SORT 40 (2) July-December 2016, 349-374

Log-ratio methods in mixture models for compositional data sets

M. Comas-Cufí, J.A. Martín-Fernández and G. Mateu-Figueras

Abstract

When traditional methods are applied to compositional data misleading and incoherent results could be obtained. Finite mixtures of multivariate distributions are becoming increasingly important nowadays. In this paper, traditional strategies to fit a mixture model into compositional data sets are revisited and the major difficulties are detailed. A new proposal using a mixture of distributions defined on orthonormal log-ratio coordinates is introduced. A real data set analysis is presented to illustrate and compare the different methodologies.

MSC: 62E99, 62G07, 62H30, 62H99.

Keywords: Compositional data, Finite Mixture, Log ratio, Model-based clustering, Normal distribution, Orthonormal coordinates, Simplex.

1. Introduction

A *finite mixture distribution* is a probability distribution with probability density function (pdf) given by the expression

$$\pi_1 f_1(\,\cdot\,;\boldsymbol{\theta}_1) + \cdots + \pi_k f_k(\,\cdot\,;\boldsymbol{\theta}_k), \tag{1}$$

where $f_1,\ldots,f_k$ are pdf's of distributions with parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ respectively, and $\pi_1,\ldots,\pi_k$ are positive numbers with $\sum_{i=1}^{k}\pi_i = 1$ (McLachlan and Peel, 2000). The pdfs $f_1,\ldots,f_k$ are typically called *mixtures components*. In this paper we assume the most common case where all the mixture components, f_i, in a mixture belong to a unique family (Gaussian, skew-normal, etc) with pdf, f, and parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ belonging to a unique set $\boldsymbol{\Theta}$.

According to Scott and Symons (1971) and McLachlan and Peel (2000), finite mixture models provide reasonable results in several multivariate techniques, for instance,

Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona. gloria.mateu@udg.edu

Received: February 2016
Accepted: October 2016

discriminant analysis, density estimation and model-based clustering (Banfield and Raftery, 1993), even for high-dimensional data (Bouveyron and Brunet-Saumard, 2014). The Gaussian mixture is the most common model thanks to its theoretical and computational simplicity (McLachlan and Peel, 2000). However, because of its simplicity, Gaussian mixtures have some significant limitations which triggered the proposal of alternative models. For example, Student t mixtures were introduced to fit distributions with heavier tails (Andrews and McNicholas, 2012, Lee and McLachlan, 2014, Lin, 2010); and skew-normal and skew-t (Azzalini and Capitanio, 1999, 2003) mixtures were proposed to fit asymmetrical distributions (Lee and McLachlan, 2011). Moreover, Browne and McNicholas (2013) introduced the Generalized Hyperbolic mixture, a more general mixture model which includes, either asymptotically or explicitly, different types of well-known families of mixture models. A crucial point to note is that all these mixture models were designed for data in real space. For data in a different sample space, there is a general agreement that other distributions should be used. For example, Bickel and Scheffer (2004) used multinomial mixture distributions for discrete data in text classification, and Bouguila (2011) proposed other extensions of multinomial mixture distributions for count data. Another example is circular data, whose sample space is the sphere. Banerjee et al. (2005) and Mardia et al. (2007) proposed mixtures of Von Mises probability distributions, defined for random vectors in the sphere.

Finite mixture modelling for compositional data (CoDa) also needs its own probability distributions because the CoDa sample space, the simplex $\mathscr{S}^D$, has a particular algebraic-geometric structure, different from the one in real space (Pawlowsky-Glahn and Egozcue, 2001). CoDa, also called D-part compositions, are vectors $\mathbf{x} = (x_1, ..., x_D)$ with all its parts strictly positive and carrying only relative information. A D-part composition is usually restricted to sum to a fixed constant κ, i.e.

$$\sum_{i=1}^{D} x_i = \kappa. \tag{2}$$

As a convention, it is usual to assume $\kappa = 1$ for proportions and $\kappa = 100$ for percentages. Because the value of κ is irrelevant, in this paper we will assume that $\kappa = 100$ for simplicity. Typical examples of CoDa are frequent in economics (income and expenditure distributions), medicine (body composition: fat, bone, muscle), the food industry (food composition: fat, sugar, etc), geochemistry and chemometrics (chemical composition), ecology (abundance of different species), sociology (time-use surveys), and genetics (genotype frequency). When a problem is compositional, one assumes that the absolute value of each part is irrelevant and the interest is focused on the ratios of the parts. Following this idea, Aitchison (1986) introduced the log-ratio methodology to deal with compositional data. According to this methodology, the compositions are expressed in terms of log-ratio coordinates and traditional techniques are applied to them. This log-ratio methodology is coherent with the algebraic-geometric structure of the simplex

introduced later by Pawlowsky-Glahn and Egozcue (2001). In the literature we find a large number of papers where a specific methodology for CoDa is developed following the log-ratio approach (e.g., Martín-Fernández et al., 2015, Vives-Mestres et al., 2014, Palarea-Albaladejo et al., 2012).

As in many other statistical methods, log-ratio methodology requires complete data sets. When measuring concentrations, some elements are often not present in sufficient concentrations and measuring instruments report them as values below detection limits. In the literature this issue is also known as the rounded zero problem. The data matrix is completed by using imputation strategies, replacing non-detected values with reasonable estimates, and by allowing the computation of log-ratios for applying to any multivariate data analysis. The interested reader can refer to Palarea-Albaladejo et al. (2014), whose work encompasses the recent advances in this area.

Another approach to the zero problem consists in transforming the data from the simplex into the real space using a transformation defined on the zero, for example the hyperspherical transformation (Neocleous et al., 2011, Wang et al., 2007). Scealy et al. (2015) recommend the square root transformation because it handles zero components. While these possibilities can exhibit good results, in practice they lack of geometric structure (see discussion in Aitchison, 1982). In this work we consider the log-ratio methodology, which can be seen as a transformation but it also provides a geometry to the simplex with its own operations.

It is difficult to find in the literature finite mixture models for CoDa that consider distributions restricted to the simplex. The exception are a few studies (e.g., Albert and Gupta, 1982, Bouguila et al., 2004, Calif et al., 2011) where finite mixture models using Dirichlet distributions, a traditional probability distribution in the simplex, are used. Nevertheless, it is more frequent to ignore the compositional nature of the CoDa data and to use mixtures models of distributions on real space (e.g., Papageorgiou et al., 2001). Recently, in practical works, the log-ratio methodology had been considered to fit a mixture model (e.g., Ferrer-Rosell et al., in press) without theoretical and methodological considerations. As a consequence, there is a methodological gap in the analysis of CoDa where the latest advances in log-ratio methods can contribute to mixture modelling. In the present work, we introduce a new technique to model CoDa using mixtures of distributions well-defined on the simplex using orthonormal log-ratio coordinates and consequently coherent with its algebraic-geometric structure. In particular we use the normal and the skew-normal distributions on the simplex (Mateu-Figueras and Pawlowsky-Glahn, 2007, Mateu-Figueras et al., 2013).

This paper is organized as follows: in Section 2 a brief introduction of CoDa analysis is provided. Section 3 describes the pros and cons of each of the traditional mixture models when applied to CoDa. Section 4 is devoted to introducing log-ratio mixture models and two real data sets are analysed in Sections 5 and 6 to compare the traditional and log-ratio approaches. Finally, Section 7 contains conclusions and final remarks. The programming of the data analyses discussed in this work has been conducted using the open-source R statistical environment (R Core Team, 2014). Computer rou-

tines implementing the methods can be obtained from the R packages `Mclust`, `Rmixmod`, `EMMIXuskew` and also from the website www.compositionaldata.com. As an accompaniment to this article, the data and the programs used to fit the mixtures in Sections 5 and 6 are provided as supplementary material.

2. Compositional data analysis

Aitchison (1986) stated that there are two basic operations in the simplex $\mathscr{S}^D$: *perturbation* ($\oplus$) and *powering* ($\odot$). *Perturbation* is defined between two compositions $\mathbf{x}$ and $\mathbf{y}$, and *powering* is defined between a composition $\mathbf{x}$ and a scalar value α as:

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, \ldots, x_D y_D), \qquad \alpha \odot \mathbf{x} = C({x_1}^{\alpha}, \ldots, {x_D}^{\alpha}), \tag{3}$$

where $C(\mathbf{x}) = \frac{\kappa}{\sum x_k}(x_1, \ldots, x_D)$ is the closure operation for rescaling a vector.

These operations respectively play analogous roles to translation and scalar multiplication in $\mathbb{R}^D$, and provide a vector space structure of dimension $D-1$ to the simplex. Pawlowsky-Glahn and Egozcue (2001) stated that the inner product

$$< \mathbf{x}, \mathbf{y} >_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \tag{4}$$

provides $\mathscr{S}^D$ with the structure of an Euclidean space of dimension $D-1$. Note that a norm and a distance can be derived from the inner product given by Equation 4. This Euclidean space structure allows us to establish the principle of working on coordinates (Mateu-Figueras et al., 2011). The idea is to express compositions in terms of their coordinates with respect to an orthonormal basis on $\mathscr{S}^D$ and apply traditional statistical methods to these coordinates. These coordinates are formed by log-ratios, therefore we use the log-ratio methodology mentioned above. Once an orthonormal basis $\mathscr{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ is fixed, any D-part composition $\mathbf{x}$ can be expressed as the linear combination

$$\mathbf{x} = (h_1 \odot \mathbf{v}_1) \oplus \cdots \oplus (h_{D-1} \odot \mathbf{v}_{D-1}).$$

The elements of vector $\mathbf{h}_{\mathscr{B}}(\mathbf{x}) = (h_1, \ldots, h_{D-1})$ are the orthonormal log-ratio coordinates of composition $\mathbf{x}$ with respect to the basis $\mathscr{B}$. Egozcue et al. (2003) introduced an example of these coordinates where

$$h_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \; i = 1, \ldots, D-1, \tag{5}$$

whose corresponding basis is $\mathscr{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ with

$$\mathbf{v}_i = C\Big(\underbrace{e^{1/\sqrt{i(i+1)}}, \ldots, e^{1/\sqrt{i(i+1)}}}_{i}, 1/e^{\sqrt{i/(i+1)}}, \underbrace{1, \ldots, 1}_{D-(i+1)}\Big).$$

In this paper we use the coordinates in Equation 5 but any other orthonormal basis can also be considered. Determining which basis or coordinates are the most appropriate to solve a specific problem, is not straightforward. Nevertheless, the sequential binary partition introduced by Egozcue and Pawlowsky (2005) is a very useful tool to construct a particular basis to increase the interpretability of the corresponding coordinates.

One can define a pdf on the simplex by a pdf over the vector of orthonormal log-ratio coordinates. Indeed, let $f^*(\cdot\,;\boldsymbol{\theta}) : \mathbb{R}^{D-1} \to \mathbb{R}^+$ be a pdf defined on real space with parameters $\boldsymbol{\theta}$. Then, $f_{\mathscr{B}}(\mathbf{x}\,;\boldsymbol{\theta}) = f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x})\,;\boldsymbol{\theta})$ defines a pdf on the simplex, $f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}) : \mathscr{S}^D \to \mathbb{R}^+$, with respect to the Aitchison measure on $\mathscr{S}^D$. For example, fixing an orthonormal basis $\mathscr{B}$, the log-ratio normal distribution with parameters μ and $\boldsymbol{\Sigma}$ is defined as

$$f_{\mathscr{B}}(\mathbf{x}\,;\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{(D-1)/2}|\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{h}_{\mathscr{B}}(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{h}_{\mathscr{B}}(\mathbf{x})-\boldsymbol{\mu})}. \tag{6}$$

Note that it is a density on the simplex with respect to the Aitchison measure. The Aitchison measure, $d\lambda_a$, is a natural measure on $\mathscr{S}^D$, compatible with its Euclidean vector space structure (see Mateu-Figueras et al., 2013, for an in-depth discussion). This measure is absolutely continuous with respect to the Lebesgue measure on real space, $d\lambda$, and the relationship between them is $|d\lambda_a/d\lambda| = (\sqrt{D}x_1x_2\cdots x_D)^{-1}$.

Figure 1 (left) shows the contour lines of three normal distributions in the simplex $\mathscr{S}^3$. Note that the distribution in the centre of the ternary diagram is similar to the cir-

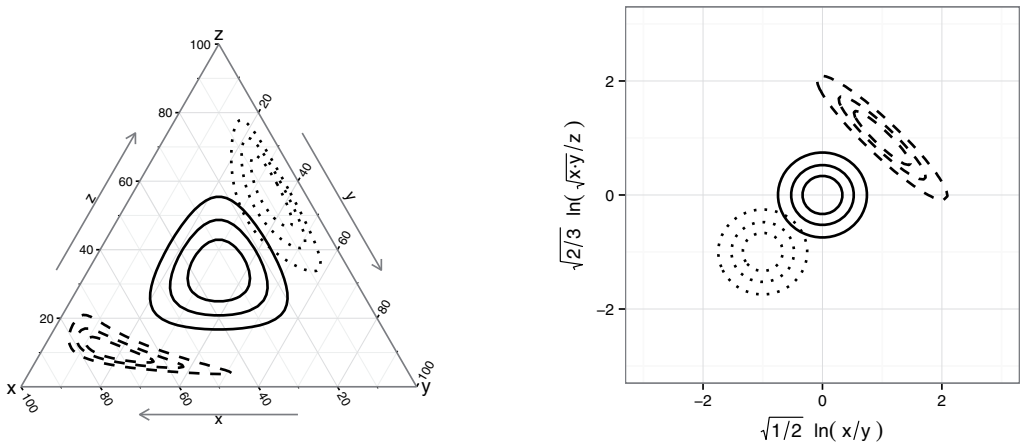

Figure 1: *Contour lines of typical log-ratio normal distribution on the simplex: (left) in the ternary diagram; (right) in log-ratio coordinates.*

cular contour lines in real space. However, note that, the farther the distribution from the centre is, the more different the contours from the traditional Gaussian shape are. These shapes are frequent in real data sets from industrial and scientific applications (Buccianti, 2011, Vives-Mestres et al., 2014). When these distributions are plotted using their orthonormal log-ratio coordinates (Figure 1 (right)) the traditional Gaussian contour lines are obtained. This idea can be applied by using other distributions on real space as, for example, the skew-normal (Mateu-Figueras and Pawlowsky-Glahn, 2007).

The well-known additive log-ratio vector (Aitchison, 1986) can be interpreted as the coordinates of a composition with respect to a non-orthogonal basis. Although the expression of the corresponding pdf is similar to Equation 6, the distances are not preserved among the additive log-ratio components and the principle of working on coordinates cannot always be applied (Mateu-Figueras et al., 2011). The equally well-known centred log-ratio vector (Aitchison, 1986) can be interpreted as the coordinates of a composition with respect to a generating system, not a basis. Despite the distances being preserved in this case, we do not recommend its use in a mixture model context because the fitted densities will be degenerate (Mateu-Figueras et al., 2011).

3. Modelling compositional data using traditional mixtures

When the goal is to fit a finite mixture model, the researcher can encounter different difficulties such as unbounded likelihood function, different local maximum, etc. The reader interested in knowing how to deal with these difficulties can consult McLachlan and Peel (2000) for an in-depth exposition. In this article we will indicate all the decisions taken in the process of fitting the finite mixtures.

3.1. Finite mixtures using traditional distributions defined on the real space

This approach assumes that $\mathscr{S}^D$ is a subset of $\mathbb{R}^D$ and its particular Euclidean space structure described in Section 2 is ignored. It is assumed that compositions are generated from a finite mixture distributions with pdf given by Equation 1 where $f(\,\cdot\,;\boldsymbol{\theta}_i) : \mathbb{R}^D \to \mathbb{R}^+$ is a pdf defined on the real space and with respect to the Lebesgue measure (e.g., a multivariate normal distribution or a t-student distribution). The main reason for using this approach is the simplicity of working without having to consider any restriction. However, this strategy exhibits some significant limitations and misleading results could be obtained.

When one uses traditional distributions defined on the real space, the mixture pdf is strictly positive in all the space, giving positive probability to impossible events. For example, the *impossible* event of having the i-th part negative has positive probability, i.e $P(\{\mathbf{x} \in \mathscr{S}^D | x_i < 0\}) > 0$. This difficulty is similar to the traditional confidence interval of a very small or very large proportion, i.e. it may provide lower or upper limit respectively beyond the restricted space.

Table 1: *CoDa set with three parts (a,b,c) from 20 compositions. (h_1, h_2) are its log-ratio coordinates. Two categorical covariates were considered: site and condition.*

a	**b**	**c**	h_1	h_2	**site**	**condition**
54.73	34.37	10.90	0.329	1.128	S1	C1
64.75	25.08	10.18	0.671	1.123	S1	C1
64.18	24.91	10.91	0.669	1.060	S1	C1
83.53	11.85	4.61	1.381	1.568	S1	C1
62.72	28.15	9.13	0.566	1.246	S1	C1
62.10	27.73	10.17	0.570	1.148	S1	C1
69.46	22.53	8.00	0.796	1.305	S1	C1
68.25	26.43	5.32	0.671	1.696	S1	C1
66.88	26.16	6.96	0.664	1.464	S1	C1
61.62	28.38	9.99	0.548	1.169	S1	C1
31.65	55.23	13.12	−0.394	0.946	S2	C1
24.32	61.47	14.21	−0.656	0.817	S2	C1
24.47	59.49	16.04	−0.628	0.708	S2	C1
18.75	68.00	13.25	−0.911	0.809	S2	C1
15.72	72.96	11.32	−1.085	0.895	S2	C1
18.83	32.85	48.32	−0.394	−0.542	S2	C2
12.11	30.61	57.27	−0.656	−0.890	S2	C2
10.75	26.14	63.10	−0.628	−1.082	S2	C2
10.31	37.38	52.31	−0.911	−0.800	S2	C2
8.15	37.81	54.05	−1.085	−0.918	S2	C2

In addition, this approach defined on the real space also ignores the constant sum constraint. Therefore, a further limitation is the collinearity that appears between parts after restricting the parts to sum a constant (Equation 2). This collinearity implies that the covariance matrix is singular, and therefore some methods can not be directly applied. Frequently, mixture models are estimated using the Expectation–Maximization (EM) algorithm (Dempster et al., 1977). In the E-step of the EM-algorithm a pdf computed from the sample is evaluated. Because most pdf depend on the inverse of the covariance matrix (e.g., multivariate normal and skew-normal), the common solution consists of removing one part of the composition for the rest of the analysis (e.g., Papageorgiou et al., 2001). However, this strategy may produce misleading results. For example, let $\mathbf{X}$ be the CoDa set recorded in Table 1. It is a simulated 3-part compositional data set representing proportions of 3 different elements, denoted a, b and c. Assume that the compositions come from two different locations, S_1 and S_2; and that they were collected under two possible weather conditions, C_1 and C_2. In addition, assume that it is well known that these weather conditions only affect part c: in condition C_1 the level of element c is lower than in condition C_2 (for example, element c is water and condition C_1 is a sunny day while condition C_2 is a rainy day). In this way, the compositions from row numbers 16 to 20 (Table 1) are the perturbed corresponding counterparts of

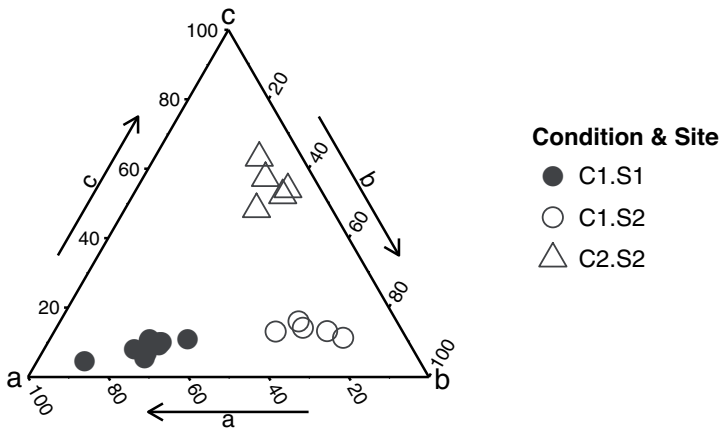

Figure 2: *CoDa set* $\mathbf{X}$ *in the ternary diagram. Filled and empty symbols are respectively used for data from location* S_1 *and* S_2*. Circles and triangles respectively correspond to condition* C_1 *and* C_2*.*

compositions from row numbers 11 to 15 after the perturbation $(1,1,r)$, where r is a random number depending on condition C_2. In this example we have modelled r as a lognormal random variable with parameters $\mu = 2$ and $\sigma = 0.25$. We have considered that condition C_1 and C_2 were an effect of the component c regardless of the magnitude of components a and b. Therefore, the effect of condition C_1 and C_2 could be modelled by means of a perturbation (Equation 3), which is a movement in the simplex with the Aitchison geometry.

The ternary diagram in Figure 2 shows that $\mathbf{X}$ is formed by three groups: the first group consists of the observations collected in site S_1 (filled circles), all of them collected under condition C_1; the second group with observations collected in site S_2 under condition C_1 (empty circles) and the third group with observations collected in site S_2 under condition C_2 (empty triangles). Suppose that an analyst, who is interested in fitting a traditional mixture model to $\mathbf{X}$, is not informed about the two different weather conditions and he or she only knows the information about the location. Because of the collinearity he/she decides to eliminate part c for the rest of the analysis. After eliminating part c, the researcher is working with the data set represented in Figure 3. This plot suggests that the analyst might conclude that $\mathbf{X}$ is formed by three mixture components as a result of the information collected in only the first two elements. This is a misleading conclusion because, by construction, we know that exclusively attending to the raw information provided by the first two elements the CoDa set $\mathbf{X}$ is formed by only two groups (one group for each location). But, when we work with proportions (a,b,c), despite part c having been eliminated, its effect (weather condition) is still present and interpretations about the nature of the groups based only on parts (a,b) may be misleading. An interested reader could find other examples about the misleadings conclusions and problems resulting from applying standard analysis to compositional data in Aitchison (1999, 2002).

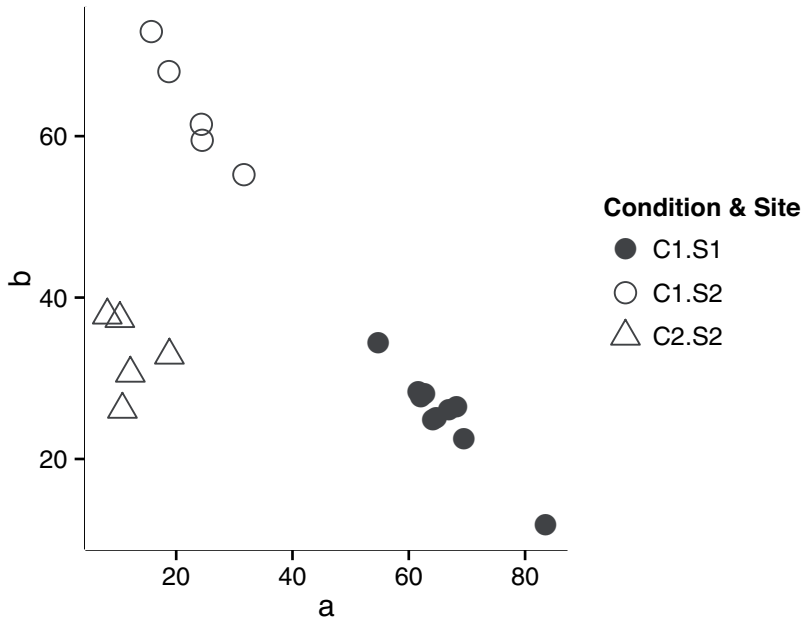

Figure 3: *Scatterplot of parts* (a,b) *of CoDa set* $\mathbf{X}$*. Filled and empty symbols are respectively used for data from location* S_1 *and* S_2*. Circles and triangles respectively correspond to condition* C_1 *and* C_2*.*

3.2. Finite mixtures using traditional distributions defined on the simplex

A finite mixture of distributions defined on the simplex is a probability distribution with pdf given by Equation 1 where $f(\,\cdot\,;\boldsymbol{\theta}) : \mathscr{S}^D \to \mathbb{R}^+$, is a pdf defined on the simplex. The Dirichlet distribution has been traditionally used as the probability distribution on $\mathscr{S}^D$. It can be obtained by the projection on the simplex of a random vector formed by independent and equally scaled gamma distributed parts. Despite its simplicity and its good mathematical properties, it has a very strong independence structure (Aitchison, 1986). In particular, any ratio x_i/x_j of two parts have to be independent from another ratio x_k/x_m formed from other two parts. In practice, such an independence structure cannot be assumed for most real data sets and consequently it heavily restricts the Dirichlet potential modelling application (Aitchison, 1986). To solve this difficulty, many generalizations of the Dirichlet distribution with less independence structure have been proposed: the Connor and Mosimann's distribution (Connor and Mosimann, 1969), the scaled Dirichlet distribution (Aitchison, 1986). In addition, Rayens and Srinivasan (1994) extend the Liouville distribution further to the generalized Liouville family. Later Smith and Rayens (2002), due to the limited applicability of the Liouville family of distributions, propose a generalization called Conditional Liouville distribution. Ongaro and Migliorati (2013) present the Flexible distribution, a generalization of the Dirichlet that exhibits greather flexibility in terms of dependence/independence structure and shape of the density. Finally, Monti et al. (2011) introduce the shifted-scaled Dirichlet distribution. This

generalized distribution is defined by adding the perturbation and powering operations (Equation 3) to the standard Dirichlet distribution. Unfortunately, all of these attempts have had limited success in fitting the general dependence structure of CoDa. Note that all these distributions are usually expressed through their density function with respect to the Lebesgue measure on $\mathscr{S}^D$ but the density with respect to the Aitchison measure could be easily obtained using the relationship between them (see Monti et al. (2011) for a detailed analysis of the implications of changing the measure).

In the literature different methods are found to estimate the parameters of a Dirichlet distribution. As it is an exponential family, the log-likelihood function is globally concave and a global optimum can be obtained. However, there is no closed form solution for the ML equations and numerical methods must be employed. According to Ng et al. (2011), the MLE via Newton-Raphson algorithm converges to the global optimum. Narayanan (1991) provides a Fortran subroutine with three different possibilities to estimate the initial parameter required. We can also obtain MLE estimates via the EM gradient methods (Ng et al., 2011). Recently the performance of different algorithms and starting value strategies to obtain the MLE of the Dirichlet parameters have been compared by Giordan and Wehrens (2015) using high-dimensional data. Nevertheless, the main problem is that final estimates can be outside the correct range for the parameters. Also, a large amount of iterations could be required to reach convergence. In practice, given a CoDa set, there is no straightforward method to fit a Dirichlet mixture or any of its generalizations. However, to obtain an approximation of the MLE estimator of a Dirichlet mixture, it is possible to apply the classification EM-algorithm (Celeux and Govaert, 1992) using any of the mentioned approaches to fit a Dirichlet model (see example in Section 5).

4. Modelling compositional data using a mixture of log-ratio distributions

To model CoDa using a finite mixture of log-ratio distributions, we consider

$$\pi_1 f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}_1) + \cdots + \pi_k f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}_k) \tag{7}$$

where $f_{\mathscr{B}}(\mathbf{x};\boldsymbol{\theta}_i)$ are pdf's defined on the simplex with parameters $\boldsymbol{\theta}_i$, that is, they are densities defined considering the particular algebraic-geometric structure of the simplex defined in Section 2 and consequently are expressed with respect to the Aitchison measure. As indicated before and according to the principle of working on coordinates, we have

$$f_{\mathscr{B}}(\mathbf{x};\boldsymbol{\theta}) = f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x});\boldsymbol{\theta})$$

where $f^*(\,\cdot\,;\boldsymbol{\theta})$ are pdf on $\mathbb{R}^{D-1}$ for the orthonormal log-ratio coordinates vectors $\mathbf{h}_{\mathscr{B}}(\mathbf{x})$. Let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ be a CoDa set. Thus fitting the parameters $\pi_1,\ldots,\pi_k$ and $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ of Equation 7 using maximum likelihood estimators is equivalent to fitting the parameters in

$$\pi_1 f^*(\,\cdot\,;\boldsymbol{\theta}_1)+\cdots+\pi_k f^*(\,\cdot\,;\boldsymbol{\theta}_k) \tag{8}$$

using the data set $\mathbf{X}^{\mathrm{T}}=\{\mathbf{h}_{\mathscr{B}}(\mathbf{x}_1),\ldots,\mathbf{h}_{\mathscr{B}}(\mathbf{x}_n)\}$, that is, the log-ratio coordinates of the data set with respect to a selected orthonormal basis $\mathscr{B}$.

Indeed, the likelihood function evaluated for the CoDa set $\mathbf{X}$ is

$$\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_{\mathscr{B}}(\mathbf{x}_i;\boldsymbol{\theta}_j)=\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x}_i);\boldsymbol{\theta}_j). \tag{9}$$

Because the likelihood functions are the same, the maximum likelihood estimators $\hat{\pi}_1,\ldots,\hat{\pi}_k,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_k$ are also the same

$$\left(\hat{\pi}_1,\ldots,\hat{\pi}_k,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_k\right)=\underset{\pi_1,\ldots,\pi_k,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\arg\max}\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_{\mathscr{B}}(\mathbf{x}_i;\boldsymbol{\theta}_j)= \tag{10}$$

$$=\underset{\pi_1,\ldots,\pi_k,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\arg\max}\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x}_i);\boldsymbol{\theta}_j). \tag{11}$$

Following this approach, we cannot obtain the misleading results shown in Section 3.1.. Taking the example from Section 3.1, we were interested in fitting a mixture to a sample $\mathbf{X}$ formed by parts a, b and c (Table 1). Instead of eliminating one part, now the analyst decides to express parts a, b and c in log-ratio coordinates. Before starting the analysis, a basis $\mathscr{B}$ of $\mathscr{S}^3$ is selected, for example

$$\mathscr{B}=\left\{C\left(e^{1/\sqrt{2}},1/e^{\sqrt{1/2}},1\right),C\left(e^{1/\sqrt{6}},e^{1/\sqrt{6}},1/e^{\sqrt{2/3}}\right)\right\}, \tag{12}$$

and the compositions of $\mathbf{X}$ are expressed in terms of their coordinates $\mathbf{X}^{\mathrm{T}}$ ($h_1=\sqrt{1/2}\ \ln(a/b)$ and $h_2=\sqrt{2/3}\ln(\sqrt{ab}/c)$) (see Table 1). Figure 4 shows the plot of these coordinates where the different effect of the location (parts a and b) and the weather conditions (part c) are highlighted. Note that the compositions from S_2 under condition C_1 take the same value in the first coordinate as their counterparts under condition C_2.

In this case the interpretations based only in terms of parts a and b will not be misleading. In fact, if the analyst also decides to remove part c, a basis $\mathscr{B}'$ of $\mathscr{S}^2$ is selected as:

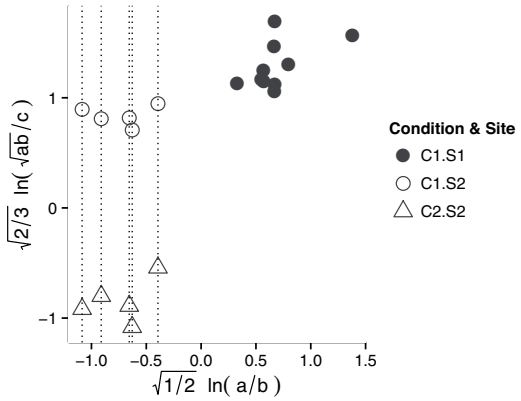

Figure 4: *Scatterplot of log-ratio coordinates for the CoDa set* **X**. *Filled and empty symbols are respectively used for data from location S_1 and S_2. Circles and triangles respectively correspond to condition C_1 and C_2.*

$$\mathscr{B}' = \left\{ C\left(e^{1/\sqrt{2}}, 1/e^{\sqrt{1/2}}\right) \right\}.$$

In this way, the corresponding coordinate h_1 is the same as before. Figure 5 shows the histograms of coordinate h_1 separated by weather conditions in two stratas. Note that, regardless of the condition, all the data collected in S_2 take the same value, forming one cluster (between -1 and 0). On the other hand, the compositions collected in S_1 are close to one.

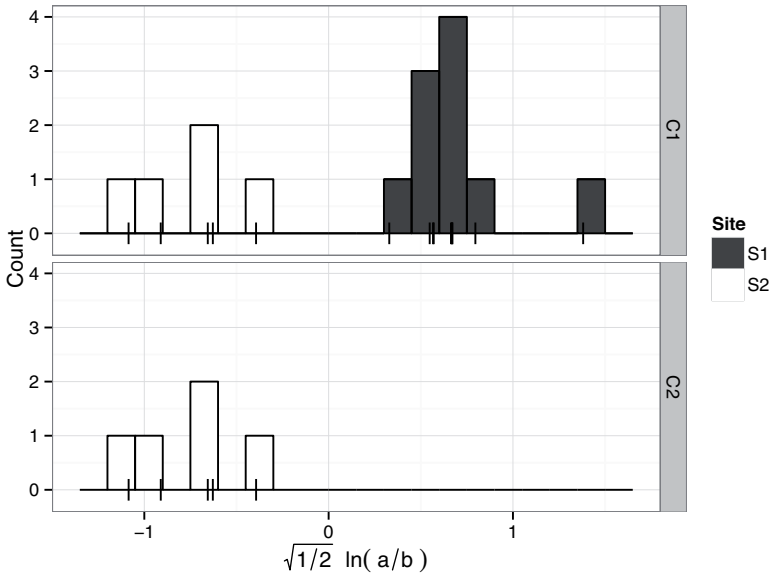

Figure 5: *Histograms of first log-ratio coordinate for CoDa set* **X**. *Two stratas correspond to weather conditions.*

In Equations 9 and 10, we fit the mixture using the coordinates $\mathbf{h}_{\mathscr{B}}(\mathbf{x})$ with respect to a specific basis $\mathscr{B}$ but any other orthonormal basis could have been chosen as well. Thus, in any compositional analysis involving coordinates, it is important to check the invariance of the results under changes of basis. When fitting a mixture of log-ratio distributions, it is enough to check that the family of distributions used to fit the mixture is basis invariant, that is, it satisfies the following definition.

Definition 1 *Let $\mathscr{B}_1$ and $\mathscr{B}_2$ be two basis on $\mathscr{S}^D$. Let $\boldsymbol{\Theta}$ be a parameter space for a probability density function $f^* : \mathbb{R}^{D-1} \to \mathbb{R}^+$. A probability density function f^* is* basis invariant *if for any two different basis $\mathscr{B}_1$, $\mathscr{B}_2$, for any parameters $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, there are parameters $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ such that*

$$f^*(\boldsymbol{h}_{\mathscr{B}_1}(\boldsymbol{x});\boldsymbol{\theta}_1) = f^*(\boldsymbol{h}_{\mathscr{B}_2}(\boldsymbol{x});\boldsymbol{\theta}_2).$$

Most common distributions are basis invariant when we do not restrict the parameters. For example, the log-ratio normal distribution (Equation 6) is formulated in terms of Mahalanobis distance and of covariance matrix determinant, that are both invariant elements under change of basis (Barceló-Vidal et al., 1999). Moreover, using the linear transformation property (Azzalini and Capitanio, 1999), it can easily be proved that the multivariate log-ratio skew-normal distribution is also invariant under change of basis.

5. A real data set: Forensic Glass

To illustrate and compare the different described approaches, we analysed the USA Forensic Science Service data set, also known as the Forensic Glass data set. This data is available from the UCI Machine Learning Repository (Bache and Lichman, 2013). The data set is composed of 214 fragments of glass samples where the percentages of eight chemical elements were measured. The fragments of glass were originally come from seven types of glass. In order to easily display the results using ternary diagrams and bivariant plots, we only consider three chemical elements: Calcium (Ca), Silica (Si) and Aluminium (Al). For simplicity, we only consider three types of glass (containers, vehicle headlamps and vehicle windows) but all types of glass could be considered and lead to similar conclusions. We call this data set the Reduced Forensic Glass data set (Table 2). Figure 6 shows this data set formed by 59 glass samples in the ternary diagram. We can see that the types of glass do not form well-separated groups and consequently there will be a weak relation between the components of the mixture and the types of glass. This was already observed by Venables and Ripley (2002) in a discriminant context.

We fit a mixture model using the normal distribution on real space, the Dirichlet distribution and the log-ratio normal and skew-normal distributions on the simplex. For all cases the index BIC indicates that $k = 3$ are the optimal number of components

Table 2: *Reduced Forensic Glass data set: parts (Ca, Si, Al) and its log-ratio coordinates. The categorical covariate (type) shows the provenance of glass.*

Ca	**Si**	**Al**	h_1	h_2	**type**
10.43	88.23	1.35	−1.510	2.541	Veh
10.12	88.26	1.63	−1.531	2.375	Veh
10.23	88.10	1.67	−1.523	2.359	Veh
10.31	88.06	1.63	−1.517	2.382	Veh
10.14	87.73	2.13	−1.526	2.155	Veh
11.60	87.39	1.01	−1.428	2.818	Veh
10.81	88.40	0.79	−1.486	2.994	Veh
10.12	88.40	1.48	−1.533	2.455	Veh
10.63	87.79	1.58	−1.493	2.418	Veh
10.36	88.12	1.52	−1.514	2.441	Veh
10.48	87.97	1.55	−1.504	2.429	Veh
11.77	87.53	0.71	−1.419	3.112	Veh
10.67	87.48	1.85	−1.488	2.290	Veh
10.69	87.33	1.98	−1.485	2.234	Veh
10.87	87.26	1.86	−1.473	2.292	Veh
10.80	88.29	0.91	−1.486	2.878	Veh
11.23	87.66	1.12	−1.453	2.721	Veh
7.41	88.18	4.42	−1.751	1.433	Con
11.92	85.88	2.20	−1.396	2.186	Con
13.29	84.89	1.82	−1.311	2.380	Con
13.41	84.78	1.80	−1.304	2.393	Con
13.26	84.84	1.90	−1.312	2.344	Con
11.84	86.03	2.13	−1.402	2.210	Con
13.15	84.81	2.04	−1.318	2.282	Con
14.23	83.94	1.84	−1.255	2.395	Con
8.65	87.57	3.78	−1.637	1.621	Con
8.59	87.66	3.74	−1.643	1.627	Con
14.51	83.87	1.63	−1.241	2.501	Con
11.54	85.88	2.58	−1.419	2.043	Con
13.08	85.17	1.75	−1.325	2.407	Con
6.78	90.96	2.26	−1.836	1.957	Head
7.31	89.89	2.80	−1.774	1.808	Head
10.71	87.80	1.49	−1.488	2.469	Head
11.89	85.60	2.51	−1.396	2.076	Head
10.72	87.65	1.63	−1.486	2.396	Head
10.38	87.48	2.14	−1.507	2.160	Head
10.38	86.80	2.82	−1.502	1.931	Head
10.60	86.13	3.27	−1.481	1.816	Head
10.21	87.40	2.39	−1.518	2.062	Head
10.17	87.47	2.36	−1.522	2.071	Head
10.65	86.20	3.15	−1.479	1.848	Head

Table 2 (cont.)

Ca	Si	Al	h_1	h_2	type
11.05	85.97	2.98	−1.451	1.908	Head
10.58	86.65	2.77	−1.487	1.953	Head
10.70	86.16	3.14	−1.475	1.853	Head
10.46	86.56	2.97	−1.494	1.891	Head
9.92	87.41	2.68	−1.539	1.957	Head
10.47	88.14	1.40	−1.506	2.513	Head
9.93	87.21	2.86	−1.536	1.903	Head
9.93	87.68	2.39	−1.540	2.052	Head
10.33	86.97	2.69	−1.506	1.968	Head
10.32	87.52	2.16	−1.512	2.150	Head
10.36	87.40	2.24	−1.508	2.121	Head
7.97	89.78	2.24	−1.712	2.025	Head
11.11	85.67	3.22	−1.444	1.845	Head
10.84	85.76	3.40	−1.463	1.791	Head
10.07	87.55	2.38	−1.529	2.061	Head
10.06	87.53	2.41	−1.530	2.050	Head
10.09	87.60	2.31	−1.528	2.086	Head
10.25	87.27	2.47	−1.514	2.036	Head

except for the Dirichlet distribution whose optimal value is for $k = 5$. For illustration purposes and in order to easily compare all described approaches, we will use $k = 3$ for all different cases. For each mixture approach, we fit the mixture 100 times using different starting points to avoid local maximums.

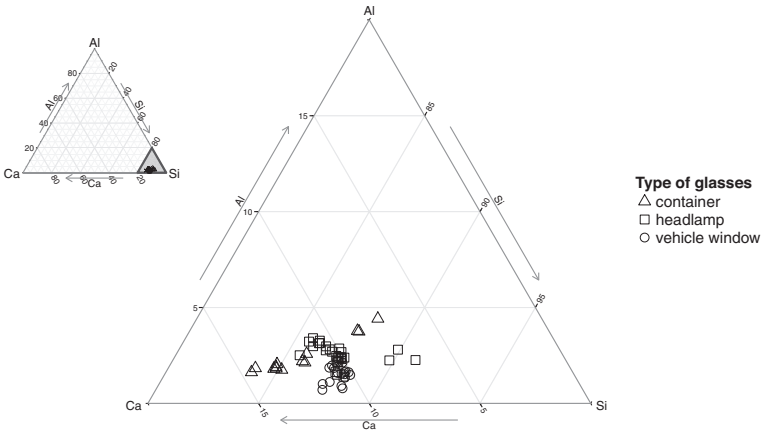

Figure 6: *Reduced Forensic Glass data set in ternary diagram: Calcium (Ca), Silica (Si) and Aluminium (Al) chemical elements. Three groups of glass: containers (circles), headlamps (triangles) and vehicle windows (squares). The large ternary diagram is a zoom of the shadow area seen in the smaller initial ternary diagram.*

Using the traditional approach introduced in Section 3.1 we fit a mixture of distributions on real space with three mixture components. In particular we choose a traditional Gaussian mixture. As mentioned, we need to eliminate one part to avoid the constant sum constraint. For example, when we removed the Calcium (Ca) part, the corresponding mixture model (BIC = -763.4) obtained is $\pi_1 f(\cdot\,;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+\pi_2 f(\cdot\,;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+\pi_3 f(\cdot\,;\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ with estimates

$$\hat{\pi}_1 = 0.12,\ \ \hat{\boldsymbol{\mu}}_1 = (88.76, 1.65),\ \ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.66 & 0.81 \\ 0.81 & 0.52 \end{pmatrix},$$

$$\hat{\pi}_2 = 0.38,\ \ \hat{\boldsymbol{\mu}}_2 = (85.85, 2.68),\ \ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.17 & 0.72 \\ 0.72 & 0.58 \end{pmatrix},$$

$$\hat{\pi}_3 = 0.5,\ \ \hat{\boldsymbol{\mu}}_3 = (87.67, 1.97) \text{ and } \ \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.16 & -0.18 \\ -0.18 & 0.27 \end{pmatrix}.$$

Figure 7 (top-left) shows the isodensity curves for the fitted mixture of Gaussian distributions. Figure 7 (top-right and bottom-left) also shows the isodensity curves of the finite mixture when the parts removed were Aluminium (Al) and Silica (Si), respectively. The dashed lines represent the limit of the simplex, i.e. the region were restrictions given by Equation 2 are held. In Figure 7 (bottom-right) the isodensity curves have been completed to be represented in the ternary diagram. Note that the distribution is giving positive probability to impossible regions.

Despite the fact that in Gaussian mixtures the maximum likelihood function is invariant whatever part is removed, we stated that in practice the numerical algorithm gets stuck in a local optimum. That is, the invariance of the results is not guaranteed, and different mixtures may be obtained depending on the part removed.

A Dirichlet probability distribution is specified by the parameters $\boldsymbol{\alpha} = \left(\alpha^1,\ldots,\alpha^D\right)$. Therefore, to fit a mixture of K Dirichlet distributions the parameters π_1, ..., π_K and $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K$ need to be estimated. To make this estimation we approximated the MLE estimator of a Dirichlet mixture using the EM-algorithm proposed by Celeux and Govaert (1992). The mixture of Dirichlet distributions obtained (BIC =-732.9) was $\pi_1 f(\cdot\,;\boldsymbol{\alpha}_1)+\pi_2 f(\cdot\,\boldsymbol{\alpha}_2)+\pi_3 f(\cdot\,;\boldsymbol{\alpha}_3)$ with estimates

$$\hat{\pi}_1 = 0.37, \qquad \hat{\boldsymbol{\alpha}}_1 = (281.2, 2343.1, 71.6),$$

$$\hat{\pi}_2 = 0.15, \qquad \hat{\boldsymbol{\alpha}}_2 = (272.9, 1777.2, 41.2),$$

$$\hat{\pi}_3 = 0.48 \text{ and } \quad \hat{\boldsymbol{\alpha}}_3 = (34.6, 304.3, 6.3).$$

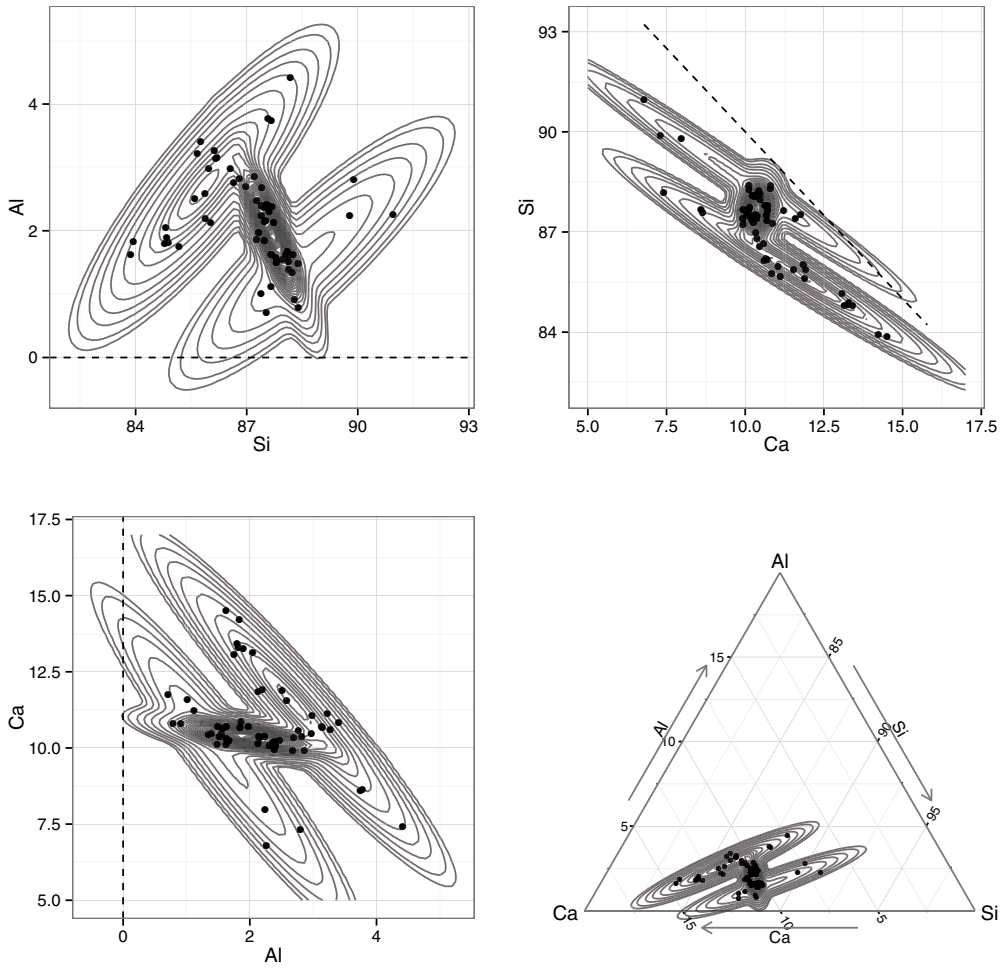

Figure 7: *Reduced Forensic Glass data set. On the top-left, top-right and bottom-left isodensity curves for mixtures of Gaussian distributions in R^2 after removing the Ca, the Al and the Si part respectively. On bottom-right the isodensity curves transformed into the simplex.*

Note that for $k = 3$ the Dirichlet BIC value is worse than the value for the normal distribution. Using the Dirichlet parameter estimates we can, respectively, obtain the centre of each mixture component in the simplex: $(10.43, 86.91, 2.66)$, $(13.05, 84.98, 1.97)$ and $(10.02, 88.15, 1.83)$, expressed in percentages.

Figure 8 shows how the Dirichlet mixture fits the data set. Due to the strong independence structure of the Dirichlet model (noted above in Section 3.2), the density can only take nearly elliptical shapes. Consequently, the mixture obtained cannot capture non-elliptical forms of variability.

Finally, we use the log-ratio approach introduced in Section 4. To fit a mixture of log-ratio distributions it is necessary first to express each composition with respect to a

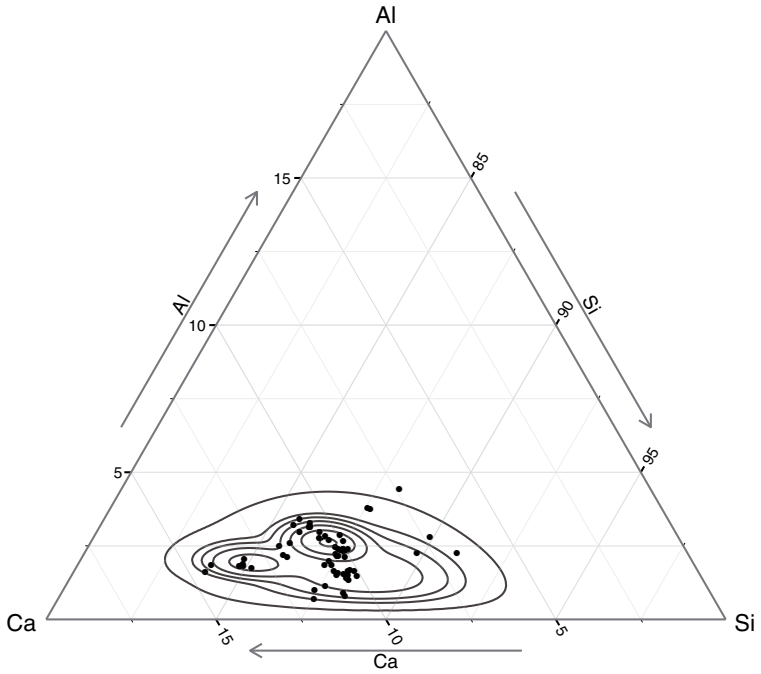

Figure 8: *Reduced Forensic Glass data set: classification given by a standard Dirichlet mixture model.*

basis of $\mathscr{S}^3$. Consider the same basis $\mathscr{B}$ defined in Equation 12. Table 2 contains the data set expressed in log-ratio coordinates with respect to basis $\mathscr{B}$, resulting in coordinates $h_1 = \sqrt{1/2}\ln(\mathrm{Ca}/\mathrm{Si})$ and $h_2 = \sqrt{2/3}\ln(\sqrt{\mathrm{Ca}\cdot\mathrm{Si}}/\mathrm{Al})$.

Fitting a Gaussian mixture to the log-ratio coordinates (BIC = -84.3) results in mixture model $\pi_1 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ with estimates

$$\hat{\pi}_1 = 0.59,\ \ \hat{\boldsymbol{\mu}}_1 = (-1.5, 2.31),\ \ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 8e-04 & 0.0059 \\ 0.0059 & 0.0949 \end{pmatrix},$$

$$\hat{\pi}_2 = 0.1,\ \ \hat{\boldsymbol{\mu}}_2 = (-1.73, 1.75),\ \ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.005 & -0.0059 \\ -0.0059 & 0.0422 \end{pmatrix},$$

$$\hat{\pi}_3 = 0.31,\ \ \hat{\boldsymbol{\mu}}_3 = (-1.39, 2.12) \text{ and } \ \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.0065 & 0.0186 \\ 0.0186 & 0.0581 \end{pmatrix}.$$

Note that the difference between the BIC value for the log-ratio normal distribution and the previous distributions seems to be unusually large. However, these values can not be directly comparable because the latter is calculated using log-ratio coordinates. In Figure 9 the isodensity curves of the log-ratio normal distribution are represented in

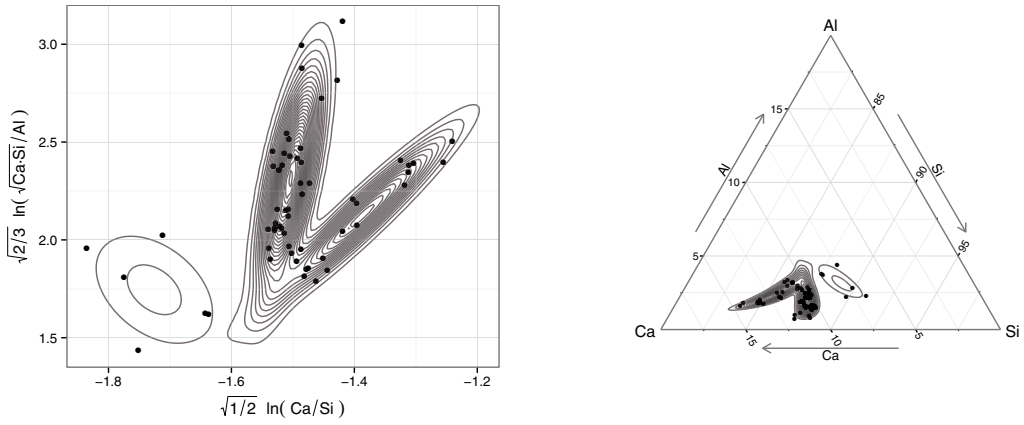

Figure 9: *Log-ratio Gaussian mixtures for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.*

the space of coordinates (left) and in the ternary diagram (right). Looking at the coordinate space, we see that this mixture can model elliptical forms of variability and consequently, on the simplex the estimated mixture is able to model those typical arc shaped forms (Figure 9 (right)). Because multivariate log-ratio normal is basis invariant (Section 4), working with another orthonormal log-ratio basis results in the same mixture as that represented in the ternary diagram (Figure 9 (right)). As noted above, there is low similarity between mixture components and types of glass. In this case the adjusted Rand index (Hubert and Arabie, 1985) is equal to 0.219.

Note that the parameters of the mixture are expressed with respect to coordinates h_1 and h_2. To better interpret the parameters of the mixture, we back-transformed the parameters $\boldsymbol{\mu}_i$ into the simplex: $(10.46, 87.75, 1.79)$, $(7.77, 89.13, 3.10)$ and $(12.02, 85.59, 2.39)$, into percentages. Note that only the centre of the first log-ratio normal mixture component is similar to the centre of the first Dirichlet mixture component. To better interpret the covariance parameter $\boldsymbol{\Sigma}_i$, Aitchison (1986) proposes using the variation matrix, that is, the variance of each log-ratio. In this case, the corresponding log-ratio variances are shown in Table 3.

The first mixture component is characterised by the highest relative variability of the ratio between the Calcium and Aluminium parts and lowest between the Calcium and

Table 3: *Forensic Glass data set: log-ratio variances for each mixture component fitted by a log-ratio Gaussian mixture.*

Mixture component	var(ln(Ca/Si))	var(ln(Ca/Al))	var(ln(Si/Al))
1	0.0016	0.1530	0.1324
2	0.0101	0.0556	0.0760
3	0.0131	0.1226	0.0582

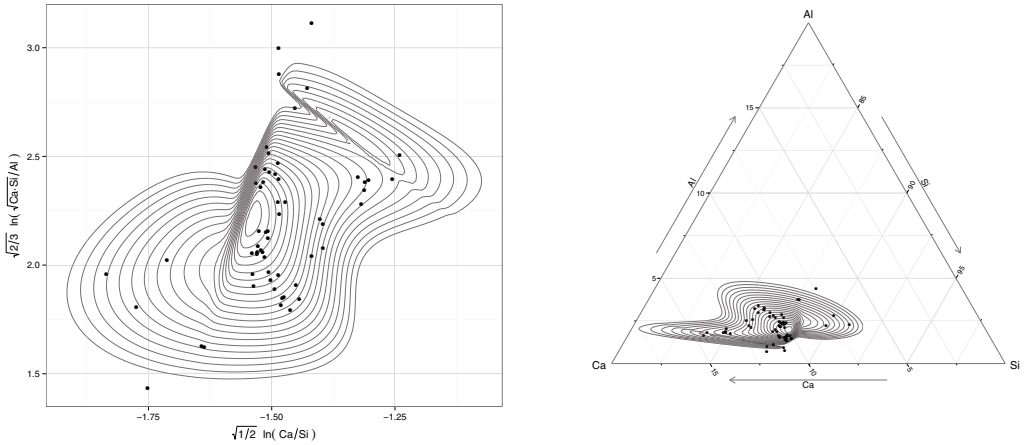

Figure 10: *Log-ratio skew normal mixture adjusted for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.*

Silica elements. Due to var(ln(Ca/Si)) being close to zero, the concentration of these elements are nearly proportional (Martín-Fernández et al., 2015). Note that this behaviour is common across the three mixture components. All the variances take small values for the second mixture component, while the third mixture component differs from the first due to the small value in the variance of ln(Si/Al).

Following an analogous approach, it is possible to fit other non-Gaussian models. For example, in Figure 10 the data set is modelled with a mixture of multivariate log-ratio skew-normal distributions using the package provided by Prates et al. (2013) (BIC = -62.3). The log-ratio skew-normal model extends the modelling possibilities because it contains the log-ratio normal model as a particular case. Nevertheless, the final model is more complex because a skew parameter is added for each density in the mixture. This complexity also contributes to the BIC value which is worse than the value for the log-ratio normal distribution. For the sake of brevity, we prefer not to give the estimated parameters here. The multivariate log-ratio skew-normal model is also basis invariant, thus working with another orthonormal log-ratio basis results in the same mixture as that represented in the ternary diagram (Figure 10 (right)). Although the adjusted Rand index increased slightly to 0.348, there is low similarity between mixture components and types of glass.

6. A second real data set: C-horizon of the Kola data set

To illustrate how to proceed when the number of parts is greater than three, we analysed a reduced data set of the C-horizon of the Kola data set (Reimann, Filzmoser). We selected a subsample formed by 69 observations belonging to three groups: Alkaline (7), Sediments (39) and Granite (23). For these samples we created the subcomposition

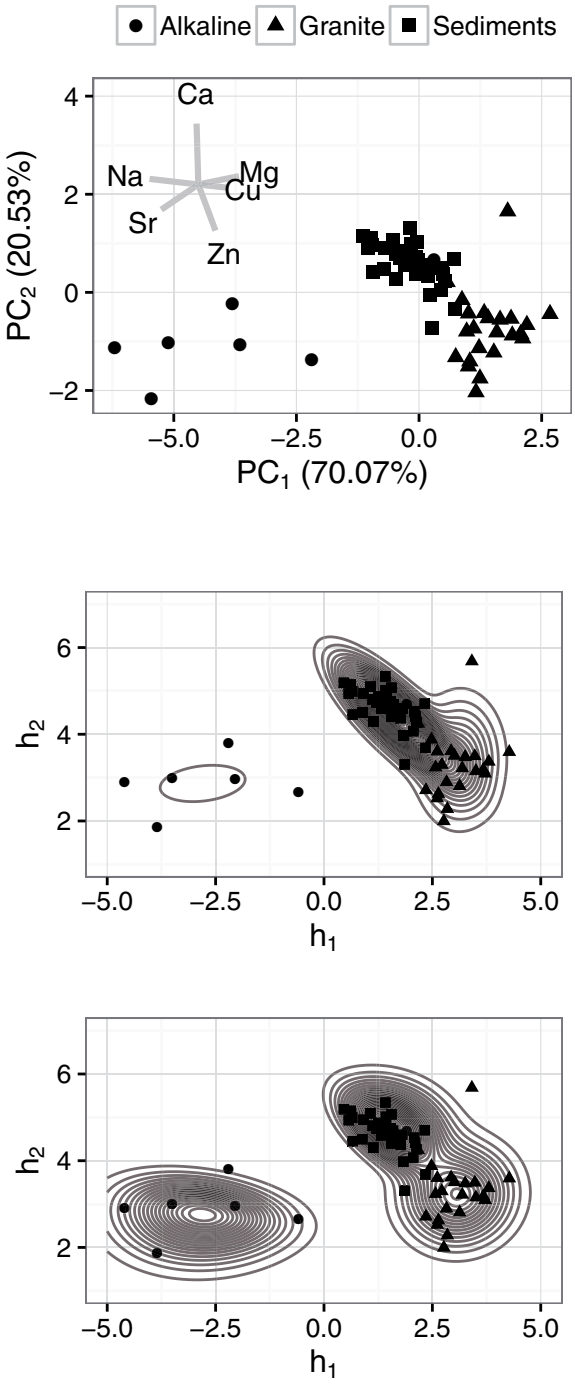

Figure 11: *Mixtures adjusted to the reduced C-horizon of Kola data set: (top) compositional biplot; (middle) marginal of the log-ratio Gaussian mixture for the two first coordinates: h_1 and h_2; (bottom) marginal of the log-ratio skew normal mixture for the two first coordinates.*

formed by the chemical elements: Calcium (Ca), Copper (Cu), Magnesium (Mg), Sodium (Na), Strontium (Sr) and Zinc (Zn).

Figure 11 (top) shows the compositional biplot, which consists of a principal component plot applied to the centred log-ratio coordinates. The two principal components explain a 90.6% variance, which is a high percentage of the total variance of the sample. The first principal axis (PC_1) is associated to the relative variation in parts Na and Sr as opposed to Mg and Cu. On the other hand, the axis of the PC_2 is associated to the relative variation of element Ca versus Zn. The group of Alkaline observations has a high concentration of elements Na and Sr with respect to the proportion in the groups Granite and Sediments that have a high concentration of Mg and Cu elements. The main differences between the groups Granite and Sediments is that the former has a higher proportion of the element Ca, whilst the latter has high concentration in the Zn part.

We fit a mixture model using the normal and the skew-normal distributions on log-ratio coordinates. For the sake of brevity, the estimated parameters are not provided. In both cases the BIC index indicates that $k = 3$ is the optimal number of components. To avoid local maximums we recalculated the parameters for each mixture until no improvement was obtained in the likelihood function during 100 simulations. To calculate the orthonormal log-ratio coordinates in this example we considered the orthonormal basis $\mathscr{B}$ formed by the directions of the principal components.

Figure 11 (middle) shows the marginal of the adjusted log-ratio normal mixture with respect to the first (h_1) and second (h_2) orthonormal log-ratio coordinates. For the log-ratio normal distributions the Rand index was 0.580, with 29 observations misclassified. In Figure 11 (bottom) the marginal (h_1, h_2) of the adjusted log-ratio skew normal mixture is shown. In this case the Rand index is better (0.760) and the misclassification rate is also improved because only 5 observations were misclassified.

7. Final remarks

Traditional distributions in finite mixtures for compositional data sets show significant difficulties. If densities for real data are used, probabilities of impossible events are obtained. Additionally, as a part of a composition is often removed to estimate the model, the results depend on that part. Dirichlet density and some generalizations on the simplex can not capture the variability of many compositional data sets due to their strong independence structure. The proposed log-ratio models are defined on the simplex using its particular algebraic-geometric structure. Consequently probabilities for impossible events are not obtained and there is no need to eliminate any part. The log-ratio normal model is a flexible model that can describe different forms of variability and dependence structures. It is a simple model and provides a rich enough parametric class of distributions on the appropriate sample space. Certainly, the model has the equivalent limitations as the traditional Gaussian mixtures in real space. Nevertheless, the proposed methodology allows different and alternative models. Indeed, any mixture model

defined on the real space can be considered to model data on the simplex space using the principle of working on coordinates. In this paper we have proposed a mixture of normal and skew-normal distributions to the log-ratio coordinates of a compositional sample. These two options extend the range of possibilities we have had up to now with the Dirichlet model or its generalizations. Interestingly, both proposed log-ratio models are invariant with respect to the orthonormal basis chosen to compute the log-ratios. The proposed log-ratio methodology could be extended by studying the possibilities of other known distributions on real space, like Student-t and skewed-t mixtures. Furthermore, in a non-parametric context, an analogy of these models with the P-spline methodology for CoDa should be explored Eilers et al. (2015).

Acknowledgments

This research was supported by the Ministerio de Economía y Competividad through the projects "METRICS" and "CoDa-RETOS" (MTM2012-33236; MTM2015-65016-C2-1-R: MINECO/FEDER,UE) and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR: 2014SGR551). The authors gratefully acknowledge the constructive comments of the anonymous referees which have undoubtedly helped to signficanlty improve the quality of the paper.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London (UK). Reprinted in 2003 by Blackburn Press.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology*, 31, 563–580.

Aitchison, J. (2002). Simplicial inference. In *Algebraic Methods in Statistics and Probability* (ed. Viana MA and Richards DS), vol 287. Contemporary Mathematics Series: American Mathematical Society, Providence, RI (USA), 1–22.

Albert, J. H. and Gupta, A. K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables. *The Annals of Statistics*, 10, 1261–1268.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Statistics and Computing*, 22, 1021–1029.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 579–602.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 367–389.

Bache, K. and Lichman, M. (2013). UCI machine learning repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6, 1345–1382.

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

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (1999). Comment on "Singularity and nonnormality in the classification of compositional data" by Bohling, G. C., Davis, J. C., Olea, R. A. and Harff, J. *Mathematical Geology*, 31, 581–585.

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In Rastogi, R., Morik, K., Bramer, M., and Wu, X., editors, *ICDM 2004, fourth IEEE International Conference on Data Mining*, 19–26, Brighton. IEEE Computer Society.

Bouguila, N. (2011). Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22, 186–198.

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

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71, 52–78.

Browne, R. P. and McNicholas, P. D. (2013). A mixture of generalized hyperbolic distributions. ArXiv e-prints arXiv:1305.1036

Buccianti, A. (2011). *Natural Laws Governing the Distribution of the Elements in Geochemistry: The Role of the Log-Ratio Approach*, 255–266. John Wiley and Sons, Ltd.

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

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14, 315–332.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194–206.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38.

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

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37, 795–828.

Eilers, P.H.C., Marx, B.D. and Durban, M. (2015). Twenty years of P-splines. *SORT*, 39, 149–186.

Ferrer-Rosell, B., Coenders, G., and Martínez-García, E. (in press). Segmentation by tourist expenditure composition. An approach with compositional data analysis and latentclasses. *Tourism Analysis*.

Giordan, M. and Wehrens, R. (2015). A comparison of computational approaches for maximum likelihood estimation of the Dirichlet parameters on high-dimensional data. *SORT*, 39, 109–126.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

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

Lee, S. X. and McLachlan, G. J. (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. ArXiv e-prints arXiv:1109.4706

Lee, S. X. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24, 181–202.

Lin, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, 20, 343–356.

Mardia, K. V., Taylor, C. C. and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63.

Martín-Fernández, J. A., Daunis-i-Estadella, J. and Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *SORT*, 39, 231–252.

Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2007). The skew-normal distribution on the simplex. *Communications in Statistics-Theory and Methods*, 36, 1787–1802.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. In *Compositional Data Analysis*, 29–42. John Wiley and Sons, Ltd.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *SORT*, 37, 29–56.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models, Willey Series in Probability and Statistics*. John Wiley and Sons, New York.

Monti, G. S., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2011). Notes on the scaled Dirichlet distribution. In *Compositional Data Analysis*, 128–138. John Wiley and Sons, Ltd.

Monti, G. S., Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The shifted-scaled Dirichlet distribution in the simplex. In Egozcue, J. J., Tolosana-Delgado, R. and Ortego, M. I., editors, *CoDaWork 2011, the 4th International Workshop on Compositional Data Analysis*, Sant Feliu de Guíxols. CIMNE.

Narayanan, A. (1991). Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 40, 365–374.

Ng, K. W., Tian, G.-L. and Tang, M.-L. (2011). *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley and Sons.

Neocleous, T., Aitken, C. and Zadora, G. (2011). Transformations for compositional data with zeros with an application to forensic evidence evaluation. *Chemometrics and Intelligent Laboratory Systems*, 109, 77–85.

Ongaro, A. and Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114, 412–426.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Buccianti, A. (2014). Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *Journal of Geochemical Exploration*, 141, 71–77.

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

Papageorgiou, I., Baxter, M. J. and Cau, M. A. (2001). Model-based cluster analysis of artefact compositional data. *Archaeometry*, 43, 571–588.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15, 384–398.

Prates, M. O., Lachos, V. H. and Cabral, C. R. B. (2013). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna, Austria.

Rayens, W. S. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the Simplex. *Journal of the American Statistical Association*, 89, 1465–1470.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R. (2011). *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. John Wiley and Sons Ltd, Chichester (UK).

Scealy, J. L., Patrice de Caritat, Grunsky, E. C., Tsagris, M. T and Welsh, A. H. (2015). Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, 136–148, DOI: 10.1080/01621459.2014.990563.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 387–397.

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

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York.

Vives-Mestres, M., Daunis-i Estadella, J. and Mártin-Fernández, J. A. (2014). Individual T-2 control chart for compositional data. *Journal of Quality Technology*, 46, 127–139.

Wang, H., Liu, Q., Mok, H. M. K., Fu, L. and Tse, W. M. (2007). A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179, 459–468.