for τ 2, 5, 10, 20, 40. The selected configurations make it possible to cover a great variety of cases. Well-separated as well as poorly separated clusters (Kullback–Leibler distance (7) between clusters ranging from 0.76 to 125), leading to unimodal and multimodal distributions, were included. Furthermore, sets of p_i's and α_i's values from very similar to extremely different were chosen, thus implying different variability and the presence of 1 up to 3 clusters.

For each sample we considered all combinations of the initialization and maximization procedures selected in Sects. 5.1 and 7.1. In particular, to compute the starting values we chose the barycenter method based on sample mean and median, and the k-means clustering applied to the regular simplicial transformation. For the latter, we considered both $k = 3$ and $k = 2$ groups.

We adopted a very strict convergence criterion for the EM algorithm, stopping when the change both in the log-likelihood and in the parameter estimates is lower than 10^{-3}.

For brevity we only report the percentages of cases in which each combination reaches the highest likelihood value (see Table 1).

The most evident conclusion is the clear superiority of the SEM+EM method over the other maximization ones, irrespective of the initialization choice. Indeed, for any given initialization method it nearly always produces the highest likelihood. Moreover, even if SEM is more time consuming than CEM, it converges quite rapidly (in the majority of cases less than 30 iterations, with a maximum of 50), and the EM, when used after the SEM, has a very fast convergence as well. The CEM+EM has a slightly better performance than the EM algorithm. By adding the EM to the CEM and SEM, a substantial improvement of the percentages in Table 1 is obtained. However, the increase in the likelihood value due to the EM is almost always negligible when applied to the SEM, while it is often quite significant in the CEM case.

The three initialization strategies 1., 3. and 4. have comparable and superior global performances. Strategy 2. is needed though, because it is the only one producing the best results in some of the two cluster cases. This phenomenon is even more clean-cut in the real data set analysis (Sect. 8). Quite remarkably, a careful investigation shows that the first three initialization strategies followed by SEM+EM are enough to reach the highest likelihood value, for each parameter configuration and real data set example, among combinations of all considered initialization and maximization procedures.

7.3 MLE and standard error simulation

In this section we present the results of a simulation study concerning the MLE of the FD model parameters and the complete-data likelihood-based evaluation of their standard errors described in Sect. 6. The estimators were derived by using the first three initializations displayed in Table 1 followed by the SEM +EM combination as discussed in Sect. 7.2.

Table 1 Simulation results with 20 parameter configurations

Initialization	Algorithm				
	CEM	SEM	EM	CEM + EM	SEM + EM
1. RS	15	45	55	60	85
2. RS_2	5	20	15	20	30
3. Mean	15	60	55	65	90
4. Median	15	60	55	70	85

Percentages of cases for which the combination transformation (row)/EM-type algorithm (column) reaches the highest likelihood value. Legend for initializations: RS 3-means with Regular Simplicial transformation, RS_2 2-means with Regular Simplicial transformation, $Mean$ barycenter method based on mean, $Median$ barycenter method based on median

Several parameter configurations and sample sizes were investigated. For space constraints, we report only four configurations (Tables 2, 3, 4 and 5, respectively) with two sample sizes ($n = 100; 500$). The chosen configurations correspond to four representative cases. In the first two we selected three moderately separated clusters (overlapping measures given by (8) equal to $k(\alpha_i, \tau) = 5.244;\ i = 1, 2, 3$) with equal α values and equal cluster weights in the first case, different weights in the second. The third case corresponded to a bimodal density (only one cluster separated from the others since $k(\alpha_1, \tau) = 7.19$, $k(\alpha_2, \tau) = 1.84$, $k(\alpha_3, \tau) = 1.18$) with different α_i's. Finally, we considered an extremely awkward setting with three strongly overlapped clusters and unequal weights producing a unimodal density ($k(\alpha_i, \tau) = 1.48;\ i = 1, 2, 3$). For each parameter configuration, we simulated $K = 1000$ replications of size n random samples, and we computed the parameter estimates, the complete-data likelihood-based standard error estimates derived from $B = 3,000$ independent bootstrap samples and the confidence intervals based on the (asymptotic) normal distribution. In each table below, rows "MLE mean" and "MLE sd" show the simulated mean and standard deviation of the MLE of $\boldsymbol{\theta}$. Row "se mean" reports the simulated mean of the standard error estimators and row "arb" reports its absolute relative bias, i.e. the mean of the absolute deviations between such estimates of the standard errors and the simulated standard deviation (row 2) divided by this last quantity. Finally, row "coverage" gives the simulated confidence levels against a 95 % nominal one.

The performance of the MLE in terms of mean and standard deviation appears unusually good for a mixture model. Indeed, in the first three cases (Tables 2, 3 and 4) even for $n = 100$ the bias is always less than 5 % and the standard deviations are reasonably low. In the last, quite critical case (Table 5), the performance is still acceptable with only an increase in bias and standard deviations especially of the p_i's

Table 2 Simulation results with $\mathbf{X} \sim \mathcal{FD}(\boldsymbol{\alpha} = (15, 15, 15), \mathbf{p} = (1, 1, 1)/3, \tau = 10)$

	p_1	p_2	α_1	α_2	α_3	τ
$n = 100$						
MLE mean	0.334	0.334	15.537	15.557	15.554	10.427
MLE sd	0.059	0.059	1.921	1.945	1.955	1.523
se mean	0.057	0.057	1.886	1.891	1.889	1.522
arb	0.078	0.079	0.078	0.080	0.082	0.052
Coverage	0.940	0.942	0.952	0.936	0.936	0.951
$n = 500$						
MLE mean	0.333	0.334	15.101	15.099	15.108	10.090
MLE sd	0.025	0.026	0.817	0.805	0.816	0.657
se mean	0.025	0.025	0.824	0.824	0.825	0.668
arb	0.034	0.035	0.035	0.038	0.035	0.026
Coverage	0.952	0.946	0.951	0.959	0.959	0.961

MLE mean MLE (simulated) mean; *MLE sd* MLE (simulated) standard deviation; *se mean* standard error mean; *arb* absolute relative bias; *coverage* confidence interval coverage

Table 3 Simulation results with $\mathbf{X} \sim \mathcal{FD}(\boldsymbol{\alpha} = (15, 15, 15), \mathbf{p} = (0.1, 0.3, 0.6), \tau = 10)$

	p_1	p_2	α_1	α_2	α_3	τ
$n = 100$						
MLE mean	0.101	0.302	15.439	15.400	15.418	10.352
MLE sd	0.041	0.058	1.834	1.860	2.015	1.478
se mean	0.038	0.055	1.821	1.848	1.968	1.506
arb	0.155	0.090	0.080	0.083	0.093	0.058
Coverage	0.919	0.932	0.955	0.953	0.955	0.951
$n = 500$						
MLE mean	0.101	0.302	15.104	15.106	15.105	10.101
MLE sd	0.017	0.025	0.772	0.807	0.860	0.661
se mean	0.017	0.025	0.799	0.811	0.862	0.661
arb	0.060	0.036	0.046	0.037	0.039	0.025
Coverage	0.955	0.947	0.963	0.959	0.956	0.951

MLE mean MLE (simulated) mean; *MLE sd* MLE (simulated) standard deviation; *se mean* standard error mean; *arb* absolute relative bias; *coverage* confidence interval coverage

estimators. This correctly reflects the expected uncertainty in detecting the clusters. These positive results, together with the fast convergence rates, corroborate the reliability and efficiency of the proposed estimation strategy.

The standard error estimates are computationally efficient and generally very accurate: the bias is always less than 4% in the first three cases; somewhat significant deviations appear only in the last case. As for the arb, it is generally lower than 10% except in the last case only for some p_i's when $n = 100$.

Very reliable results are obtained in terms of confidence interval coverages except for the last case with $n = 100$.

Table 4 Simulation results with $\mathbf{X} \sim \mathcal{FD}(\boldsymbol{\alpha} = (10, 50, 80), \mathbf{p} = (1, 1, 1)/3, \tau = 10)$

	p_1	p_2	α_1	α_2	α_3	τ
$n = 100$						
MLE mean	0.337	0.332	10.425	52.190	83.615	10.549
MLE sd	0.062	0.100	1.446	7.390	11.884	1.996
se mean	0.064	0.096	1.391	7.153	11.467	1.947
arb	0.131	0.231	0.097	0.090	0.088	0.069
Coverage	0.949	0.929	0.933	0.941	0.942	0.942
$n = 500$						
MLE mean	0.335	0.334	10.077	50.349	80.661	10.096
MLE sd	0.029	0.044	0.588	2.981	4.854	0.860
se mean	0.028	0.042	0.599	3.079	4.964	0.846
arb	0.069	0.095	0.042	0.047	0.040	0.034
Coverage	0.938	0.942	0.952	0.948	0.946	0.933

MLE mean MLE (simulated) mean; *MLE sd* MLE (simulated) standard deviation; *se mean* standard error mean; *arb* absolute relative bias; *coverage* confidence interval coverage

Table 5 Simulation results with $\mathbf{X} \sim \mathcal{FD}(\boldsymbol{\alpha} = (15, 15, 15), \mathbf{p} = (0.1, 0.3, 0.6), \tau = 5)$

	p_1	p_2	α_1	α_2	α_3	τ
$n = 100$						
MLE mean	0.118	0.317	15.960	15.900	15.999	5.913
MLE sd	0.123	0.162	2.535	2.636	3.093	1.531
se mean	0.109	0.144	2.557	2.712	3.092	1.808
arb	0.498	0.486	0.222	0.236	0.268	0.253
Coverage	0.904	0.842	0.920	0.929	0.905	0.921
$n = 500$						
MLE mean	0.098	0.308	15.151	15.113	15.171	5.122
MLE sd	0.050	0.076	1.144	1.215	1.416	0.783
se mean	0.055	0.082	1.252	1.310	1.513	0.887
arb	0.395	0.384	0.160	0.181	0.172	0.192
Coverage	0.922	0.942	0.946	0.949	0.948	0.957

MLE mean MLE (simulated) mean; *MLE sd* MLE (simulated) standard deviation; *se mean* standard error mean; *arb* absolute relative bias; *coverage* confidence interval coverage

8 An application to a real data set

The applicative potential of the FD can be better illustrated by a real data set analysis. To this end, we shall focus on a large data set consisting of 572 samples of Italian olive oil produced in nine different areas, which can be naturally aggregated into three geographical macro-areas: southern Italy, Sardinia and northern Italy. On each sample the composition of eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction is reported. Olive oils mostly consist of oleic acid (more than 72 % on average), the palmitic acid accounts

Fig. 3 Histograms and estimated densities of 2-part compositions

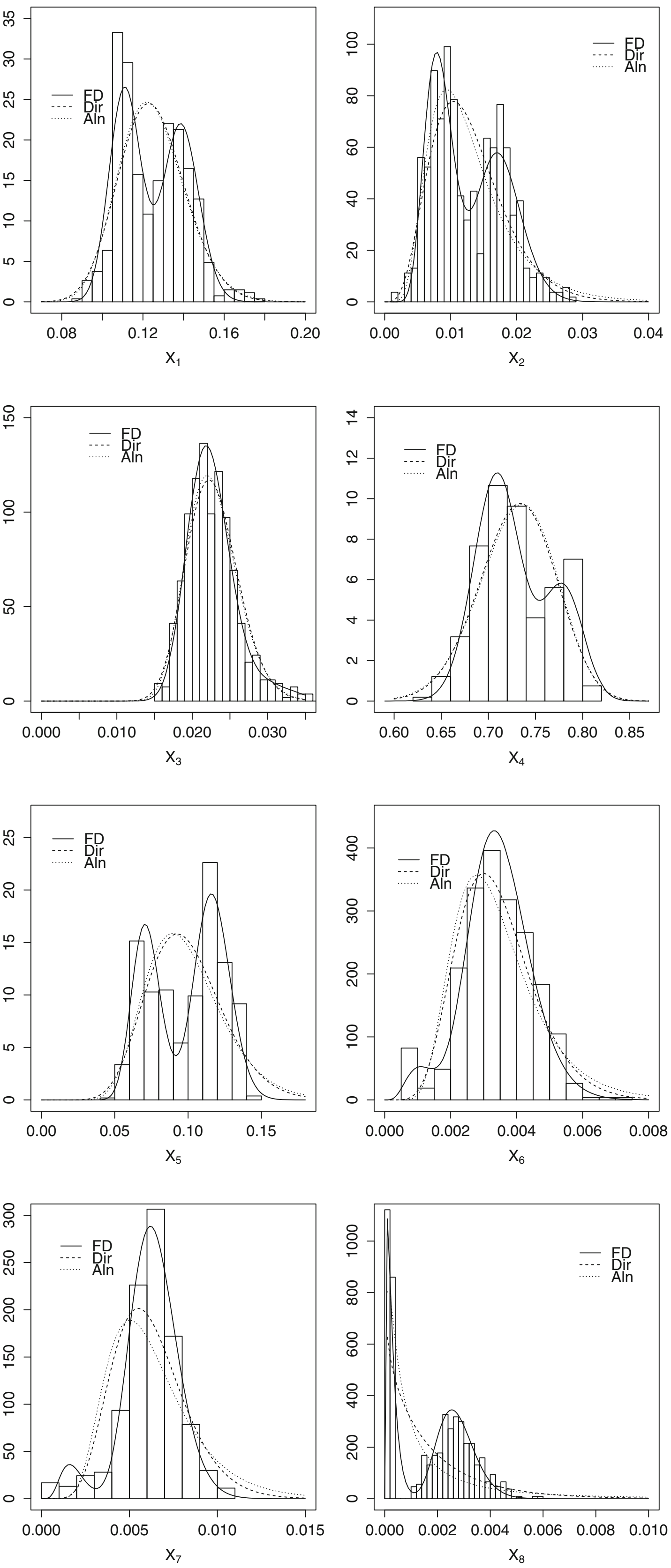

Table 6 2-part compositions: goodness of fit measures

Variables		Dir	FD	ALN
X_1 = palmitic	AIC	−2885	−2932	−2889
	BIC	−2876	−2915	−2880
X_2 = palmitoleic	AIC	−4101	−4141	−4073
	BIC	−4093	−4123	−4064
X_3 = stearic	AIC	−4556	−4581	−4569
	BIC	−4547	−4564	−4560
X_4 = oleic	AIC	−1905	−1970	−1900
	BIC	−1896	−1953	−1892
X_5 = linoleic	AIC	−2415	−2606	−2356
	BIC	−2407	−2589	−2388
X_6 = linolenic	AIC	−5746	−5828	−5676
	BIC	−5738	−5811	−5668
X_7 = arachidic	AIC	−5127	−5353	−4990
	BIC	−5119	−5336	−4981
X_8 = eicosenoic	AIC	−5736	−6182	−5653
	BIC	−5728	−6165	−5644

Table 7 Parameter estimates, standard errors of estimators and measures (8) of cluster distance for FD model

Variables	(X_2, X_3)
$\hat{\boldsymbol{\alpha}} = (15.1, 40.3, 1719.4)$	$\text{SE}(\hat{\boldsymbol{\alpha}}) = (0.79, 2.04, 89.09)$
$\hat{\boldsymbol{p}} = (0.465, 0.032, 0.503)$	$\text{SE}(\hat{\boldsymbol{p}}) = (0.024, 0.012, 0.026)$
$\hat{\tau} = 16.58$	$\text{SE}(\hat{\tau}) = 1.01$
$k(\boldsymbol{\alpha}, \tau) = (12.57, 5.77, 0.16)$	

for about 12 %, the linoleic acid for about 10 % and each of the remaining acids represents less than 2.3 %. This data set was originally presented by Forina et al. (1983). It was subsequently analyzed by various authors and it is made available, for example, in Azzalini et al. (2012). Here, our aim is mainly to assess the validity of the FD as a general model for compositional data. To this end, we decided to focus on the fitting performances of the FD model when applied to data displaying a large variety of different patterns. The olive oil data set has been chosen since it does meet the requirement, as we shall see. In order to achieve the maximum variety of different data configurations, we chose to focus on all 2-part and 3-part compositions, i.e. the one and two-dimensional marginals. These compositions have been chosen because the graphical representations available in these low-dimensional cases yield a deeper understanding of the model and data features. A detailed cluster analysis of the data set in relation to the three geographical macro-areas could also be performed by selecting proper amalgamations and/or subcompositions, but it goes beyond the scope of the present paper (though we shall give some comments on that).

The data set presented 37 zero values relative to the 6-th (linolenic) or to the 7-th (arachidic) variables. This could be dealt with by amalgamation of such variables with some others. Nevertheless, in order to preserve the explicative potential of such two variables, we preferred to neglect the sampling units because they represent a small fraction of the total sample.

In a comparative perspective, two other models for compositional data were considered: the Dirichlet (Dir) and the additive logistic normal (ALN). Although not specifically designed for cluster purposes, they are among the most well-known and widespread models. The ALN is obtained by assigning a multivariate normal distribution to the variables $Y_i = \log(X_i/X_D)$ $(i = 1, \ldots, D-1)$ (Aitchison 2003). Other logratio models might have been considered, e.g. the ones obtained by replacing alr transformations with isometric ilr ones. However, to our purposes (comparing estimated densities and likelihood-based measures of fitting) these models would have produced identical results. This is because ilr transformations are linearly related to alr ones, and the normal distribution is invariant under linear transformations, so that the isometric models can be viewed as a re-parametrization of the ALN one.

We first analyze the eight 2-part compositions obtained by choosing a single variable and amalgamating the remaining ones. These eight marginals cover many different patterns including symmetry and asymmetry, perfect unimodality as well as bimodality with different degrees of cluster separation.

The histograms of the variables and the densities of the fitted models (Fig. 3) highlight the good fitting of the FD model in both the unimodal and the bimodal cases. In particular, the inability of the Dirichlet and of the ALN to capture bimodality is evident. Moreover, even in the unimodal data sets, the FD, through its two-cluster structure, allows to better fit both the tails and the central body.

The better fit of the FD is confirmed by the goodness of fit measures AIC and BIC reported in Table 6.

Incidentally, note that the two clusters highlighted by the FD density in the plot of variable X_8, nearly perfectly separate the southern macro-area from the other two. In the other bimodal FD density plots the southern macro-area data are spread over the whole range of the variables; in the plots of variables X_1 and X_2 Sardinia and northern Italy data belong to the same cluster, while in the plots of variables X_4 and X_5 they belong to separated ones.

Let us now focus on 3-part compositions obtained by choosing two variables and amalgamating the remaining ones. The 28 data sets thereby obtained show many different shapes with various location, variability and dependence behaviors, as well as the presence of unimodality and multimodality. A peculiar feature of the data set is the unusually large number of positive correlations. Indeed, the unitary sum

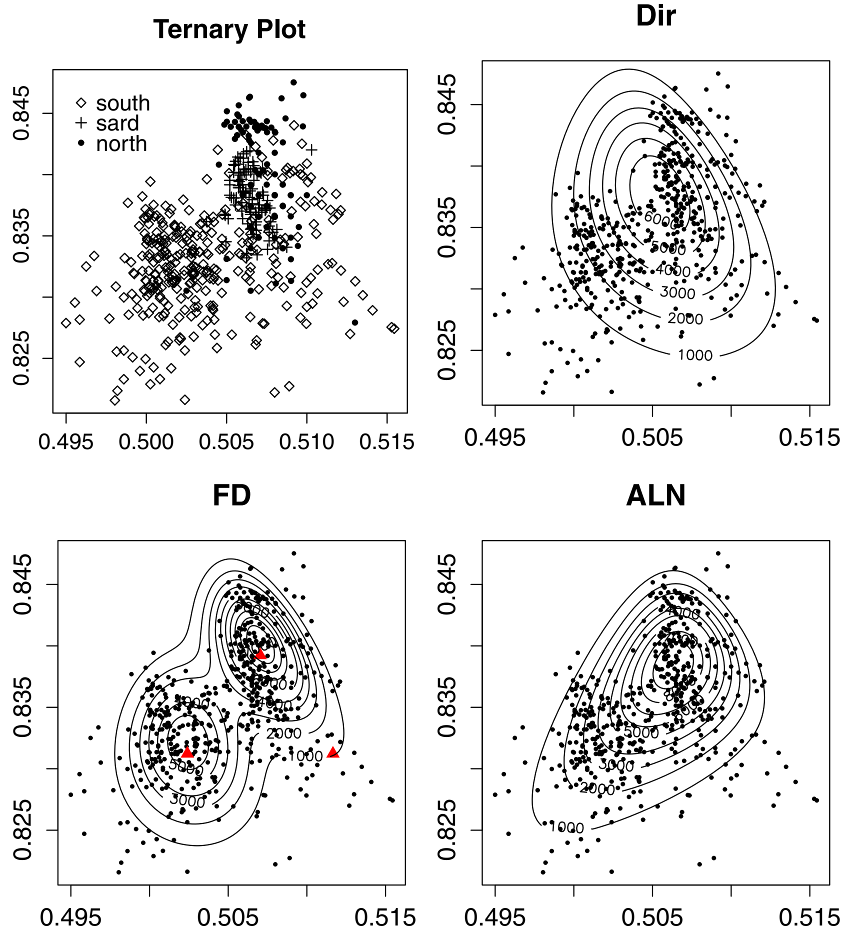

Fig. 4 Zoomed areas of ternary diagrams representing plot and contours of estimated Dirichlet, FD and ALN models of 3-part composition X_2 = palmitoleic and X_3 = stearic

Table 8 Goodness of fit measures of 3-part composition X_2 and X_3

	Dir	FD	ALN
AIC	−8379	−8706	−8656
BIC	−8366	−8681	−8634

Table 9 Parameter estimates, standard errors of estimators and measures (8) of cluster distance for FD model

Variables	(X_6, X_7)
$\hat{\boldsymbol{\alpha}} = (9.6, 5.1, 2815.8)$	$\text{SE}(\hat{\boldsymbol{\alpha}}) = (0.43, 0.54, 125.99)$
$\hat{\boldsymbol{p}} = c(0, 0.945, 0.055)$	$\text{SE}(\hat{\boldsymbol{p}}) = (0.002, 0.011, 0.011)$
$\hat{\tau} = 12.29$	$\text{SE}(\hat{\tau}) = 0.75$
$k(\boldsymbol{\alpha}, \tau) = (11.96, 17.97, 0.06)$	

constraint of compositional data typically induces negative correlation, so that correlation cannot be given the ordinary interpretation. Specifically, in any 3-part composition at most 1 out of the 3 possible correlation coefficients can be positive. Remarkably, our data set presents such a positive correlation in 24 out of 28 cases.

Although the FD model only entails negative correlations, it performs better than the other two models, in terms both of AIC and of BIC, in 18 out of 28 cases. In the remaining ones the ALN outperforms the other models. Moreover, the FD shows a better fit than the Dirichlet 27 times according to AIC and 25 according to BIC.

A careful graphical inspection highlights that the FD performs better in both the unimodal and multimodal cases, unless a high positive correlation is present and/or the location of the clusters (when present) is clearly incompatible with the one entailed by the FD, which are somewhat related issues. More specifically, the FD always outperforms the ALN when correlations are all negative, in 13 cases out of 16 when there is a small (< 0.4) positive correlation, and in 1 case out of 8 with high (≥ 0.4) correlation. The 3 data sets with a small positive correlation and better performances of the ALN display 3 clusters roughly centered along the same line, which is strongly at odds with the FD model. The only case where the Dirichlet performs better (in terms both of AIC and BIC) than the FD displays an extraordinarily high correlation equal to 0.85.

Because of space constraints, we only report 3 two-dimensional marginals which exemplify some interesting

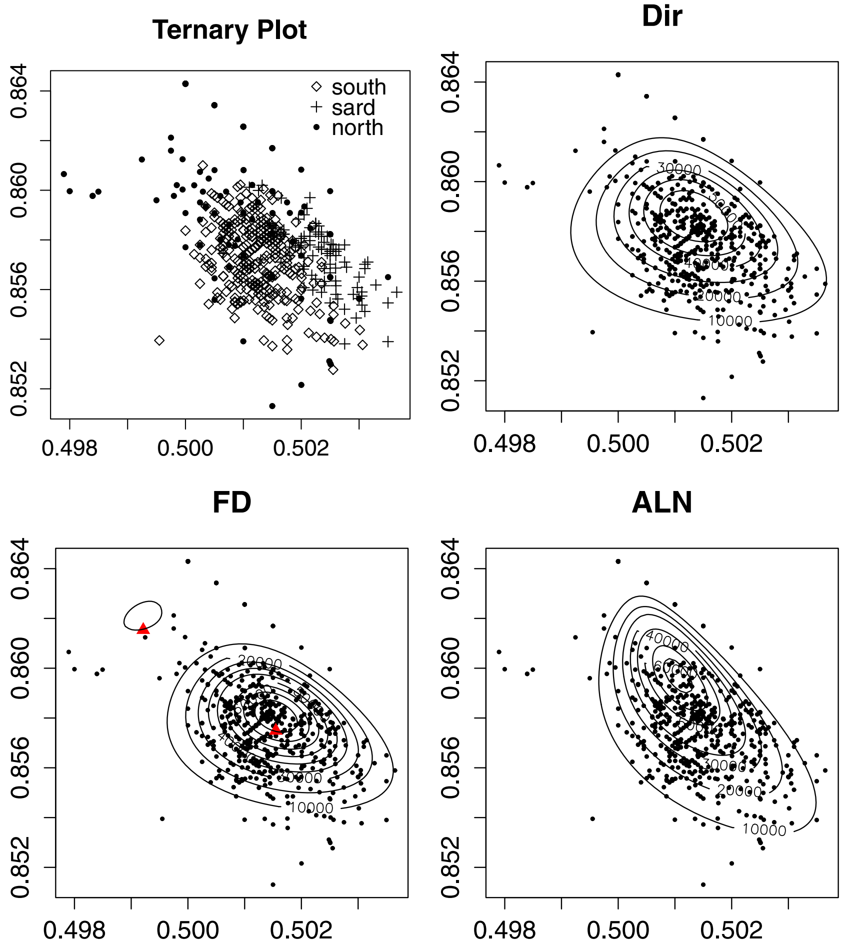

Fig. 5 Zoomed areas of ternary diagrams representing plot and contours of estimated Dirichlet, FD and ALN models of the 3-part composition X_6 = linolenic and X_7 = arachidic

Table 10 Goodness of fit measures of 3-part composition X_6 and X_7

	Dir	FD	ALN
AIC	−10,834	−11,090	−10,826
BIC	−10,821	−11,065	−10,805

situations. All scatterplots are represented on the (relevant part of) the ternary diagram RSP^2, i.e. only areas of the ternary triangles where data are concentrated are shown.

Let us first consider the pair X_2 = palmitoleic and X_3 = stearic (and third part equal to $1 - X_2 - X_3$). The correlation coefficients (−0.21, −0.8, −0.42) are all negative, which is coherent with the correlation structure of the FD model. The maximum likelihood estimates of the FD parameters are reported in Table 7.

The scatterplot highlights the presence of two main clusters. One cluster contains nearly all data from Sardinia and northern Italy and part of the southern Italy ones, and the other one the remaining southern Italy data. Only the FD is able to capture this feature (see Fig. 4).

The better performance of the FD is confirmed by the goodness of fit measures reported in Table 8.

Let us now consider a roughly unimodal data set given by the pair X_6 = linolenic and X_7 = arachidic. The correlation coefficients are (+0.38, −0.73, −0.91). The maximum likelihood estimates, the isodensity contour plots and the goodness of fit measures are reported in Table 9, Fig. 5 and Table 10 respectively.

Despite the positive correlation between X_6 and X_7, the FD allows for a better fitting of the central body of the data, resulting in remarkably lower values of the AIC and the BIC. Moreover, the FD is less influenced by extreme data.

Finally, it is of some interest to focus on the pair X_7 = arachidic and X_8 = eicosenoic, because it displays a special type of bimodality. Indeed, one of the clusters (see Fig. 6) has a typical bell shape, but the other one looks very squeezed along the $X_8 = 0$ axis. The FD captures the latter cluster by placing two very close components along this axis. The bell shape cluster, which is well modeled by the FD, corresponds almost perfectly to the southern Italy group. The northern Italy and Sardinia data are quite overlapped, with the Sardinia ones more concentrated on the left. This roughly agrees with the two FD components placed along the $X_8 = 0$ axis. The maximum likelihood estimates and the goodness of fit measures are reported in Tables 11 and 12, respectively.

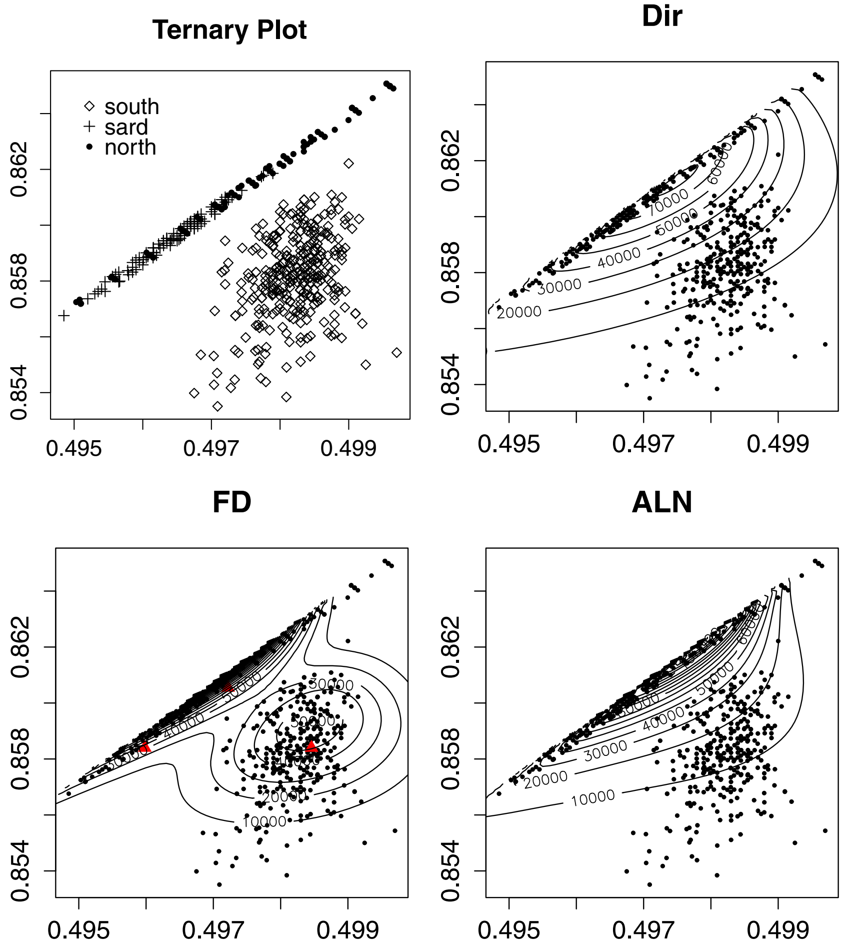

Fig. 6 Zoomed areas of ternary diagrams representing plot and contours of estimated Dirichlet, FD and ALN models of the 3-part composition X_7 = arachidic and X_8 = eicosenoic

Table 11 Parameter estimates, standard errors of estimators and measure (8) of cluster distance for FD model

Variables	(X_7, X_8)
$\hat{\boldsymbol{\alpha}} = (13.5, 0.9, 2244.5)$	$\mathrm{SE}(\hat{\boldsymbol{\alpha}}) = (0.67, 0.05, 119.68)$
$\hat{\boldsymbol{p}} = (0.061, 0.584, 0.355)$	$\mathrm{SE}(\hat{\boldsymbol{p}}) = (0.030, 0.022, 0.037)$
$\hat{\tau} = 5.61$	$\mathrm{SE}(\hat{\tau}) = 0.33$
$k(\boldsymbol{\alpha}, \tau) = (2.02, 14.61, 0.01)$	

Table 12 Goodness of fit measures of 3-part composition X_7 and X_8

	Dir	FD	ALN
AIC	−10,746	−11,105	−10,669
BIC	−10,733	−11,079	−10,647

9 Concluding remarks

When standard unimodal models fail properly to fit real data sets, as it is often the case of the Dirichlet, a fruitful option is to resort to properly designed structured mixtures, like the FD. Our results show that the latter greatly expands the modeling potential of the Dirichlet, without demanding the intricacy of a general unstructured mixture.

The FD can be viewed as a basic mixture of Dirichlet distributions with common precision parameter. The core element of the FD, namely the structure of its component means, is characterized by strong links on the one hand, and a straightforward interpretation on the other. As regards the former aspect, there is only one degree of freedom dictating the component relationships. Thus, the FD is among the most parsimonious Dirichlet mixture models with general mixing weights. This feature makes it possible to keep the theoretical and computational issues of FD estimation at a quite uncomplicated level, in many respects even comparable to common non-mixture models. Furthermore, simulations show that the resulting estimators display an unusually accurate behavior compared to general unstructured mixture model ones.

Despite the strong component links, the FD is able to grasp the main features of a variety of data sets, including unimodal and multimodal cases. In particular, the real data set analysis indicates a clear general superiority over the Dirichlet model and a better fitting than the ALN for many data patterns.

Thus, the FD appears to efficiently exploit the added flexibility of its simple mixture structure, achieving an optimal compromise between ability of data modeling and inferential tractability.

However, not all types of interesting compositional data sets can be adequately described by the FD mixture structure. Thus, it seems promising to look for more flexible structured mixture models, though still inferentially tractable. In particular, our analysis underlines the opportuneness of handling more general component mean locations, as well as positive correlations. This will be the object of future research.

Acknowledgements We are grateful to the referees and to the editor for their constructive comments, which helped to improve the paper. Research partially financially supported by the Italian Ministry of University and Research, Grants F.A. 2014 from the University of Milano-Bicocca.

Appendix

Appendix 1: Vertexes of *RSP*^{D} (see Sect. 3.1)

Having chosen as one dimensional simplex RSP^1 the line segment with vertexes $\mathbf{v}_0 = 0$ and $\mathbf{v}_1 = 1$, let us recursively determine the vertexes of RSP^n ($n \leq D-1$). Suppose we know the vertexes of RSP^{n-1}, that is $\mathbf{v}_0^{n-1}, \mathbf{v}_1^{n-1}, \ldots, \mathbf{v}_{n-1}^{n-1} \in \mathbb{R}^{n-1}$, $n \geq 2$. Then, RSP^n can be obtained by adding to the n vertexes of RSP^{n-1} a new vertex $\mathbf{v}_n$ with the same distance 1 from all old vertexes. Therefore, the first n vertexes $\mathbf{v}_0^n, \mathbf{v}_1^n, \ldots, \mathbf{v}_{n-1}^n \in \mathbb{R}^n$ of RSP^n are obtained by adding a further coordinate equal to 0 to the RSP^{n-1} vertexes (geometrically the first n vertexes of RSP^n coincide with the vertexes of RSP^{n-1}).

The last vertex $\mathbf{v}_n^n$, having the same distance from all previous vertexes, has the first $n-1$ coordinates equal to the barycenter $\mathbf{B}^{n-1}$ of RSP^{n-1}. The last coordinate is then obtained by imposing that the distance between $\mathbf{v}_n^n$ and one of the previous vertexes is one. In particular, we can choose $\left\|\mathbf{v}_n^n - \mathbf{v}_0^n\right\| = \left\|\mathbf{v}_n^n\right\| = 1$.

It is left to determine the coordinates of the generic barycenter $\mathbf{B}^n$. Again we will proceed recursively, starting from $\mathbf{B}^1 = 1/2$. As $\mathbf{B}^n$ has the same distance from all vertexes of RSP^n, its first $n-1$ coordinates must be equal to $\mathbf{B}^{n-1}$. The last coordinate can be obtained by imposing that $\left\|\mathbf{B}^n - \mathbf{v}_n^n\right\|$ is equal to the distance between $\mathbf{B}^n$ and one of the first n vertexes of RSP^n. For example, we can set $\left\|\mathbf{B}^n - \mathbf{v}_n^n\right\| = \left\|\mathbf{B}^n - \mathbf{v}_0^n\right\| = \|\mathbf{B}^n\|$. As the first $n-1$ coordinates of $\mathbf{v}_n^n$ are equal to $\mathbf{B}^{n-1}$ and $\left\|\mathbf{v}_n^n\right\| = 1$, we can write

$$\begin{aligned}\left\|\mathbf{B}^n - \mathbf{v}_n^n\right\| &= B_n^n - \sqrt{1 - \left\|\mathbf{B}^{n-1}\right\|^2} \\ &= \sqrt{\left\|\mathbf{B}^{n-1}\right\|^2 + (B_n^n)^2} = \|\mathbf{B}^n\| \qquad (16)\end{aligned}$$

where B_n^n is the last coordinate of $\mathbf{B}^n$. One can then find B_n^n as a function of $\left\|\mathbf{B}^{n-1}\right\|$ by using the second equality in (16). By plugging this expression in the last equality of (16), after some manipulation one arrives at the following recursive relation:

$$\|\mathbf{B}^n\|^2 = \left[4\left(1 - \left\|\mathbf{B}^{n-1}\right\|^2\right)\right]^{-1}, \qquad n \geq 2.$$

This recursive equation, together with the initial value $\left\|\mathbf{B}^1\right\|^2 = 1/4$, admits the explicit solution given by $\|\mathbf{B}^n\|^2 = n/[2(1+n)]$. By (16) one then has

$$B_n^n = [2n(1+n)]^{-1/2}$$

which coincides with the quantity a_i with $i = n$ defined in Sect. 3.1. As the first $n-1$ coordinates of $\mathbf{B}^n$ are equal to $\mathbf{B}^{n-1}$, by induction one explicitly determines $\mathbf{B}^n$: its i-th coordinate is $B_i^n = [2i(1+i)]^{-1/2} = a_i$, $i = 1, \ldots, n$, $n \geq 1$. By applying recursively the above described procedure for the derivation of the vertexes $\mathbf{v}_0^n, \mathbf{v}_1^n, \ldots, \mathbf{v}_n^n$ of RSP^n from the vertexes $\mathbf{v}_0^{n-1}, \mathbf{v}_1^{n-1}, \ldots, \mathbf{v}_{n-1}^{n-1}$ of RSP^{n-1} one obtains the result.

Appendix 2: Lemma 1—Rate of divergence of Dirichlet log-likelihood

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an i.i.d. sample from a Dirichlet $\mathcal{D}(\mathbf{x}; \boldsymbol{\alpha})$. Then the log-likelihood diverges to $-\infty$ when at least one of the α_i's goes to zero.

Moreover, suppose $n \geq 2$. Then a.s. the log-likelihood:

1. is bounded from above;
2. diverges to $-\infty$ when at least one of the α_i's diverges to $+\infty$, with a rate of divergence not smaller than $n\,\alpha^+ \log \sum_{i=1}^D x_i^{(0)}$.

Here $x_i^{(0)} = \prod_{j=1}^n x_{ji}^{1/n}$ denotes the geometric mean of the i-th element of the observations.

Suppose now $n = 1$. Then the log-likelihood:

1. is unbounded (from above and from below);
2. a necessary condition for its divergence to $+\infty$ is that all α_i's diverge to $+\infty$ and the rate of divergence is not larger than $\frac{1}{2}(D-1)\log \alpha^+$;
3. diverges to $-\infty$ if at least one of the α_i's diverges to $+\infty$ and at least one does not.

Proof The Dirichlet log-likelihood $l(\boldsymbol{\alpha})$ can be written as

$$\frac{1}{n} l(\boldsymbol{\alpha}) = \log \Gamma(\alpha^+) - \sum_{i=1}^D \log \Gamma(\alpha_i) + \sum_{i=1}^D \alpha_i \log x_i^{(0)}.$$

As $\log \Gamma(y) \approx -\log y$ as $y \to 0$, then $l(\boldsymbol{\alpha}) \to -\infty$ when one or more of the α_i's go to 0 and the others are fixed. Because $l(\boldsymbol{\alpha})$ is a regular function, this implies that it is bounded on the set $\{\alpha^+ \in (0, k]\}$ for any $k > 0$ and, therefore, it may diverge only if $\alpha^+ \to +\infty$. Thus, suppose M $(1 \le M \le D)$ α_i's go to $+\infty$, say $\alpha_1, \ldots, \alpha_M$ without loss of generality, and denote α_1^+ their sum and $\alpha_2^+ = \alpha^+ - \alpha_1^+$. Then, the following approximation holds

$$\frac{1}{n} l(\boldsymbol{\alpha}) \approx \alpha_1^+ \sum_{i=1}^{M} \eta_i \log \frac{x_i^{(0)}}{\eta_i} + \alpha_2^+ \log \alpha_1^+ + \frac{1}{2}\left(\sum_{i=1}^{M} \log \alpha_i - \log \alpha_1^+\right) \tag{17}$$

where $\eta_i = \alpha_i / \alpha_1^+$. Formula (17) can be obtained by means of a careful expansion of the terms of $l(\boldsymbol{\alpha})$ based on the two following approximations valid as $y \to \infty$:

$$\log \Gamma(y) = \left(y - \frac{1}{2}\right) \log y - y + \frac{1}{2} \log \pi + O\left(\frac{1}{y}\right)$$
$$\frac{\Gamma(y+a)}{\Gamma(y)} \approx y^a$$

the latter holding for fixed positive a.

The relation between geometric and arithmetic means implies that

$$\prod_{i=1}^{M} \left(\frac{x_i^{(0)}}{\eta_i}\right)^{\eta_i} \le \sum_{i=1}^{M} \frac{x_i^{(0)}}{\eta_i} \eta_i$$

and therefore:

$$\sum_{i=1}^{M} \eta_i \log \frac{x_i^{(0)}}{\eta_i} \le \log \sum_{i=1}^{M} x_i^{(0)}.$$

Now, suppose $n \ge 2$. Then all elements of the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are distinct a.s.. Thus $\sum_{i=1}^{M} x_i^{(0)} < \sum_{i=1}^{M} \bar{x}_i \le 1$. It follows that $l(\boldsymbol{\alpha})$ goes to $-\infty$ when at least one $\alpha_i \to +\infty$ and the rate of divergence is not smaller than $n\, \alpha^+ \log \sum_{i=1}^{D} x_i^{(0)}$. It also follows that $l(\boldsymbol{\alpha})$ is bounded a.s..

Suppose now that $n = 1$, so that $x_i^{(0)} = \bar{x}_i = x_{1i}$ $(i = 1, \ldots, D)$. If $1 \le M < D$ then $\sum_{i=1}^{M} x_i^{(0)} = \sum_{i=1}^{M} x_{1i} < 1$, and therefore the log-likelihood decreases to $-\infty$. If $M = D$, then $\sum_{i=1}^{D} \eta_i \log \frac{x_{1i}}{\eta_i}$ achieves its maximum equal to zero if we set $\eta_i = x_{1i}$ $(i = 1, \ldots, D)$. Thus, if $M = D$, the behavior of $l(\boldsymbol{\alpha})$ is determined by the term $(\sum_{i=1}^{D} \log \alpha_i - \log \alpha^+)/2$. The latter can be shown to be smaller or equal to $(D-1)\log \alpha^+ - D \log D$ again by using the relation between geometric and arithmetic means. Hence, the rate of divergence of $l(\boldsymbol{\alpha})$ is not larger than $(D-1)\log \alpha^+$.This rate can be exactly achieved if $\eta_i = x_{1i}$ $(i = 1, \ldots, D)$, which also shows that $l(\boldsymbol{\alpha})$ is indeed unbounded.

Note that the above arguments also imply that the log-likelihood diverges to $-\infty$ when at least one of the α_i's goes to zero for any n even if $M < D$ of the other α_i's diverge to $+\infty$. □

Appendix 3: Proof of Proposition 3

To prove a.s. boundedness of the log-likelihood $l(\boldsymbol{\theta})$ and the existence of a maximum we shall use the following upper bound:

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log \sum_{i=1}^{D} p_i f_D(\mathbf{x}_j; \boldsymbol{\alpha_i}) \le \max_{I_j, j=1,\ldots,n} \sum_{j=1}^{n} \log f_D(\mathbf{x}_j; \boldsymbol{\alpha}_{I_j}) = l_U(\boldsymbol{\alpha}, \tau) \tag{18}$$

where $I_j \in \{1, \ldots, D\}$ can be interpreted as the cluster to which observation $\mathbf{x}_j$ has been assigned.

For ease of exposition, let us show boundedness first. By formula (18), we have:

$$\sup_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \le \sup_{\boldsymbol{\theta}} \max_{I_j, j=1,\ldots,n} \sum_{j=1}^{n} \log f_D(\mathbf{x}_j; \boldsymbol{\alpha}_{I_j}) = \max_{I_j, j=1,\ldots,n} \sup_{\boldsymbol{\theta}} \sum_{j=1}^{n} \log f_D(\mathbf{x}_j; \boldsymbol{\alpha}_{I_j}).$$

Therefore, it is enough to show that, for any given allocation of the observations to the D Dirichlet clusters (i.e. for any I_j, $j = 1, \ldots, n$), the corresponding log-likelihood is bounded from above. Indeed, such log-likelihood coincides with the classified log-likelihood given by (14) and can be viewed as the sum of D Dirichlet log-likelihoods:

$$\sum_{j=1}^{n} \log f_D(\mathbf{x}_j; \boldsymbol{\alpha}_{I_j}) = \sum_{i=1}^{D} \sum_{j \in A_i} \log f_D(\mathbf{x}_j; \boldsymbol{\alpha}_i)$$

where $A_i = \{j : z_j = i\}$ identify the observations assigned to cluster i. By Lemma 1, a necessary condition for any of the D Dirichlet log-likelihoods to be unbounded is that there exists at least one cluster with only one observation and all α_i's diverge to $+\infty$ except at most one (in which case τ goes to $+\infty$ as well). In this case the sum of the log-likelihoods of the clusters (at most $D-1$) with only one observation diverges with a rate not larger than $c_1 \log(\alpha^+ + \tau)$ where c_1 is a positive constant. On the other hand, there must exist at least one cluster with two or more observations. Therefore, when some α_i's go to $+\infty$, by Lemma 1 the corresponding log-likelihood tends a.s.

to $-\infty$ with a rate not smaller than $-c_2(\alpha^+ + \tau)$ where c_2 is a positive constant. Thus, in this case, the classified likelihood diverges to $-\infty$ which implies that $l(\boldsymbol{\theta})$ is a.s. bounded.

Let us now prove existence of a maximum. As the log-likelihood $l(\boldsymbol{\theta})$ is a regular and differentiable function, the existence of a global maximum can be proved by showing that the supremum is not reached at the boundary of the parameter space. More precisely, consider the frontier of $\boldsymbol{\Theta}$ defined as the set of boundary points which are not actually in $\boldsymbol{\Theta}$. We shall show that, when $\boldsymbol{\theta}$ tends to the frontier (i.e. $\alpha_i \to 0$ or $\alpha_i \to \infty$, $p_i \to 1$ $i = 1, \ldots, D$, $\tau \to 0$ or $\tau \to \infty$), then the log-likelihood tends either to $-\infty$ or to values not larger than the log-likelihood (based on the whole sample) of the Dirichlet distribution. As this distribution corresponds to interior points of the parameter space of the FD, such limiting values are dominated by $l(\boldsymbol{\theta})$ computed at those interior points. They can therefore be discarded.

To obtain the above limits, we shall study the upper bound $l_U(\boldsymbol{\alpha}, \tau)$ of $l(\boldsymbol{\theta})$ given in (18).

Suppose first that at least one of the α_i's goes to $+\infty$, irrespectively of the behavior of the other parameters. For any given allocation of the observations to clusters (i.e. for any given $I_1, \ldots, I_n$), there must exist at least one cluster with two or more observations. By Lemma 1, the corresponding Dirichlet log-likelihood tends to $-\infty$ with rate dominating the log-likelihood of possible one-observation clusters. Thus, for any given allocation, the classified log-likelihood tends to $-\infty$ and so does the upper bound $l_U(\boldsymbol{\alpha}, \tau)$. An analogous argument shows that $l(\boldsymbol{\theta})$ tends to $-\infty$ even when τ tends to $+\infty$.

Suppose now that two or more α_i's go to zero. Then, whatever the allocation of the observations, in all Dirichlet cluster log-likelihoods there exists one parameter going to zero. Therefore, each Dirichlet log-likelihood goes to $-\infty$, implying that $l_U(\boldsymbol{\alpha}, \tau)$ diverges as well.

Consider, instead, the case of a single α, say α_1, going to zero. Then, for all allocations with at least one observation not assigned to the first cluster, the corresponding classified log-likelihood tends to $-\infty$. This is because the term corresponding to the first cluster tends to a finite value while all the others tend to $-\infty$. On the other hand, if all observations are assigned to the first cluster, then the classified log-likelihood tends to a Dirichlet log-likelihood, computed on the whole sample, with parameter $(\tau, \alpha_2, \ldots, \alpha_D)$. It follows that $l_U(\boldsymbol{\alpha}, \tau)$ tends to the same limit as well. The latter limit is dominated by $l(\boldsymbol{\theta})$ computed at an interior point of $\boldsymbol{\theta}$.

Finally, if $p_i \to 1$, $i = 1, \ldots, D$, or $\tau \to 0$, then it is straightforward to see that $l(\boldsymbol{\theta})$ converges to a Dirichlet log-likelihood and, again, it is dominated by the value of $l(\boldsymbol{\theta})$ at an interior point.

Appendix 4: Score statistic and information matrix of the complete-data likelihood

The elements $s_c(\theta_r) = \partial \log L_c(\boldsymbol{\theta}) / \partial\theta_r$ $(r = 1, \ldots, 2D)$ of the score statistic $\mathbf{S}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ computed from (12) have the form:

$$
\begin{aligned}
s_c(p_i) &= \frac{z_{.i}}{p_i} - \frac{z_{.D}}{p_D} \qquad (i = 1, \ldots, D-1) \\
s_c(\alpha_i) &= n\psi(\alpha^+ + \tau) - z_{.i}\left[\psi(\alpha_i + \tau) - \psi(\alpha_i)\right] \\
&\quad -n\psi(\alpha_i) + \sum_{j=1}^{n} \log x_{ji} \quad (i = 1, \ldots, D) \\
s_c(\tau) &= n\psi(\alpha^+ + \tau) - \sum_{i=1}^{D} z_{.i}\psi(\alpha_i + \tau) \\
&\quad + \sum_{i=1}^{D}\sum_{j=1}^{n} z_{ji} \log x_{ji}
\end{aligned}
\qquad (19)
$$

where $z_{.i} = \sum_{j=1}^{n} z_{ji}$, $(i = 1, \ldots, D)$.

The elements $i_c(\theta_r, \theta_p) = \partial^2 \log L_c(\boldsymbol{\theta}) / \partial\theta_r \partial\theta_p$ $(r, p = 1, \ldots, 2D)$ of the $2D \times 2D$ matrix $\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ assume the following form:

$$
\begin{aligned}
i_c(p_i, p_i) &= \frac{z_{.i}}{p_i^2} + \frac{z_{.D}}{p_D^2} \quad (i = 1, \ldots, D-1) \\
i_c(p_i, p_h) &= \frac{z_{.D}}{p_D^2} \quad (i \neq h;\ \ i, h = 1, \ldots, D-1) \\
i_c(p_i, \alpha_h) &= 0 \quad (i = 1, \ldots, D-1) \quad (h = 1, \ldots, D) \\
i_c(p_i, \tau) &= 0 \quad (i = 1, \ldots, D-1) \\
i_c(\alpha_i, \alpha_i) &= -n\psi'(\alpha^+ + \tau) + z_{.i}\left[\psi'(\alpha_i + \tau) - \psi'(\alpha_i)\right] \\
&\quad +n\psi'(\alpha_i) \quad (i = 1, \ldots, D) \\
i_c(\alpha_i, \alpha_h) &= -n\psi'(\alpha^+ + \tau) \quad (i \neq h;\ \ i, h = 1, \ldots, D) \\
i_c(\alpha_i, \tau) &= -n\psi'(\alpha^+ + \tau) + z_{.i}\psi'(\alpha_i + \tau)(i = 1, \ldots, D) \\
i_c(\tau, \tau) &= -n\psi'(\alpha^+ + \tau) + \sum_{i=1}^{D} z_{.i}\psi'(\alpha_i + \tau)
\end{aligned}
\qquad (20)
$$

where $\psi'(\cdot)$ denotes the trigamma function.

References

Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (2003)

Azzalini A, Menardi G, Rosolin T (2012) R package pdfCluster: cluster analysis via nonparametric density estimation (version 1.0-0). Università di Padova, Italia. http://cran.r-project.org/web/packages/pdfCluster/index.html

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

Barndorff-Nielsen, O., Jørgensen, B.: Some parametric models on the simplex. J. Multivar. Anal. **39**(1), 106–116 (1991)

Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Stat. Data Anal **41**, 561–575 (2003)

Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **4**, 315–332 (1992)

Celeux, G., Chauveau, D., Diebolt, J.: Stochastic versions of the EM algorithm: an experimental study in the mixture case. J. Stat. Comput. Simul. **55**, 287–314 (1996)

Connor, R.J., Mosimann, J.E.: Concepts of independence for proportions with a generalization of the dirichlet distribution. J. Am. Stat. Assoc. **64**(325), 194–206 (1969)

Coxeter, H.: Regular Polytopes. Dover Publications, New York (1973)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser B **39**(1), 1–38 (1977)

Diebolt, J., Ip, E.: Stochastic EM: method and application. In: WR Gilks, S.R., Spiegelhalter, D. (eds.) Markov Chain Monte Carlo in Practice, pp. 259–273. Chapman & Hall, London (1996)

Efron, B.: Missing data, imputation, and the bootstrap. J. Am. Stat. Assoc. **89**(426), 463–475 (1994)

Favaro, S., Hadjicharalambous, G., Prunster, I.: On a class of distributions on the simplex. J. Stat. Plan. Inference **141**(426), 2987–3004 (2011)

Feng, Z., McCulloch, C.: Using bootstrap likelihood ratio in finite mixture models. J. R. Stat. Soc. B **58**, 609–617 (1996)

Forina M, Armanino C, Lanteri S, Tiscornia E (1983) Classification of olive oils from their fatty acid composition. In: Martens, Russwurm (eds) Food Research and Data Anlysis, Dip. Chimica e Tecnologie Farmaceutiche ed Alimentari, University of Genova

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, New York (2006)

Gupta, R.D., Richards, D.S.P.: Multivariate liouville distributions. J. Multivar. Anal. **23**, 233–256 (1987)

Gupta, R.D., Richards, D.S.P.: Multivariate liouville distributions, II. Probab. Math. Stat. **12**, 291–309 (1991)

Gupta, R.D., Richards, D.S.P.: Multivariate liouville distributions, III. J. Multivar. Anal. **43**, 29–57 (1992)

Gupta, R.D., Richards, D.S.P.: Multivariate liouville distributions, IV. J. Multivar. Anal. **54**, 1–17 (1995)

Gupta, R.D., Richards, D.S.P.: Multivariate liouville distributions, V. In: NL Johnson, N.B. (ed.) Advances in the Theory and Practice of Statistics: A Volume in Honour of Samuel Kotz, pp. 377–396. Wiley, New York (1997)

Gupta, R.D., Richards, D.S.P.: The covariance structure of the multivariate liouville distributions. Contemp. Math. **287**, 125–138 (2001a)

Gupta, R.D., Richards, D.S.P.: The history of the Dirichlet and Liouville distributions. Int. Stat. Rev. **69**(3), 433–446 (2001b)

Hathaway, R.J.: A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann. Stat. **13**(2), 795–800 (1985)

Kiefer, J., Wolfowitz, J.: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. **27**(4), 887–906 (1956)

Lehmann, E., Casella, G.: Theory of Point Estimation. Springer, New York (1998)

Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44**(2), 226–233 (1982)

McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

Meilijson, I.: A fast improvement to the EM algorithm on its own terms. J. R. Stat. Soc. Ser. B **51**(1), 127–138 (1989)

Meng, X.L., Rubin, D.B.: Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. J. Am. Stat. Assoc. **86**(416), 899–909 (1991)

O'Hagan, A., Murphy, T.B., Gormley, I.C.: Computational aspects of fitting mixture models via the expectation-maximization algorithm. Comput. Stat. Data Anal. **56**(12), 3843–3864 (2012)

Ongaro, A., Migliorati, S.: A generalization of the Dirichlet distribution. J. Multivar. Anal. **114**, 412–426 (2013)

Palarea-Albaladejo, J., Martín-Fernández, J., Soto, J.: Dealing with distances and transformations for fuzzy c-means clustering of compositional data. J. Classif. **29**, 144–169 (2012)

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R.: Modeling and Analysis of Compositional Data. Wiley, New York (2015)

Peters, B.C., Walker, H.F.: An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. SIAM J. Appl. Math. **35**(2), 362–378 (1978)

R Development Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rayens, W.S., Srinivasan, C.: Dependence properties of generalized Liouville distributions on the simplex. J. Am. Stat. Assoc. **89**(428), 1465–1470 (1994)

Redner, R.: Note on the consistency of the maximum likelihood estimate for non-identifiable distributions. Ann. Stat. **9**, 225–228 (1981)

Rothenberg, T.: Identification in parametric models. Econometrica **39**(3), 577–591 (1971)

Smith, B., Rayens, W.: Conditional generalized Liouville distributions on the simplex. Statistics **36**(2), 185–194 (2002)

Wald, A.: Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. **20**, 595–601 (1949)