

# APORTACIONS DE L'ANÀLISI COMPOSICIONAL A LES MIXTURES DE DISTRIBUCIONS

**Marc Comas Cufi**

Per citar o enllaçar aquest document:  
Para citar o enlazar este documento:  
Use this url to cite or link to this publication:  
<http://hdl.handle.net/10803/664902>



<http://creativecommons.org/licenses/by/4.0/deed.ca>

Aquesta obra està subjecta a una llicència Creative Commons Reconeixement

Esta obra está bajo una licencia Creative Commons Reconocimiento

This work is licensed under a Creative Commons Attribution licence



TESI DOCTORAL

Aportacions de  
l'anàlisi composicional a les  
mixtures de distribucions

MARC COMAS CUFÍ  
2018





TESI DOCTORAL

**Aportacions de  
l'anàlisi composicional a les  
mixtures de distribucions**

MARC COMAS CUFÍ  
2018

Programa de Doctorat en Tecnologia

Directors

Dra. Glòria Mateu Figueras

i

Dr. Josep A. Martín Fernández

Memòria presentada per optar al títol de doctor per la Universitat de Girona





La Dra. Glòria Mateu Figueras i el Dr. Josep Antoni Martín Fernández, professors del departament d'Informàtica, Matemàtica Aplicada i Estadística de la Universitat de Girona,

#### DECLAREM

Que el treball titulat *Aportacions de l'anàlisi composicional a les mixtures de distribucions*, que presenta Marc Comas Cufí per a l'obtenció del títol de doctor, ha estat realitzat sota la nostra direcció.

I perquè així consti i tingui el efectes oportuns, signem aquest document.

Signatures,

Glòria Mateu Figueras

Josep Antoni Martín Fernández

Girona, 6 de juny de 2018.



*A la Lia,  
el petit cigró de primavera;  
i a la Laia,  
perquè amb tu tot és possible.*





# Agraïments

Aquesta història va començar fa temps, i durant aquest temps moltes són les persones a qui vull agrair el seu temps dedicat. Essent conscient que em deixo gent, seré breu.

A les amistats: en especial a la Nuri, tenies raó «ja la tenim aquí». Gràcies pels ànims donats en els moments de més dubtes. Santi, Pepus, Carles, Vera, Marina i Ivan amb vosaltres ha estat un plaer descobrir el món CoDa. Martín i Glòria vosaltres dos m'heu mostrat la passió, l'ordre i l'empenta necessaris per tirar endavant aquesta tesi. Juanjo i Javier, moltes gràcies pels coneixements compartits, sens dubte sou referents a seguir. En general, gràcies als amics del carrer, als amics de la carrera i als amics de can Gol; Rafel, gràcies per deixar-me compaginar tantes escapades estadístiques.

A la família: als meus pares, vosaltres heu estat els que vau plantar i regar la llavor d'aquest fruit. Papa, m'hauria agradat que hi fossis per poder-ho compartir amb tu. Carlos, Olga, David, Laura, *tiu* i tia, gràcies per estar aquí quan les coses es torcen; el camí amb vosaltres sempre es fa menys feixuc. Finalment, gràcies Laia per entendre'm i ajudar-me, sense tu això no hauria estat possible.



# Publicacions

Aquesta tesi es presenta com a compendi dels següents articles:

- Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G. (2016), **Log-ratio methods in mixture models for compositional data sets**. *Statistics and Operations Research Transactions*, 40 (2), pp. 349–374. Factor d'impacte: 1.333, segon quartil (Q2) del *Journal Citation Report (JCR)* de l'*Institute of Scientific Information*.
- Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G. (2017), **Merging the components of a finite mixture using posterior probabilities**. *Statistical Modelling*, a impremta. Factor d'impacte: 0.932, segon quartil (Q2) del *Journal Citation Report (JCR)* de l'*Institute of Scientific Information*.
- Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G., Palarea-Albaladejo, J. (2018), **Modelling count data with the logistic-normal-multinomial distribution**. *Computational Statistics & Data Analysis*, en revisió. Factor d'impacte: 1.693, primer quartil (Q1) del *Journal Citation Report (JCR)* de l'*Institute of Scientific Information*.

A part de les publicacions, de l'anàlisi composicional aplicada a les mixelures de distribucions quan aquestes estan definides en el Símplex se'n desprenen *quatre* aportacions en forma de presentació oral a congressos:

- 5th International Conference of the ERCIM WG on COMPUTING & STATISTICS (ERCIM 2012). **Model-based clustering via Gaussian mixture models for compositional data: protein consumption in Europe**. Marc Comas-Cufí, Glòria Mateu-Figueras, Santiago Thió-Henestrosa, Josep Antoni Martín-Fernández. Oviedo, Spain.

- 5th International Workshop on Compositional Data Analysis (CODAWORK 2013). **Model-based clustering via Gaussian mixture models: a compositional sensitivity analysis.** Marc Comas-Cufí, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández. Vorau, Austria.
- 6th International Conference of the ERCIM WG on COMPUTING & STATISTICS (ERCIM 2013). **Compositional entropies in model based clustering.** Marc Comas-Cufí, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández. London, UK.
- IAMG 2015, the 17th annual conference of the International Association for Mathematical Geosciences. **Finite mixtures of distributions: compositional model-based clustering.** Marc Comas-Cufí, Antonella Buccianti, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras. Freiberg, Germany.

De la utilització de l'anàlisi composicional aplicada a les mixtures de distribucions per a la classificació basada en models paramètrics s'en desprenen les següents *tres* aportacions a congressos:

- I International Workshop on Proximity Data, Multivariate Analysis and Classification (I AMyC). **Merging classes.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras. Granada, Spain.
- 6th International Workshop on Compositional Data Analysis (CODAWORK 2015). **A compositional approach for merging finite mixture components.** Marc Comas-Cufí, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández. l'Escala, Spain.
- Conference of the International Federation of Classification Societies (IFCS 2015). **An integrated formulation for merging mixtures components based on posterior probabilities.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras. Bologna, Italy.

Finalment, de l'anàlisi composicional aplicada a les mixtures de distribucions per a la definició d'una nova distribució en l'espai de comptatges s'en desprenen les següents *sis* aportacions a congressos:

- Royal Statistical Society Annual meeting (RSS 2016). **A parametric approach to count zeros imputation in compositional data sets.** Marc Comas-Cufí, Javier Palarea-Albaladejo, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras. Manchester, UK.
- XXXVI Congreso Nacional de Estadística e Investigación Operativa (SEIO 2016). **Un enfoque paramétrico para el tratamiento de ceros de conteo en conjuntos de datos composicionales.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, Javier Palarea-Albaladejo. Toledo, Spain.
- II International Workshop on Proximity Data, Multivariate Analysis and Classification (II AMyC). **A parametric approach to count data in compositional data sets.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, Javier Palarea-Albaladejo. Barcelona, Spain.
- 7th International Workshop on Compositional Data Analysis (CODAWORK 2017). **Maximum likelihood estimation for the logistic-normal-multinomial distribution.** Marc Comas-Cufí, Glòria Mateu-Figueras, Josep Antoni Martín-Fernández, Javier Palarea-Albaladejo. Abbadia San Salvatore, Italy.
- Conference of the International Federation of Classification Societies (IFCS 2017). **Model-based clustering of count data based on the logistic-normal-multinomial distribution.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, Javier Palarea-Albaladejo. Tokyo, Japan.
- III International Workshop on Proximity Data, Multivariate Analysis and Classification (III AMyC). **Estimating the parameters of a logistic-normal-multinomial distribution.** Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, Javier Palarea-Albaladejo. Valladolid, Spain.



# Llista d'abreviatures

alr	Additive Log-Ratio
clr	Centered Log-Ratio
CoDa	Compositional Data
DM	Dirichlet-Multinomial
ilr	Isometric Log-Ratio
LNМ	Log-ratio-Normal-Multinomial
SBP	Sequential Binary Partition





# Índex de figures

3.1	Representacions equivalents de composicions de tres parts a (a) $\mathbb{R}^3$ i (b) al diagrama ternari. . . . .	19
3.2	Representació gràfica de l'operació clausura. . . . .	20
3.3	Subcomposició $\mathbf{x}' \in \mathcal{S}^2$ representada com a projecció lineal de $\mathbf{x} \in \mathcal{S}^3$ . . . . .	21
3.4	A l'esquerra, pertorbació de les composicions inicials $*$ per $p = (0.1, 0.1, 0.8)$ que resulten en $\circ$ . A la dreta, potència de les composicions inicials $*$ per $\alpha = 0.2$ resultant en $\circ$ . . . . .	23
3.5	Per visualitzar les relacions, angles, distàncies, . . . cal representar les dades en coordenades. . . . .	28
3.6	Rectes paral·leles al símplex. A l'esquerra, $\log x_2 - \log x_3 = k$ per a $k = -2, 0, 2$ . A la dreta, $\log x_1 - 2 \log x_2 + \log x_3 = k$ per a $k = -4, -2, 0, 2, 4$ . . . . .	28
3.7	Rectes ortogonals a $\mathcal{S}^3$ . A l'esquerra, $r_1 : x_2 = x_3$ i $r_2 : 2 \log x_1 - \log x_2 - \log x_3 = 0$ . A la dreta, $r_1 : \log x_1 - 3 \log x_2 + 2 \log x_3 = 0$ i $r_2 : 5 \log x_1 - \log x_2 - 4 \log x_3 = 0$ . . . . .	28
3.8	Circumferències a $\mathcal{S}^3$ de radi $r = 0.5, 1, 2$ . A l'esquerra amb centre ( $\circ$ ) a $(1/3, 1/3, 1/3)$ que és el baricentre del triangle i a la dreta a $(2/6, 1/6, 3/6)$ . . . . .	29
3.9	Gràfica de la funció de distribució d'algunes mixtures finites de distribucions normals. . . . .	32
3.10	Exemple del procés de combinació de les components d'una mixtura de vuit components. . . . .	37
5.1	Observacions de tres components provinents de vidres de diferents zones. . . . .	134
5.2	Mixtura finita gaussiana definida a $\mathbb{R}^2$ a través de l'eliminació d'una component d'una mostra definida a $\mathcal{S}^3$ . . . . .	135

---

5.3	Mixtura finita de distribucions Dirichlet ajustada al conjunt reduït de vidres. . . . .	136
5.4	Mixtura finita de distribucions normals sobre $\mathcal{S}^3$ ajustada al conjunt reduït de vidres. A l'esquerra la representació de la mixtura a l'espai de coordenades. A la dreta la representació de la mixtura al diagrama ternari. . . . .	137
5.5	Mixtura de quatre distribucions normals i la representació de les probabilitats a posteriori a $\mathcal{S}^4$ . . . . .	138
5.6	Probabilitats reals d'una mostra en equilibri de Hardy-Weinberg i les estimacions fetes amb la distribució Dirichlet-multinomial i la logquocient-normal multinomial. . . . .	140
5.7	Mostra en equilibri de Hardy-Weinberg i generació aleatòria feta amb la distribució Dirichlet-multinomial i la logquocient-normal multinomial ajustades a la mostra original. . . . .	140

# Índex de taules

2.1	Relació entre objectius del nucli central de la tesi i els articles publicats o enviats. . . . .	16
3.1	Exemple de partició seqüencial binària (SBP): coordenades ilr $\mathbf{y} = (y_1, y_2)$ i base $\Psi$ . . . . .	27



# Índex

<b>Resum</b>	<b>1</b>
<b>Resumen</b>	<b>3</b>
<b>Abstract</b>	<b>5</b>
<b>1 Introducció</b>	<b>9</b>
1.1 Motivació . . . . .	9
1.2 Situació dins la recerca . . . . .	9
1.3 Presentació dels articles . . . . .	12
1.4 Estructura de la tesi . . . . .	13
<b>2 Objectius</b>	<b>15</b>
2.1 Objectius del nucli central de la tesi . . . . .	15
<b>3 Metodologia</b>	<b>17</b>
3.1 Dades Composicionals . . . . .	17
3.1.1 Conceptes bàsics . . . . .	18
3.1.2 Principis de l'anàlisi de dades composicionals . . . . .	21
3.1.3 El símplex com a espai vectorial . . . . .	22
3.1.4 Transformacions del Símplex a l'espai real basades en logquocients . . . . .	22
3.1.5 Geometria al Símplex . . . . .	26
3.1.6 Models de distribució sobre el Símplex . . . . .	29
3.2 Models de mixtura . . . . .	30
3.2.1 Estimadors de màxima versemblança dels paràmetres d'un model de mixtura finita . . . . .	32
3.2.2 Classificació paramètrica . . . . .	34
3.2.3 Combinació de components . . . . .	36

<b>4</b>	<b>Articles</b>	<b>39</b>
4.1	Statistics and Operations Research Transactions . . . . .	43
4.2	Statistical Modelling . . . . .	71
4.3	Computational Statistics & Data Analysis . . . . .	105
<b>5</b>	<b>Resultats i discussió</b>	<b>133</b>
5.1	Mixtures finites de distribucions definides en el Símplex . . .	133
5.2	Combinació de les components d'una mixtura . . . . .	137
5.3	La distribució logquocient-normal multinomial . . . . .	139
5.4	Conclusions . . . . .	141
5.5	Futures línies d'investigació . . . . .	141
	<b>Bibliografia</b>	<b>143</b>

# Resum

La present tesi representa un compendi de *tres* treballs originals realitzats durant els anys 2014-2018. Aquests treballs comparteixen un nexa comú: tots ells són diferents aportacions de l'anàlisi composicional a l'estudi dels models basats en mixtures de distribucions de probabilitat. D'una forma molt breu, podríem dir que l'*anàlisi composicional* és una metodologia consistent en estudiar una mostra de mesures estrictament positives des d'un punt de vista relatiu. *Les mixtures de distribucions*, també anomenades barreges de distribucions, són un tipus particular de distribucions de probabilitat definides com la combinació lineal convexa d'altres distribucions.

En el primer treball que forma part d'aquesta tesi, es van analitzar quines opcions existien per a definir mixtures finites de distribucions de probabilitat dins l'espai mostral de les dades composicionals (Símplex) considerant la seva particular estructura algebraica. Entre les diferents opcions existents, es va constatar que, o bé les mixtures de distribucions no estaven ben definides en el Símplex, o bé les mixtures de distribucions no eren prou riques en quant a la capacitat per a modelar conjunts de dades composicionals reals. Això, portà a considerar la metodologia logquocient com a eina per a resoldre les problemàtiques existents. Mitjançant l'anàlisi composicional basada en logquocients es va proposar una metodologia per a la construcció de mixtures de distribucions de probabilitat ben definides en el Símplex, les qual són tant riques com les distribucions que existents per a modelar dades reals multivariants.

En general, els models basats en mixtures de distribucions s'ajusten amb l'algoritme EM. Aquest algoritme obté els paràmetres de les distribucions que intervenen en la mixtura i els paràmetres de la pròpia barreja. A part d'aquests paràmetres, l'algoritme també calcula la probabilitat de que cadascuna de les observacions hagi estat generada per cada una de les components que conformen la mixtura de distribucions. Aquestes probabilitats, anomenades probabilitats a posteriori, permeten classificar cadascuna de les observacions en la component més probable, convertint aquest procés en



un mètode d'agrupació molt popular. Alguns autors han proposat utilitzar aquestes probabilitats no només per classificar les observacions sinó també per definir una estructura jeràrquica de les components d'una mixtura de distribucions. En el segon treball d'aquesta tesi, es presentà un model que integrava totes les propostes trobades en la literatura, les quals basaven la construcció d'aquesta jerarquia en els vectors de probabilitats a posteriori. A més d'aquest nou model integrador, es van introduir nous mètodes per a la creació de jerarquies utilitzant mesures coherents, des d'un punt de vista composicional, per als vectors de probabilitats.

Les mixtures més freqüents, emergeixen de compondre una distribució categòrica amb una altra distribució de probabilitat, generalment definida a l'espai real. Així, al considerar la distribució categòrica composta amb una funció de distribució de probabilitat obtenim una mixtura finita d'aquesta distribució concreta amb pesos donats pels paràmetres de la distribució categòrica. En aquest cas, es diu que la distribució categòrica és la *distribució barreja*, l'altre distribució s'anomena la *distribució nucli*. Aquest procés es pot realitzar sempre que existeixi un mecanisme que permeti definir els paràmetres de la distribució nucli a partir dels valors observats d'una variable aleatòria amb la distribució pes. Concretament, si considerem la distribució multinomial com a distribució nucli i la distribució logquocient normal en el Símplex com a distribució de barreja, tindrem el que es coneix com la distribució de probabilitat logquocient normal multinomial. En el tercer i últim treball d'aquest compendi es deriven diferents propietats d'aquesta distribució, es presenta un nou mètode per estimar-ne els paràmetres i es mostra la seva capacitat per a modelar dades de comptatges enfront la distribució de probabilitat Dirichlet-multinomial, una de les més populars en aquest context.

# Resumen

La presente tesis representa un compendio de *tres* trabajos originales realizados durante los años 2014-2018. Estos trabajos comparten un nexo común: todos ellos son diferentes aportaciones del análisis composicional al estudio de los modelos basados en mixturas de distribuciones de probabilidad. Brevemente, podríamos decir que el *análisis composicional* es una metodología consistente en estudiar una muestra de medidas estrictamente positivas desde un punto de vista relativo. *Las mixturas de distribuciones*, también llamadas mezclas de distribuciones, son un tipo particular de distribuciones de probabilidad definidas como la combinación lineal convexa de otras distribuciones.

En el primer trabajo que forma parte de esta tesis, se analizó qué opciones existían para definir mixturas de distribuciones de probabilidad dentro del espacio muestral de los datos composicionales (Símplex) considerando su particular estructura algebraica. Entre las diferentes opciones existentes, se consideró que, o bien las mixturas de distribuciones no estaban bien definidas en el Símplex, o bien las mixturas de distribuciones no eran suficientemente ricas en cuanto a la capacidad para modelar conjuntos de datos composicionales reales. Eso, llevó a considerar la metodología logcociente como una herramienta para resolver las problemáticas existentes. Mediante el análisis composicional basado en logcocientes se propuso una metodología para la construcción de mixturas de distribuciones de probabilidad bien definidas en el Símplex, las cuales son tan ricas como las distribuciones existentes para modelar datos reales multivariantes.

En general, los modelos basados en mixturas de distribuciones se ajustan con el algoritmo EM. Este algoritmo obtiene los parámetros de las distribuciones que intervienen en la mixtura y los parámetros de la propia mezcla. A parte de estos parámetros, el algoritmo también calcula las probabilidades de que cada una de las observaciones hayan sido generadas por cada una de las componentes que conforman la mixtura de distribuciones. Estas probabilidades, llamadas probabilidades a posteriori, permiten clasificar cada una de

las observaciones en la componente más probable, convirtiendo este proceso en un método de agrupación muy popular. Algunos autores han propuesto utilizar estas probabilidades no solo para clasificar las observaciones sino también para definir una estructura jerárquica de las componentes de una mixtura de distribuciones. En el segundo trabajo que conforma esta tesis, se presentó un modelo que integraba todas las propuestas encontradas en la literatura, las cuales basaban la construcción de esta jerarquía en los vectores de probabilidad a posteriori. Además de este nuevo modelo integrador, se introdujeron nuevos métodos para la creación de jerarquías utilizando medidas coherentes, des de un punto de vista composicional, para los vectores de probabilidades.

Las mixturas más frecuentes, emergen de componer una distribución categórica con una distribución de probabilidad, generalmente definida en el espacio real. Así, al considerar la distribución categórica compuesta con una función de distribución de probabilidad obtenemos una mixtura finita de esta distribución concreta con pesos dados por los parámetros de la distribución categórica. En este caso, se dice que la distribución categórica es la *distribución de mezcla*, la otra se llama la *distribución núcleo*. Este proceso se puede realizar siempre que exista un mecanismo que permita definir los parámetros de la distribución núcleo a partir de los valores observados de un variable aleatoria siguiendo la distribución de mezcla. En particular, si consideramos la distribución multinomial como distribución peso y la distribución de probabilidad logcociente normal en el Simplex como distribución núcleo, tenemos lo que se conoce como la distribución de probabilidad logcociente normal multinomial. En el tercer y último artículo de este compendio se derivan diferentes propiedades de esta distribución, se presenta un nuevo método para estimar sus parámetros i se muestra su capacidad para modelar datos de cuentas frente la distribución de probabilidad Dirichlet-multinomial, una de las más populares en este contexto.

# Abstract

The present thesis is a compendium of *three* original works produced between 2014 and 2018. The papers have a common link: they are different contributions made by compositional data analysis to the study of the models based on mixtures of probability distributions. In brief, we could say that *compositional data analysis* is a methodology that consists of studying a sample of measures that are strictly positive from a relative point of view. *Mixtures of distributions* are a specific type of probability distribution defined to be the convex linear combination of other distributions.

In the first work that makes up this thesis, the available options for defining mixture of probability distributions within the sample space of compositional data (simplex) are analysed, considering their specific algebraic structure. Among the different available options, it is observed that either mixtures of distribution are not well defined in the simplex, or that they were not rich enough in terms of their capacity to model sets of real compositional data, leading us to consider the log-ratio approach as a tool to solve existing problems. By means of compositional data analysis based on log-ratios, a method for constructing mixtures of distributions that are well defined in the simplex and are as rich as existing distributions for modelling real multivariate data is proposed.

Generally, the models based on mixtures of distributions are adjusted with the EM algorithm, which obtains the parameters of the distributions that intervene in the mixture and the parameters of the mixture itself. Apart from these parameters, the algorithm also calculates the probability that each of the observations has been generated by each of the components that make up the mixture distribution. These probabilities, called posterior probabilities, allow for classifying each of the observations in the most probable component, making this process a very popular clustering method. Some authors have proposed using these probabilities not only to cluster observations, but also to define a hierarchical structure of the components of mixture distribution. In the second work that makes up this thesis, a

model is presented that integrates all the proposals found in the literature that base the construction of this hierarchy on the vectors of posterior probabilities. Apart from this new integrating model, new methods for creating hierarchies using coherent measures for vectors of probabilities, from a compositional point of view, are introduced

The most frequent mixtures emerge from putting a categorical distribution together with another probability distribution, which is generally defined in the real space. Thus, by means of considering the categorical distribution compounded with a function of probability distribution, we obtain a finite mixture of distributions of this specific distribution with weights given by the parameters of the categorical distribution. In this case, it is said that the categorical distribution is the weighting distribution; the other distribution is called the kernel distribution. This process can be carried out whenever there is a mechanism that allows the parameters of the kernel distribution to be defined from the observed values of a random variable following the weighting distribution. More specifically, if we consider the multinomial distribution as the kernel distribution and the logarithm quotient-normal distribution in the simplex as the *mixing distribution*, we will have what is known as the logarithm quotient-normal-multinomial probability distribution. In the third and last work of this compendium, different properties of this distribution are derived, a new method for estimating the parameters of the distribution is presented and the capacity improvement for modelling counting data compared with the Dirichlet-multinomial distribution, one of the most popular in this context, is demonstrated.





# Capítol 1

## Introducció

### 1.1 Motivació

El treball d'investigació a desenvolupar durant la tesi s'emmarca dins d'una de les línies de recerca principals del projecte coordinat CODA-RETOS “Análisis de datos composicionales y métodos relacionados” (Ref: MTM2015-65016-C2-1-R; Ministerio de Economía y Competitividad), més concretament dins el subprojecte CODA-TESC on un dels objectius principals és la “Caracterització dels elements químics en els diferents materials geològics: agrupació per mixtures”. Així doncs, el tema d'aquesta tesi doctoral forma part d'aquesta sublínia específica dedicada a les mixtures de distribucions.

Aquesta tesi parteix de la pregunta: *és possible aplicar l'anàlisi composicional a l'estudi de les mixtures de distribucions?* Alguns paràmetres que apareixen en la definició d'una mixtura de distribucions poden ser considerats elements del Símplex. Per una banda podem tenir mixtures de distribucions definides en el símplex, però per l'altra, les probabilitats a posteriori d'una mixtura de qualsevol distribució finita també poden ser considerades del mateix espai. Per tant, tots aquests paràmetres són susceptibles de ser analitzats mitjançant les eines logquocient provinents de l'anàlisi composicional.

### 1.2 Situació dins la recerca

Les mixtures de distribucions són models de probabilitat que permeten modelar heterogeneïtat dins d'un conjunt de dades. Aquesta heterogeneïtat s'aconsegueix assumint que una certa població està formada per diferents subpoblacions cada una d'elles modelada per un model de probabilitat par-



ticular.

Els models de mixtura han estat utilitzats en diferents parts de l'estadística. Històricament, l'aplicació més utilitzada dels models de mixtura apareix dins dels algorismes d'agrupació paramètrica. També, amb la incorporació de noves tècniques de computació, podem trobar una àmplia aplicació d'aquests models dins l'anàlisi de models lineal amb efectes mixtes. Finalment, els models de mixtura també poden ser utilitzats per a la construcció de nous models de probabilitat.

Les dades composicionals (CoDa de l'anglès *Compositional Data*) es troben àmpliament a la indústria química, petroquímica, farmacèutica, alimentària, etc. També les trobem en moltes i diverses aplicacions: com és l'anàlisi de l'ús del temps (sociologia), la composició de minerals a les roques (geologia), l'abundància d'espècies (biologia), la distribució dels recursos d'una empresa entre departaments (economia), percentatges de població (demografia), etc. Tots aquests exemples tenen en comú que les dades descriuen quantitativament les parts d'un total.

Aquesta tesi s'ha centrat en tres situacions on els models de mixtura han pogut enriquir-se de les eines de l'anàlisi de dades composicionals:

- A la literatura és difícil trobar mixtures de distribucions per dades composicionals que considerin distribucions definides en el Símplex. Les excepcions són alguns estudis (Albert i Gupta, 1982; Bouguila *et al.*, 2004; Calif *et al.*, 2011) que han considerat mixtures de distribucions Dirichlet, segurament la distribució més tradicional definida en el Símplex. No obstant això, en altres estudis s'ignora la naturalesa composicional de les dades i es tracten com dades definides a l'espai real (Papageorgiou *et al.*, 2001). També, en certs treballs pràctics, per exemple a Ferrer-Rossell *et al.* (2016), s'ha utilitzat la metodologia logquocient per ajustar models de mixtura sense gaires consideracions teòriques ni metodològiques. Per tant, a la literatura apareix un forat on la metodologia logquocient pot contribuir a la modelització de dades composicional amb nous models de mixtura.
- Quan els models de mixtura són utilitzats per agrupar observacions, no necessàriament composicionals, podem trobar que grups definits per diferents components d'una mixtura no formin, des de cert punt de vista, dos grups diferents. En aquest cas, diríem que sembla més natural que els dos grups formin un únic grup i que aquest grup estigui modelat per la mixtura de dues components. A la literatura existeixen diferents enfocaments per combinar les components d'una

mixtura basant-se en diferents criteris: Ray i Lindsay (2005) construeix un criteri basant-se en la modalitat de la mixtura formada per dues components, Melnykov (2016) construeix un mètode basant-se el nivell de superposició obtingut a través de la simulació de noves observacions utilitzant els paràmetres ajustats a la mostra. Per contra, altres enfocaments han utilitzat les probabilitats a posteriori obtingudes després d'ajustar la mixtura Baudry *et al.* (2010); Longford i Bartosova (2014). A Hennig (2010) es fa un resum de mètodes existents i s'hi proposen altres aproximacions basades en diferents criteris. En el cas particular dels criteris basats en les probabilitats a posteriori, no s'ha considerat que aquestes estimacions són elements del Símplex, i per tant, són susceptibles de ser analitzats amb la metodologia logquocient. Sembla raonable pensar que la metodologia logquocient permetrà introduir nous criteris per a combinar les components d'una mixtura, i possiblement, definir nous criteris de parada en el procés de combinar components.

- Les dades multivariants provinents de comptatges poden ser modelades amb una distribució multinomial, i el principal paràmetre d'aquesta distribució és un vector de probabilitats definit en el Símplex. Així doncs, resulta natural considerar una distribució definida en el Símplex com a distribució de barreja en la definició d'una mixtura amb la distribució multinomial. Normalment, aquesta mixtura de distribucions resultant s'anomena una composició (de l'anglès *compounding*) de distribucions. El cas més conegut és la composició de la distribució Dirichlet amb la distribució multinomial, més coneguda com la distribució Dirichlet-multinomial (DM), o la distribució de Pólya-Eggenberger multivariant (Johnson *et al.*, 1997). El principal problema d'utilitzar la distribució Dirichlet és que imposa una forta estructura d'independència a les components (Aitchison, 1986). A l'espai real, la distribució més utilitzada per a modelar variacions al voltant d'un centre és la distribució normal, per tant, sembla natural considerar la distribució equivalent definida en el Símplex (Mateu-Figueras *et al.*, 2013) com la distribució preferida per a ser composta amb la distribució multinomial. Aquest enfocament ha estat utilitzat anteriorment per Billheimer *et al.* (2001) des d'un punt de vista bayesià i per Xia *et al.* (2013) amb un mètode d'estimació ad-hoc basat amb mètodes de cadenes de Markov. El principal problema d'aquesta nova composició de distribucions (normal en el Símplex amb la distribució multinomial que hem anomenat logquocient normal multinomial) és

que la funció de probabilitat no admet una forma analíticament tractable. Per tant, apareix un forat en quant a la introducció de nous mètodes d'estimació. Igualment, sembla interessant realitzar una comparació entre aquesta distribució i la distribució DM en quant a la seva capacitat per a modelar dades provinents de comptatges.

Tot el que s'ha explicat obre la porta a la presentació de tres articles que representen aportacions de l'anàlisi composicional a les mixtures de distribucions. Les aportacions són originals i es focalitzen en diferents aspectes dels models de mixtura.

### 1.3 Presentació dels articles

La present tesi introdueix tres escenaris on l'anàlisi composicional s'ha aproximat a les mixtures de distribucions. Aquesta recerca ha culminat amb les següents aportacions innovadores:

- El primer article que conforma aquesta tesi es titula **Log-ratio methods in mixture models for compositional data sets** i ha estat publicat a la revista *Statistical & Operational Research Transactions* (SORT). Una transcripció de l'article es troba a la pàgina 43. En aquest article s'expliquen els enfocaments existents per a definir una mixtura de distribucions en el Símplex i s'enumeren els principals problemes. Tot seguit s'introdueix la proposta d'utilitzar la metodologia logquocient per a la construcció de mixtures definides en el Símplex i finalment, s'il·lustren i es comparen les diferents metodologies amb un exemple senzill però didàctic.
- El segon article es titula **Merging the components of a finite mixture using posterior probabilities** i ha estat publicat a la revista *Statistical Modelling*. Una transcripció de l'article es troba disponible a la pàgina 71. En aquest article es presenta una nova formulació per a decidir quines components d'una mixtura finita caldrien ser considerades com una única component. A continuació es mostra com la nova proposta generalitza altres enfocaments existents i es proposen dues noves alternatives basades en l'aplicació de la metodologia logquocient en el vector de probabilitats a posteriori. Finalment, es presenten alguns nous mètodes heurístics per a decidir quan parar el procés de combinació de components.

- El darrer article es titula **Modelling count data using the logratio-normal-multinomial distribution** i actualment es troba en revisió a la revista *Computational Statistics & Data Analysis*. Una transcripció de l'article està disponible a la pàgina 105. En aquest article es presenta una nova distribució definida per a dades multivariants provinents de comptatges resultant de compondre la distribució multinomial pels comptatges i la distribució logquocient normal pel vector de probabilitats del model multinomial. Aquesta nova distribució l'anomenem logquocient normal multinomial (LNM). Tot seguit es presenten algunes propietats de la nova distribució i es proposa un mètode d'estimació dels seus paràmetres. Finalment, es realitza una comparació amb la distribució Dirichlet multinomial (DM) en quant a la capacitat que tenen per a modelar diferents escenaris simulats i reals.

## 1.4 Estructura de la tesi

Un cop situada la tesi dins la recerca actual i presentats els articles, els següents capítols s'estructuren de la següent manera. Es descriuen els objectius generals i específics al Capítol 2. El Capítol 3 presenta breument i sintètica els aspectes metodològics bàsics de les dades composicionals i dels models de mixtura relacionats amb els treballs presentats. En tots tres casos es fa un repàs de la literatura més destacada a la que el lector pot recórrer en cas de desitjar informació més detallada. El Capítol 4 conforma el nucli central de la tesi, i conté una còpia dels articles publicats o una transcripció de l'article enviat. La tesi continua (Capítol 5) amb una síntesi dels principals resultats i una discussió d'aquests. Per últim, en finalitzar el capítol, es presenten les conclusions així com les futures línies d'investigació.



## Capítol 2

# Objectius

La present tesi té com a objectiu general l'aplicació de l'anàlisi composicional a les mixtures de distribucions de probabilitat. La forma d'introduir les dades composicionals a l'estudi de mixtures de distribucions ha constatat de tres aproximacions diferents que han donat lloc a tres aportacions en diferents àrees de recerca: modelització de dades composicionals, agrupació paramètrica basada en models de mixtura i modelització de dades de comptatge.

En la primera aproximació es proposa introduir l'anàlisi composicional per modelar dades definides en el Símplex a través de model de mixtura. En la segona aproximació es proposa utilitzar l'anàlisi composicional per decidir quins grups provinents d'una agrupació automàtica basada en models de mixtura poden ser considerats un únic grup. Finalment, en la tercera aproximació es proposa utilitzar l'anàlisi composicional per definir una nova distribució definida a l'espai de comptatges.

### 2.1 Objectius del nucli central de la tesi

Els objectius d'aquesta tesi poden ser separats respecte *on* apareixen les dades composicionals en els models de mixtura. Seguint aquesta classificació, podem separar els objectius en dos:

- Obj. 1. Aportacions de l'anàlisi composicional a les mixtures de distribucions quan les observacions estan definides sobre el Símplex.
- Obj. 2. Aportacions de l'anàlisi composicional a les mixtures de distribucions quan els paràmetres del model de mixtura estan definits sobre el Símplex.

Si en lloc de centrar-nos en l'objecte analitzat (observacions/paràmetres), ens centrem en la motivació que ens porta a treballar amb els models de mixtura, ens sorgeixen els següents *tres* objectius:

- Obj. A. Aportacions de l'anàlisi composicional a les mixtures de distribucions per a l'agrupació paramètrica.
- Obj. B. Aportacions de l'anàlisi composicional a les mixtures de distribucions per a la modelització de dades de composicionals.
- Obj. C. Aportacions de l'anàlisi composicional a les mixtures de distribucions per a la modelització de dades de comptatges.

Aquests objectius han estat abordats en tres publicacions diferents que han estat o estan essent revisades per revisors externs. A la Taula 2.1 tenim resumit, quina és l'aportació de cadascuna de les publicacions als objectius plantejats.

Article	Objectiu				
	1	2	A	B	C
Log-ratio methods in mixture models for compositional data sets	✓		✓	✓	
Merging the components of a finite mixture using posterior probabilities	✓	✓	✓	✓	✓
Modelling count data using the logratio-normal-multinomial distribution		✓			✓

Taula 2.1: Relació entre objectius del nucli central de la tesi i els articles publicats o enviats.

## Capítol 3

# Metodologia

En aquest capítol es fa una revisió dels principals aspectes referents a les principals temàtiques que tracta la tesi; són d'una banda les dades composicionals (CoDa de l'anglès *Compositional Data*) i d'una altra banda les mixtures de distribucions. Pel que fa a les dades composicionals, veurem els principis bàsics de la metodologia logquocient, l'estructura algebraico-geomètrica del símplex i diferents famílies de distribucions. Pel que fa a les mixtures, ens centrarem en la seva estimació i com s'utilitzen en la classificació paramètrica.

### 3.1 Dades Composicionals

Aquesta secció és un resum dels aspectes més bàsics de l'anàlisi de dades composicionals (anàlisi CoDa de l'anglès *Compositional Data*). Els exemples, notació i organització dels cinc primers punts d'aquesta secció han estat extrets de la tesi doctoral Vives-Mestres (2014) amb el permís explícit de l'autora qui basà el seu text en Pawlowsky-Glahn *et al.* (2010) i les tesis doctorals de Mateu-Figueras (2003) i Martín-Fernández (2001). En cas d'interès, el lector pot obtenir en aquestes referències més informació. També, per a més aprofundiment en l'anàlisi de dades composicionals, es suggereixen els llibres Pawlowsky-Glahn i Buccianti (2011) i Pawlowsky-Glahn *et al.* (2015)

Les dades composicionals són descripcions de les parts d'un tot, i per tant, habitualment solen expressar-se en tant per  $u$ , percentatges, ppm o concentracions. La restricció de suma constant que acostuma a caracteritzar les dades composicionals pot complicar l'anàlisi estadística així com les interpretacions que se'n derivin. Per exemple, si ens restringim a la



suma constant, el fet d'augmentar una de les parts, necessàriament implicarà reduir una de les altres parts. Això, obliga a replantejar-se el concepte d'independència entre parts quan es treballa amb vectors aleatoris composicionals.

També, el coeficient de correlació clàssic entre dues components, no es pot interpretar de la forma habitual. De fet, va ser Pearson mateix el primer a assenyalar que components amb un mateix denominador introdueixen una correlació falsa o espúria entre elles (Pearson, 1897). Aquest fet dificulta l'aplicació de l'anàlisi estadístic estàndard per analitzar dades composicionals.

Aitchison (1982, 1986) fou el primer a desenvolupar una metodologia específica amb la principal idea que les dades composicionals representen parts, i que per tant, l'única informació que contenen és relació relativa. És a dir, l'única de manera d'obtenir informació d'una part és a través de comparar-la amb una altra part. Això el portà a l'anàlisi dels quocients entre parts, i per manejabilitat, a l'anàlisi dels logquocients. Per tant, podríem dir que l'anàlisi de dades composicional és l'anàlisi dels logquocients entre parts d'una composició.

### 3.1.1 Conceptes bàsics

Tot i que la introducció de la metodologia de dades composicionals es podria fer a través de classes d'equivalència tal com està explicat a Barceló-Vidal *et al.* (2016), en aquesta tesi s'ha optat per la introducció de les dades composicionals a partir dels representats lineals que sumen una certa constant, que com veurem es poden definir a través de l'operació clausura. Recordem però, que una composició es defineix com un vector de components que representen parts d'un total, i que no tenen per què sumar una constant.

D'ara endavant ens referirem als elements d'una composició com a parts o components.

**Definició 3.1** Una composició amb  $D$ -parts és un vector  $(D \times 1)$  els components del qual  $x_1, x_2, \dots, x_D$  són nombres reals estrictament positius (i.e.  $x_1 > 0, x_2 > 0, \dots, x_D > 0$ ), que sumen un valor constant  $x_1 + x_2 + \dots + x_D = \kappa$  i que contenen informació relativa.

Habitualment  $\kappa = 1$  o  $\kappa = 100$  quan les mesures s'han transformat a proporcions o percentatges, respectivament. Recalcar altre cop, que encara que no siguin de suma constant, també serien composicions les dades mesurades en unitats de concentracions, com ara mg/l o molaritats.

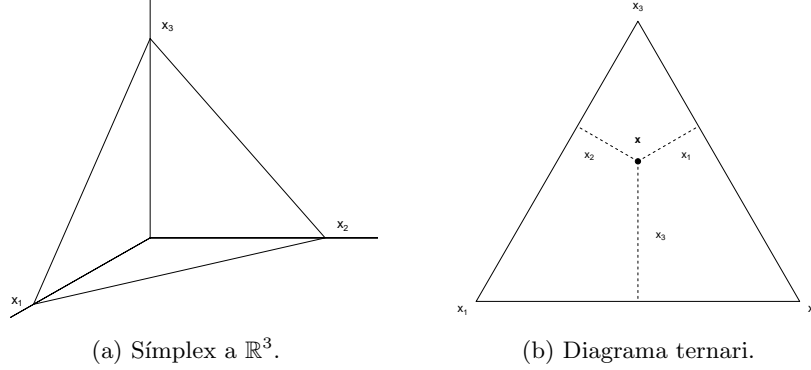


Figura 3.1: Representacions equivalents de composicions de tres parts a (a)  $\mathbb{R}^3$  i (b) al diagrama ternari.

**Definició 3.2** L'espai mostral natural de les composicions és el símplex  $\mathcal{S}^D$ , definit com

$$\mathcal{S}^D = \{(x_1, x_2, \dots, x_D) | x_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^D x_i = \kappa\}.$$

Les composicions de 3 parts amb  $\kappa = 1$  ( $D = 3$ ) es troben inscrites en un triangle equilàter a  $\mathbb{R}^3$ , situat al pla perpendicular al vector  $(1, 1, 1)$  (Figura 3.1a). No obstant això, és més habitual representar les dades al diagrama ternari (Figura 3.1b), que és una representació equivalent. Un diagrama ternari és un triangle equilàter tal que la mostra genèrica  $\mathbf{x} = (x_1, x_2, x_3)$  es troba a una distància  $x_1$  del costat oposat al vèrtex  $X_1$ , a una distància  $x_2$  del costat oposat al vèrtex  $X_2$  i a una distància  $x_3$  del costat oposat al vèrtex  $X_3$ . En el cas de  $D = 4$  el símplex es representa en un tetraedre regular d'alçada unitat.

Per aconseguir que les parts sumin una certa constant  $\kappa$ , una composició s'ha de dividir per la suma de totes les parts i multiplicar els quocients per  $\kappa$ . Aquesta operació s'anomena clausura.

**Definició 3.3** Per qualsevol vector de  $D$  components real positius  $\mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}_+^D$  la clausura de  $\mathbf{x}$  es defineix com

$$\mathcal{C}(\mathbf{x}) = \left( \frac{\kappa \cdot x_1}{\sum_{i=1}^D x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^D x_i}, \dots, \frac{\kappa \cdot x_D}{\sum_{i=1}^D x_i} \right)$$

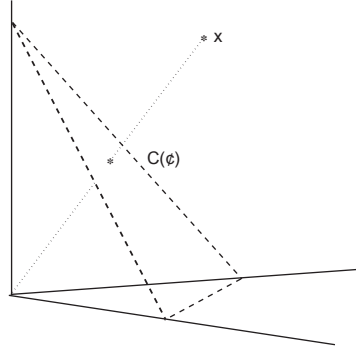


Figura 3.2: Representació gràfica de l'operació clausura.

La clausura és un escalament de les parts d'una composició per tal que sumin la constant  $\kappa$ . Des d'un punt de vista composicional, o sigui, des d'un punt de vista de logquocients, cal fer notar que l'operació de clausura no modifica en absolut la informació relativa (logquocients) entre les parts d'una composició. Tal com s'exposa a Barceló-Vidal *et al.* (2016), l'únic que s'està fent és canviar el representant dins la classe d'equivalència. La interpretació gràfica de l'operació de clausura es mostra a la Figura 3.2: la clausura de  $\mathbf{x}$  mou el punt al llarg de la recta (classe d'equivalència) que va des de l'origen fins a  $\mathbf{x}$  fins a la intersecció amb el pla  $\sum x_i = \kappa$ . En lloc de  $\mathbf{x}$ , el nou representant de la classe d'equivalència serà  $\mathcal{C}(\mathbf{x})$ .

En tant que l'interès d'una composició dins l'anàlisi composicional rau únicament en la informació relativa (logquocients), és fàcil veure que tots els logquocients d'una composició amb  $D$ -parts es poden obtenir a partir del coneixement de per exemple els  $D - 1$  quocients  $x_i/x_D$  per a  $i = 1, 2, \dots, D - 1$  (Aitchison, 1986). Això ens porta a justificar que la dimensió d'una composició amb  $D$ -parts és  $D - 1$ .

Quan només ens interessin algunes parts (no totes) de la composició  $\mathbf{x} \in \mathcal{S}^D$  diem que treballem amb una subcomposició. Com que l'anàlisi composicional basada en logquocients únicament es basa en la informació relativa, el fet de passar a estudiar una subcomposició no hauria de modificar els resultats obtinguts amb la composició completa. Aquesta propietat és coneguda com a *coherència subcomposicional* i és un dels punts que fa preferible l'estudi dels logquocients per l'anàlisi de dades composicionals.

Podem visualitzar la subcomposició dins del Símplex, com una composició de dimensió inferior obtinguda a través de projecció. A la Figura 3.3 es mostra com la subcomposició  $\mathbf{x}' \in \mathcal{S}^2$  formada amb les dues primeres parts

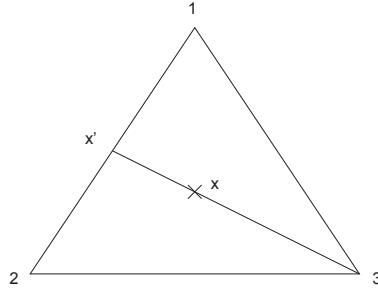


Figura 3.3: Subcomposició  $\mathbf{x}' \in \mathcal{S}^2$  representada com a projecció lineal de  $\mathbf{x} \in \mathcal{S}^3$ .

de  $\mathbf{x} \in \mathcal{S}^3$  és el resultat de la projecció de  $\mathbf{x}$  sobre el costat 12 des del vèrtex 3.

### 3.1.2 Principis de l'anàlisi de dades composicionals

Qualsevol mètode estadístic aplicat a una composició hauria de complir certes propietats, anomenades principis. Aquests fan que l'anàlisi d'un conjunt composicional sigui coherent amb la seva estructura. Aquests principis són: invariància per escala, invariància per permutació i coherència subcomposicional (Aitchison, 1986).

El principi d'invariància per canvi d'escala postula que els resultats d'una anàlisi han de ser els mateixos siguin quines siguin les unitats de la composició. L'anàlisi de quocients compleix aquest principi, ja que el quocient  $x_1/x_2 = (\lambda x_1)/(\lambda x_2)$  perquè les unitats es cancel·len. No obstant això, el quocient depèn de l'ordre de les parts, és a dir  $x_1/x_2 \neq x_2/x_1$ . Una transformació adequada utilitza logquocients, de la forma  $\log x_1/x_2$ . D'aquesta manera la inversió dels components produeix un canvi de signe, cosa que dóna una simetria respecte a l'ordre de les parts.

El principi d'invariància per permutació postula que les conclusions d'una anàlisi composicional no han de dependre de l'ordre de les parts. Els resultats obtinguts han de ser els mateixos si canviem l'ordre de les parts d'una composició.

Com ja hem comentat en l'apartat anterior, el principi de coherència subcomposicional diu que la inferència sobre subcomposicions ha de ser consistent, independentment de si la inferència es basa en la subcomposició o la composició completa. A l'espai real aquest principi es tradueix en que la inferència sobre un subconjunt de variables ha de ser la mateixa inde-

pendentment de si basem la inferència en un subconjunt de variables o el conjunt complet.

### 3.1.3 El símplex com a espai vectorial

Després dels treballs d'Aitchison, es va mostrar com la metodologia CoDa basada en l'anàlisi dels logquocient era equivalent a definir una estructura d'espai euclidià dins el Símplex (Barceló-Vidal *et al.*, 2016; Egozcue i Pawlowsky-Glahn, 2006). Presentem primer l'estructura d'espai vectorial.

Dins el Símplex es defineixen dues operacions bàsiques que anomenem: pertorbació i potència. La primera, definida per dues composicions  $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$  fa el paper d'operació interna (l'equivalent a la suma de l'espai real), i la segona, definida per una composició  $\mathbf{x} \in \mathcal{S}^D$  i un escalar  $\alpha \in \mathbb{R}$  farà el paper d'operació externa (l'equivalent a la multiplicació per un escalar a l'espai real).

**Definició 3.4** Siguin  $\mathbf{x}, \mathbf{x}^*$  dues composicions amb  $D$  parts. Llavors l'operació

$$\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1x_1^*, x_2x_2^*, \dots, x_Dx_D^*)$$

s'anomena pertorbació.

**Definició 3.5** Sigui  $\mathbf{x}$  una composició amb  $D$  parts i sigui  $\alpha$  un escalar de  $\mathbb{R}$ . Llavors l'operació

$$\alpha \otimes \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)$$

s'anomena potència.

Les operacions de pertorbació i potència, denotades  $\oplus$  i  $\otimes$  respectivament, doten l'espai del Símplex amb estructura d'espai vectorial sobre el cos  $\mathbb{R}$ . La Figura 3.4 mostra visualment el resultat d'aquestes operacions en un conjunt de composicions a  $\mathcal{S}^3$ .

### 3.1.4 Transformacions del Símplex a l'espai real basades en logquocients

En aquest apartat presentem les transformacions més típiques que ens permetran treballar directament sobre l'espai de coordenades. Aquestes transformacions són: logquocient additiva (alr de l'anglès *additive log-ratio*), logquocient centrada (clr de l'anglès *centred log-ratio*), logquocient isomètrica

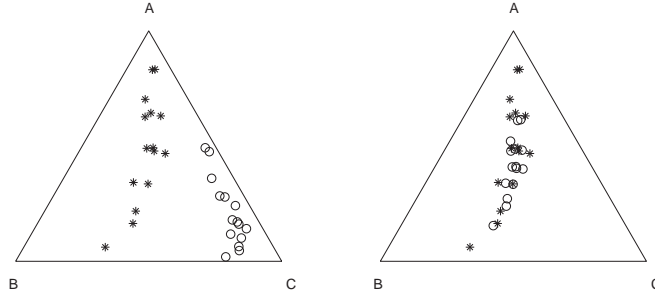


Figura 3.4: A l'esquerra, pertorbació de les composicions inicials  $*$  per  $p = (0.1, 0.1, 0.8)$  que resulten en  $\circ$ . A la dreta, potència de les composicions inicials  $*$  per  $\alpha = 0.2$  resultant en  $\circ$ .

(il·l de l'anglès *isometric log-ratio*). Tot i que les dues primeres van ser introduïdes merament com a transformacions, com veurem a la següent secció, les imatges d'aquestes transformacions representen l'expressió en coordenades respecte a una base particular d'una composició. Denotarem les coordenades fruit de les tres transformacions per  $\mathbf{w}$ ,  $\mathbf{z}$  i  $\mathbf{y}$  respectivament.

Al llarg de la tesi, es farà servir la notació  $\log$  per referir-nos al logaritme natural o neperià (en base  $e$ ) seguint la tendència marcada pels programes estadístics que fan servir per defecte la base natural.

### Transformació alr

Aitchison (1986) defineix la transformació logquocient additiva com

**Definició 3.6** Donada una composició amb  $D$  parts, la transformació logquocient additiva de  $\mathbf{x} \in \mathcal{S}^D$  a  $\mathbf{w} \in \mathcal{R}^{D-1}$  es defineix com

$$\mathbf{w} = \text{alr}(\mathbf{x}) = \left( \log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \dots, \log \frac{x_{D-1}}{x_D} \right)$$

La transformació alr és bijectiva i la seva inversa és l'  $\text{alr}^{-1}$  que es defineix com

$$x_i = \frac{\exp w_i}{\sum_{j=1}^{D-1} \exp w_j + 1} \quad (i = 1, 2, \dots, D-1),$$

$$x_D = 1 - \left( \sum_{i=1}^{D-1} x_i \right) = \frac{1}{\sum_{j=1}^{D-1} \exp w_j + 1}.$$

Un dels inconvenients de la transformació alr és la seva falta de simetria, ja que la component que figura en el denominador de cada logquocient adquireix un protagonisme especial respecte de la resta de components. Certament podríem escollir qualsevol altra component com a comú denominador.

### Transformació clr

Aitchison (1986) defineix la transformació logquocient centrada com:

**Definició 3.7** Donada una composició amb  $D$  parts, la transformació clr de  $\mathbf{x} \in \mathcal{S}^D$  a  $\mathbf{z} \in \mathbb{R}^{D-1}$  es defineix com

$$\mathbf{z} = \text{clr}(\mathbf{x}) = \left( \log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \dots, \log \frac{x_D}{g(\mathbf{x})} \right)$$

on  $g(\mathbf{x}) = (x_1 \cdot x_2 \cdots x_D)^{1/D}$  és la mitjana geomètrica de les  $D$  components de  $\mathbf{x}$ .

En aquest cas, la transformació és simètrica entre les parts. Les dades transformades se situen a l'hiperplà  $V$  de  $\mathbb{R}^D$  que passa per l'origen i és ortogonal al vector d'unitats  $(1, 1, \dots, 1)$ , és a dir,  $V = \text{clr}(\mathcal{S}^D) = \{\mathbf{z} \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$ . Això comporta una nova dificultat, ja que la suma de les components del vector transformat és igual a 0.

La transformació clr és bijectiva entre el Símplex i l'hiperplà  $V$ ; la seva inversa és la  $\text{clr}^{-1}$  que es defineix com

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{z}) = \mathcal{C}(e^{z_1}, e^{z_2}, \dots, e^{z_D})$$

L'inconvenient de la falta de simetria a la definició de la transformació alr no apareix a la definició de la transformació clr. No obstant això, l'estructura de covariàncies associada a aquesta transformació no s'allibera de l'inconvenient de la singularitat d'aquesta matriu.

A diferència de la transformació alr, la transformació clr compleix els principis d'invariància per permutació. És per aquest motiu que és utilitzada per a definir una distància entre composicions que compleixi tots els principis descrits a 3.1.2. Així doncs, es defineix la distància d'Aitchison entre dues composicions  $\mathbf{x}, \mathbf{x}^*$  com la distància euclidiana que tenen les seves respectives coordenades clr.

### Transformació ilr

Utilitzant la distància d'Aitchison descrita, Egozcue *et al.* (2003) defineixen una isometria entre els espais  $\mathcal{S}^D$  i  $\mathbb{R}^{D-1}$ . La motivació principal d'aquesta nova transformació és superar els inconvenients de les dues transformacions anteriors: la no invariància per permutació de la transformació alr (i.e. obliquïtat respecte la distància d'Aitchison definida a partir de la transformació clr) i la singularitat de la matrius de covariància de les coordenades clr. Utilitzant la distància d'Aitchison descrita, Egozcue *et al.* (2003) defineixen una isometria entre els espais  $\mathcal{S}^D$  i  $\mathbb{R}^{D-1}$ . La motivació principal d'aquesta nova transformació és superar els inconvenients de les dues transformacions anteriors: la no invariància per permutació de la transformació alr (i.e. obliquïtat respecte a la distància d'Aitchison definida a partir de la transformació clr) i la singularitat de la matriu de covariància de les coordenades clr.

La transformació isomètrica sorgeix de manera natural si observem la transformació clr. La condició  $\sum z_k = 0$  que satisfan les components dels vectors del subespai  $V = \text{clr}(\mathcal{S}^D)$  ens indica les coordenades clr es troben localitzades a l'hiperplà amb vector normal  $(1, 1, \dots, 1)$ . Per tant, com a subespai de dimensió  $D-1$  de  $\mathbb{R}^D$  podem escollir-hi una base ortonormal per identificar-hi qualsevol coordenada clr. Dit d'altra manera, donada aquesta base ortonormal, qualsevol composició queda completament identificada per les coordenades de la seva representació al subespai  $V$ .

Aquest procediment, transformació clr seguida d'un canvi de base ortonormal i de la projecció ortogonal sobre el subespai  $V$ , dóna lloc a una isometria entre els espais  $\mathcal{S}^D$  i  $\mathbb{R}^{D-1}$  considerant la distància d'Aitchison definida anteriorment a partir de la transformació clr. Egozcue *et al.* (2003) defineixen aquesta transformació i la denoten per ilr.

**Definició 3.8** Donada una base ortonormal del símplex  $\mathcal{S}^D$ ,  $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1})$ ,

i la matriu d'ordre  $(D-1 \times D)$  a  $\mathbb{R}^{D-1}$   $\Psi = \begin{pmatrix} \text{clr}(\mathbf{e}_1) \\ \text{clr}(\mathbf{e}_2) \\ \dots \\ \text{clr}(\mathbf{e}_{D-1}) \end{pmatrix}$ , es defineix la

transformació ilr d'una composició  $\mathbf{x} \in \mathcal{S}^D$  a un vector  $\mathbf{y} \in \mathbb{R}^{D-1}$  com

$$\mathbf{y} = \text{ilr}(\mathbf{x}) = \text{clr}(\mathbf{x}) \cdot \Psi'$$

A l'igual que la transformació clr, la transformació ilr compleix els tres principis de l'anàlisi CoDa.



La transformació isomètrica no és única, donat que en la seva definició no queda especificada la base ortonormal de  $\mathcal{S}^D$  i per tant tenim la llibertat d'escollir-la. Egozcue *et al.* (2003) proposen definir-la a partir d'una partició seqüencial binària (SBP de l'anglès *sequential binary partition*).

Una SBP és una jerarquia de les parts d'una composició: en un primer pas, la composició es divideix en dos grups; i en els passos següents, cada grup es divideix al seu torn en dos grups. A cada pas, el nombre de parts  $(x_{j_1}, \dots, x_{j_r})$  en el primer grup, codificades per  $+1$ , s'enregistra a  $r$  i el nombre de parts  $(x_{k_1}, \dots, x_{k_s})$  en el segon grup, codificades per  $-1$ , s'enregistra a  $s$ . Les coordenades ilr  $y_i$  obtingudes al pas  $i$  de la SBP i el corresponent element de la base  $\psi_i$  es calculen de la següent manera:

$$y_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \log \frac{(x_{j_1} x_{j_2} \dots x_{j_r})^{1/r_i}}{(x_{k_1} x_{k_2} \dots x_{k_s})^{1/s_i}}, \quad \psi_i = (\psi_1, \dots, \psi_D) \begin{cases} \psi_j = +\sqrt{\frac{s_i}{r_i(r_i+s_i)}} \\ \psi_k = -\sqrt{\frac{r_i}{s_i(r_i+s_i)}} \\ \psi_0 = 0 \end{cases},$$

on  $\psi_j$  és el coeficient per cada part  $x_{j_1}, \dots, x_{j_r}$  al numerador de  $y_i$  (codificat  $+1$  al SBP),  $\psi_k$  és el coeficient per cada part  $x_{k_1}, \dots, x_{k_s}$  al denominador de  $y_i$  (codificat  $-1$  al SBP) i  $\psi_0$  és el coeficient per les parts que no intervenen.

Com ja hem dit anteriorment, la suma dels elements de  $\psi_i$  és zero perquè el vector es troba a l'hiperplà  $V$ . A més a més, com que formen una base ortonormal,  $\psi_i \cdot \psi_l = 0$ , per  $i, l = 1, \dots, D-1$ ,  $i \neq l$ , i  $\|\psi_i\| = 1$ .

Les coordenades de la composició en la base  $\Psi$  s'anomenen balanços ( $y_i$ ) i les composicions de la base ( $\mathbf{e}_i$ ) s'anomenen *balancing elements*. Cada element ilr de la base  $\psi_i$  és un logcontrast, és a dir, una combinació lineal de logaritmes de les dades composicionals amb coeficients de suma zero.

A la Taula 3.1 es mostra un exemple de SBP pel cas de  $D = 3$ .

Podem utilitzar les operacions estàndards de l'espai real, treballar amb la distància euclidiana i aplicar el producte escalar ordinari sobre les dades ilr transformades.

Egozcue *et al.* (2003) donen les relacions entre les tres transformacions: alr, clr i ilr.

### 3.1.5 Geometria al Símplex

Es mostren a continuació algunes figures que pretenen mostrar gràficament que la geometria al símplex és diferent de la geometria euclidiana amb la qual estem acostumats a treballar a l'espai real.

La Figura 3.5 mostra rectes composicionals al símplex  $\mathcal{S}^3$  i les rectes equivalents a l'espai de coordenades ortonormals. És evident com les rectes

Taula 3.1: Exemple de partició seqüencial binària (SBP): coordenades ilr  $\mathbf{y} = (y_1, y_2)$  i base  $\Psi$ .

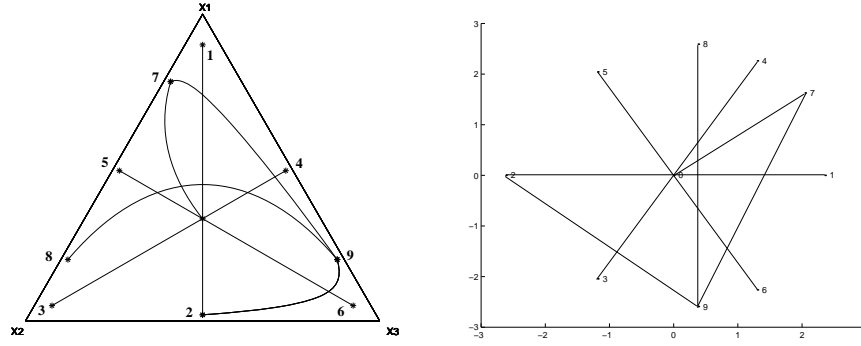
Ordre	$x_1$	$x_2$	$x_3$	$r$	$s$	Coordenada
1	-1	-1	1	1	2	$y_1 = \sqrt{\frac{1 \cdot 2}{1+2}} \log \frac{x_3}{\sqrt{x_1 x_2}}$
2	-1	1	0	1	1	$y_2 = \sqrt{\frac{1 \cdot 1}{1+1}} \log \frac{x_2}{x_1}$
$\Psi$	$-\sqrt{\frac{1}{6}}$ $-\sqrt{\frac{1}{2}}$	$-\sqrt{\frac{1}{6}}$ $+\sqrt{\frac{1}{2}}$	$+\sqrt{\frac{2}{3}}$ 0			

perpendiculars de l'espai real  $\overline{12}$  i  $\overline{89}$  queden deformades un cop representades a l'espai restringit. El mateix passa amb els angles. Vegeu com l'angle recte entre els segments  $\overline{50}$  i  $\overline{07}$  de l'espai real (Figura 3.5b) queda deformat en la seva representació al símplex de la Figura 3.5a.

Les Figures 3.6 i 3.7 mostren exemples de famílies de rectes paral·leles i ortogonals en el símplex  $\mathcal{S}^3$ . A partir d'aquests gràfics resulta evident el fet que les imatges gràfiques que tenim de recta, paral·lelisme i ortogonalitat procedents de l'espai real no són vàlides a l'espai de les composicions, malgrat ser ambdós espais mètrics euclidians.

Així, per exemple, observant les rectes de la Figura 3.6, resulta clar que el camí més curt entre dos punts del símplex no sempre és el segment rectilini entès en la forma “estàndard”. Naturalment, però, prenem les coordenades ilr o clr de totes les rectes de representades a les Figures 3.6 i 3.7, obtindríem imatges estàndard de rectes paral·leles i ortogonals contingudes en el pla  $z_1 + z_2 + z_3 = 0$  de  $\mathbb{R}^3$ .

Per acabar, la Figura 3.8, mostra les gràfiques d'unes quantes circumferències representades sobre  $\mathcal{S}^3$ . Igual com passava amb les rectes, els perfils d'aquestes circumferències composicionals no tenen res a veure amb els perfils estàndard d'aquestes figures. La proximitat a la frontera del símplex provoca distorsions en els perfils, des d'un punt de vista euclidià. Això és pel fet que la distància entre dos punts molt “propers” entre si (en el sentit estàndard del terme) situats gairebé tocant la frontera del triangle és molt més gran que la distància de dos punts amb la mateixa proximitat situats en la zona central del símplex.



(a) Rectes composicionals al símplex  $\mathcal{S}^3$  (b) Equivalència de les rectes composicionals a l'espai real  $\mathbb{R}^2$ .

Figura 3.5: Per visualitzar les relacions, angles, distàncies, ... cal representar les dades en coordenades.

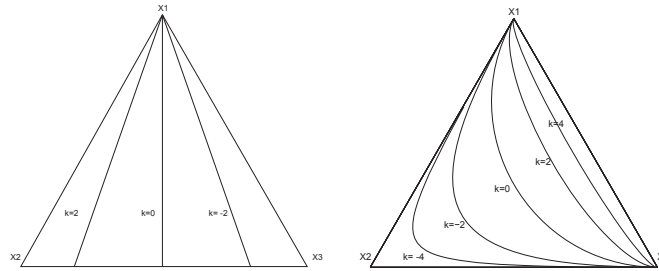


Figura 3.6: Rectes paral·leles al símplex. A l'esquerra,  $\log x_2 - \log x_3 = k$  per a  $k = -2, 0, 2$ . A la dreta,  $\log x_1 - 2 \log x_2 + \log x_3 = k$  per a  $k = -4, -2, 0, 2, 4$ .

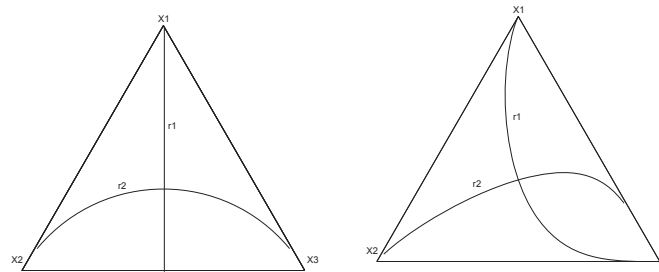


Figura 3.7: Rectes ortogonals a  $\mathcal{S}^3$ . A l'esquerra,  $r_1 : x_2 = x_3$  i  $r_2 : 2 \log x_1 - \log x_2 - \log x_3 = 0$ . A la dreta,  $r_1 : \log x_1 - 3 \log x_2 + 2 \log x_3 = 0$  i  $r_2 : 5 \log x_1 - \log x_2 - 4 \log x_3 = 0$

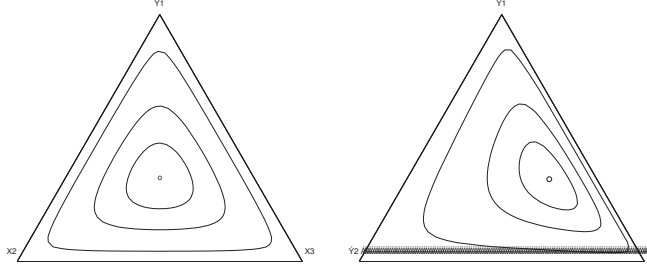


Figura 3.8: Circumferències a  $\mathcal{S}^3$  de radi  $r = 0.5, 1, 2$ . A l'esquerra amb centre (o) a  $(1/3, 1/3, 1/3)$  que és el baricentre del triangle i a la dreta a  $(2/6, 1/6, 3/6)$ .

### 3.1.6 Models de distribució sobre el Símplex

Aquest part és una breu enumeració d'algunes de les distribucions més conegudes que estan definides sobre el Símplex. La major part del contingut ha estat extret de (Mateu-Figueras, 2003).

#### Distribució de Dirichlet

La distribució més coneguda definida sobre  $\mathcal{S}^D$  és la distribució Dirichlet (Aitchison, 1986). La distribució Dirichlet queda completament determinada a partir d'un vector positiu amb  $D$  components  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)$ , anomenat el paràmetre de concentració. La funció de densitat és

$$f(\mathbf{x} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^D \alpha_i)}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D x_i^{\alpha_i - 1}.$$

Podem obtenir una composició aleatòria  $\mathbf{x}$  que segueixi una distribució de Dirichlet a partir de la clausura d'un vector aleatori positiu  $\mathbf{w} = (w_1, \dots, w_D)$ , i.e.  $\mathbf{x} = \mathcal{C}(\mathbf{w})$ , amb les variables  $w_i$  independents amb distribució  $\text{Gamma}(\alpha_i, 1)$ .

Les principals limitacions d'una distribució Dirichlet ja van ser exposades a Aitchison (1986). Entre elles, cal destacar el fet que qualsevol parell de quocients fet amb quatre components diferents d'una distribució Dirichlet és independent. Això, si estem realitzant una anàlisi composicional basat amb logquocients resulta ser una forta assumpció.

### Distribució normal a $\mathcal{S}^D$

La distribució normal a  $\mathcal{S}^D$  ha rebut diferents noms en funció de les característiques en quant a la definició i la mesura considerada al definir-ne el càlcul de probabilitats.

A (Aitchison i Shen, 1980) s'introdueix la distribució com la distribució normal logística additiva, anomenada aln (de l'anglès *additive logistic normal*). La distribució es defineix fàcilment a partir de la distribució normal: diem que una composició aleatòria  $\mathbf{x}$  segueix una distribució *normal logística additiva* si el vector transformat  $\text{alr}(\mathbf{x})$  segueix una distribució  $\mathcal{N}_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

El principal avantatge que ofereix la distribució aln és que una variable aleatòria seguint una distribució aln seguirà seguint una distribució aln després d'aplicar-li les operacions de pertorbació i potència. Com a principal limitació, tenim que la distribució depèn de l'ordenació de les components, això és, si canviem la darrera component, podríem obtenir resultats diferents.

A (Mateu-Figueras, 2003; Mateu-Figueras *et al.*, 2013) es defineix la distribució com la distribució normal sobre  $\mathcal{S}^D$  introduïda. Al igual que la distribució aln també la podem definir a partir de la distribució normal. En aquest cas, direm que una composició aleatòria  $\mathbf{x}$  segueix distribució normal a  $\mathcal{S}^D$  si les coordenades isomètriques,  $\mathbf{h} = \text{ilr}(\mathbf{x})$ , segueixen una distribució  $\mathcal{N}_{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

Tot i que a  $\mathcal{S}^D$  la família de distribucions normals a  $\mathcal{S}^D$  coincideix amb la família de distribucions aln. La normal a  $\mathcal{S}^D$  a l'estar expressada respecte a una base ortonormal, farà que els resultats que obtinguem siguin invariants per permutacions.

## 3.2 Models de mixtura

Els models de mixtura, o simplement mixtures per breuetat, són distribucions de probabilitat obtingudes a partir d'altres distribucions de probabilitat. En general, podríem dir que una variable aleatòria  $\mathbf{X}$  prové d'una mixtura si la seva funció de densitat o funció de masses es pot escriure com

$$m(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\phi})g(\mathbf{z} \mid \boldsymbol{\theta})d\mathbf{z}, \quad (3.1)$$

on  $f(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\phi})$  és una certa funció de densitat o masses amb paràmetres  $\boldsymbol{\phi}(\mathbf{Z})$  i  $\mathbf{Z}$  és una variable aleatòria amb funció de densitat o masses  $g(\mathbf{z} \mid \boldsymbol{\theta})$  amb paràmetre  $\boldsymbol{\theta}$ . Respectivament,  $f$  i  $g$  són anomenades la funció nucli i

la funció de barreja del model de mixtura. També, a la literatura és comú referir-se als models de mixtures com distribucions compostes o de barreja.

Les mixtures han estat àmpliament estudiades a la literatura, les referències més bàsiques són els llibres de Everitt i Hand (1981), Maritz i Lwin (1989), Titterington *et al.* (1986), McLachlan i Basford (1988), Lindsay (1995), Böhning (1999), McLachlan i Peel (2000), Aitkin *et al.* (2009) i McNicholas (2016).

Si la variable aleatòria  $Z$  és discreta, sobre un conjunt de valors  $z_1, \dots, z_k$  amb probabilitats  $\boldsymbol{\theta} = (\pi_1, \dots, \pi_k)$ , aleshores diem que la distribució de  $\mathbf{X}$  és un *model de mixtura finita*. En aquest cas, la funció de barreja és la funció de masses de la distribució categòrica amb paràmetres  $\boldsymbol{\theta} = (\pi_1, \dots, \pi_k)$  i l'Equació 3.1 se sol escriure com

$$m(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{j=1}^k \pi_j f(\mathbf{x} \mid \mathbf{z}_j, \boldsymbol{\phi}). \quad (3.2)$$

En aquest escenari, la notació més usual és escriure  $\boldsymbol{\phi}_j := \boldsymbol{\phi}(\mathbf{z}_j)$ , i per tant, es sol escriure la funció de densitat o masses d'un model de mixtura finita com

$$m(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{j=1}^k \pi_j f(\mathbf{x} \mid \boldsymbol{\phi}_j). \quad (3.3)$$

El model de mixtura finita més comú és la mixtura finita de distribucions normals. Aquest és el cas particular en què  $\pi_1, \dots, \pi_k$  són els paràmetres de la distribució categòrica i  $f$  és la funció de densitat d'una distribució normal,  $\mathcal{N}$ , i  $\boldsymbol{\phi}_j$  són els paràmetres  $\boldsymbol{\mu}_j$  i  $\Sigma_j$  de la distribució normal. En aquest cas particular, podem escriure l'Equació 3.2 com

$$m(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{j=1}^k \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j)$$

on  $\boldsymbol{\phi} = (\boldsymbol{\mu}_1, \Sigma_1, \dots, \boldsymbol{\mu}_k, \Sigma_k)$ . A la Figura 3.9 podem veure diferents exemples extrets de McLachlan i Basford (1988). Aquests exemples mostren com les mixtures de distribucions normals permeten la modelització de moltes característiques que podríem trobar en una mostra qualsevol: biaix, bimodalitat, curtosis, etc. Els quatre exemples mostrats s'han construït amb les mixtures:

- Unimodal amb biaix:  $\frac{1}{5}N(0, 1) + \frac{1}{5}N(\frac{1}{2}, \frac{2}{3}) + \frac{3}{5}N(\frac{13}{15}, \frac{5}{9})$ ,
- Fort biaix:  $\sum_{i=0}^7 \frac{1}{8}N(3\{(\frac{2}{3})^i - 1\}, (\frac{2}{3})^i)$ ,

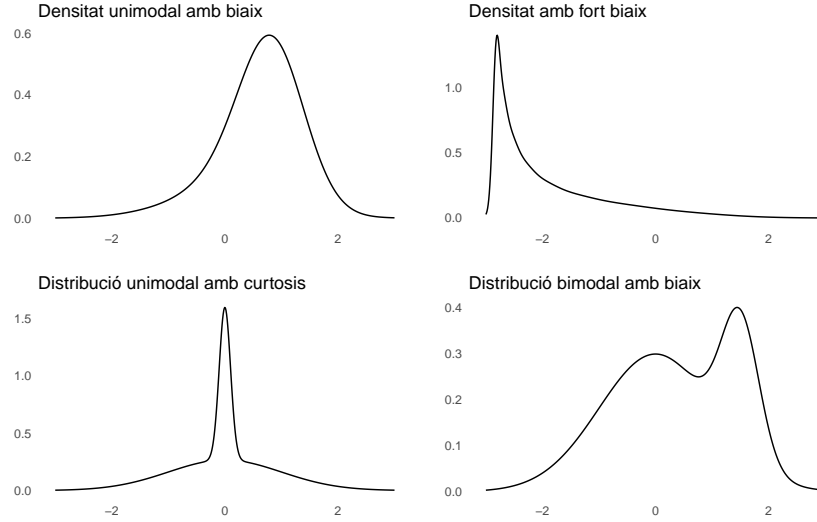


Figura 3.9: Gràfica de la funció de distribució d'algunes mixtures finites de distribucions normals.

- Unimodal amb curtosis:  $\frac{2}{3}N(0, 1) + \frac{1}{3}N(0, \frac{1}{10})$  i
- Bimodal amb biaix:  $\frac{3}{4}N(0, 1) + \frac{1}{4}N(\frac{3}{2}, \frac{1}{3})$

### 3.2.1 Estimadors de màxima versemblança dels paràmetres d'un model de mixtura finita

Donada una mostra  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  independent i idènticament distribuïda segons un model de mixtura finita, la seva versemblança vindrà donada per

$$L := L(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^k m(\mathbf{x}_i | \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^k \sum_{j=1}^k \pi_j f(\mathbf{x} | \boldsymbol{\phi}_j). \quad (3.4)$$

Aquesta funció mesura com de “creïbles” són els paràmetres  $\boldsymbol{\theta}$  i  $\boldsymbol{\phi}$  havent vist la mostra  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , com més gran sigui el valor  $L$  més “creïbles” seran els paràmetres. A la pràctica, en lloc d'optimitzar la funció  $L$ , s'optimitza el seu logaritme, anomenada la logversemblança, la qual vé donada per

$$\ell := \ell(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^k \log m(\mathbf{x}_i | \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^k \log \left( \sum_{j=1}^k \pi_j f(\mathbf{x} | \boldsymbol{\phi}_j) \right). \quad (3.5)$$

El punt on la funció  $\ell$  és màxima s'anomena l'estimador de màxima versemblança (EMV) de  $\boldsymbol{\theta}$  i  $\boldsymbol{\phi}$ , denotats  $\hat{\boldsymbol{\theta}}$  i  $\hat{\boldsymbol{\phi}}$ , aquest EMV el podem trobar com a solució de les equacions

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbf{0} \quad \text{i} \quad \frac{\partial}{\partial \boldsymbol{\phi}} \ell(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbf{0}. \quad (3.6)$$

Les equacions 3.6 poden ser manipulades de forma que  $\hat{\boldsymbol{\theta}}$  i  $\hat{\boldsymbol{\phi}}$  han de complir

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^n \tau_j(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) \quad (3.7)$$

i

$$\sum_{j=1}^k \sum_{i=1}^n \tau_j(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) \frac{\partial \log f(\mathbf{x}_i \mid \hat{\boldsymbol{\phi}}_j)}{\partial \boldsymbol{\phi}} = \mathbf{0} \quad (3.8)$$

on

$$\tau_j(\mathbf{x}_i \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{\pi_j f(\mathbf{x}_i \mid \boldsymbol{\phi}_j)}{\sum_{h=1}^k \pi_h f(\mathbf{x}_i \mid \boldsymbol{\phi}_h)} \quad (3.9)$$

és la probabilitat a posteriori que l'observació  $x_i$  provingui de la component  $j$ -èssima de la mixtura.

Si ens fixem amb les equacions 3.7 i 3.8, aquestes suggereixen una aproximació iterativa per a l'obtenció dels paràmetres  $\hat{\boldsymbol{\theta}}$  i  $\hat{\boldsymbol{\phi}}$ . De fet, Hasselblad (1966, 1969), Wolfe (1965, 1967, 1970) i Day (1969) proposaren aquest mètode iteratiu en casos particulars. Ara, vist amb perspectiva, aquest mètode iteratiu es pot identificar com una aplicació directa de l'algoritme EM proposat per Dempster *et al.* (1977).

L'algoritme EM és un algoritme iteratiu que en cada iteració millora el valor de l'Equació 3.4 mitjançant l'estimació de nous valors pels paràmetres  $\boldsymbol{\theta}$  i  $\boldsymbol{\phi}$ . L'avantatge principal de l'algoritme EM és que optimitza la funció  $\ell(\boldsymbol{\theta}, \boldsymbol{\phi})$  a través de l'anomenada funció de logversemblança completa,  $\ell_c(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$ . Aquesta funció difereix de la funció  $\ell$  en el fet que s'ha pogut observar la variable  $z \in \{z_1, \dots, z_k\}$  i per tant, en molts casos resulta en una funció més senzilla per treballar. En aquest cas, podem veure que  $\ell_c$  es pot escriure com

$$\ell_c = \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}[z = z_j] (\log \pi_j + \log f(\mathbf{x} \mid \boldsymbol{\phi}_j)).$$

Bàsicament, l'algoritme EM itera de forma repetida dues etapes ben diferenciades anomenades: l'etapa del càlcul del valor esperat (pas E) i



l'etapa de maximització dels paràmetres (pas M). En el pas E es calculen les probabilitats a posteriori de l'equació 3.9 quan  $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$  i  $\boldsymbol{\phi} = \boldsymbol{\phi}^{(t)}$  obtenint la funció

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}) = \sum_{i=1}^n \sum_{j=1}^k \tau_j(\mathbf{x}_i | \boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}) (\log \pi_j + \log f(\mathbf{x} | \boldsymbol{\phi}_j)).$$

En el pas M es maximitza l'expressió  $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)})$  respecte dels paràmetres  $\boldsymbol{\theta}$  i  $\boldsymbol{\phi}$  per obtenir noves estimacions  $\boldsymbol{\theta}^{(t+1)}$  i  $\boldsymbol{\phi}^{(t+1)}$  d'aquests. Podem resumir les dues fases de l'algoritme EM com

**Fase del valor esperat** En aquesta fase es considera que els paràmetres  $\boldsymbol{\theta}^{(t)}$  i  $\boldsymbol{\phi}^{(t)}$  de la distribució són fixes. Amb els paràmetres fixats es calcula la funció

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}) = \mathbb{E}_{Z|\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}} [\ell_c(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) | \mathbf{x}_1, \dots, \mathbf{x}_n]. \quad (3.10)$$

**Fase de maximització** En aquesta fase es busca quins paràmetres  $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\phi}^{(t+1)})$  maximitzen l'equació 3.10, i.e.

$$(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\phi}^{(t+1)}) = \arg \max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}).$$

L'algoritme EM presenta algunes limitacions. La primera d'elles és que tot i que està demostrat que en cada iteració es millora el valor de l'Equació 3.5, en el cas de les mixtures finites *no està garantit que aquesta optimització acabi convergint a un màxim global*. O sigui, és molt normal que l'algoritme retorni els paràmetres que optimitzen l'Equació 3.5 en un entorn local del paràmetre estimat  $\hat{\boldsymbol{\theta}}$  i  $\hat{\boldsymbol{\phi}}$ . Una altra de les limitacions que presenta l'algoritme EM és la lentitud amb què molts cops es convergeix al màxim local. Finalment, un altre problema existent en els models de mixtura és l'aparició de singularitats provocant que l'Equació 3.5 no estigui acotada.

### 3.2.2 Classificació paramètrica

Tot i que és àmpliament acceptat que el primer treball que basa la classificació en models de mixtura gaussiana fou Wolfe (1963), cal dir que aquesta tesi es basà en el treball de Tiedeman (1955) el qual introduí les bases de la classificació a través de models de mixtura. Exactament a Tiedeman (1955) es diu

Considera  $G$  matrius d'observacions cadascuna generant una funció de densitat de la forma donada per [1]. Descarta el tipus d'identificació de cada conjunt d'observacions i tindràs una sèrie mixta amb una densitat d'una forma desconeguda.

On [1] era l'equació de la distribució gaussiana. Després seguí amb el que ja coneixem com el problema de la classificació basada en models de mixtura

... resoldre el problema de reconstruir les  $G$  funcions de densitat dels tipus originals. Aquest és el problema; si és possible, la seva solució és extremadament difícil.

Això, portarà al llarg de les dues següents dècades als treballs Wolfe (1963, 1965, 1967) que resolen el problema en el cas gaussià. Tot i això, amb perspectiva, es podria dir que el primer treball que mostra l'enorme utilitat dels models de mixtura com a eina per a la classificació paramètrica fou McLachlan i Basford (1988), el qual mostrà una extensa varietat de dissenys experimentals. A partir d'aquest punt, els models de mixtura i la seva utilització com a principal eina d'agrupació basada en models es popularitzà. Banfield i Raftery (1993) mostraren com resoldre el problema de classificació basada en models de mixtura gaussianes multivariants en 8 descomposicions de la matriu de covariància. Celeux i Govaert (1995) ampliaren a 14 parametritzacions diferents. Finalment, fou amb la introducció del software MCLUST per S-Plus (Fraley i Raftery, 1999) que els models de mixtura es popularitzaren. Tal fou l'impacte, que durant molts anys el terme *agrupació basada en models* s'anomenava MCLUST (McNicholas, 2016). Actualment, per tal de modelar de forma més precisa els diferents grups, s'han introduït model de mixtures de diferents tipus; com els basats en la distribució  $t$ -multivariant, skew-normal multivariants o altres distribucions més generals (Lee i McLachlan, 2014; Andrews i McNicholas, 2012; Browne i McNicholas, 2015; Lee i McLachlan, 2013; Lin, 2010; Lee i McLachlan, 2011).

En aquest enfocament de classificació paramètrica s'assumeix que les dades provenen d'una mixtura amb funció de densitat donada per l'equació (3.3), on per defecte, es fixa el nombre de components  $k$ . Aleshores, cada una de les components s'assumeix que defineix un grup. Un cop estimats els paràmetres,  $\hat{\theta}$  i  $\hat{\phi}$ , s'assigna cada observació  $\mathbf{x}_i$  a aquella component  $j$  amb major probabilitat a posterior  $\tau_j(\mathbf{x}_i | \hat{\theta}, \hat{\phi})$ . Així doncs, en la classificació paramètrica a través de mixtures finites, podem distingir les següents tres etapes:

1. Donada la mostra  $\mathbf{x}_1, \dots, \mathbf{x}_n$  i el nombre de components  $k$ , s'estimen els paràmetres de màxima versemblança,  $\hat{\boldsymbol{\theta}}$  i  $\hat{\boldsymbol{\phi}}$ , de  $\boldsymbol{\theta}$  i  $\boldsymbol{\phi}$  respectivament.
2. Es calculen les probabilitats a posteriori

$$\tau_1(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}), \dots, \tau_k(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) \text{ per } i = 1 \dots n.$$

3. Cada observació  $\mathbf{x}_i$  és assignada a la component  $j$ -èssima amb

$$j = \arg \max_{h \in \{1, \dots, k\}} \tau_h(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}).$$

### 3.2.3 Combinació de components

En l'enfocament original de la classificació basada en models de mixtura comentats en la Secció 3.2.2, s'assumeix que una única distribució defineix un únic grup. Això que en molts casos pot tenir sentit, pot deixar de tenir sentit en el cas que algunes de les distribucions que conformen les components de la mixtura tinguin una localització semblant. Per exemple, a la Figura 3.10 es pot veure un escenari on el model de mixtura que millor s'ajusta a les dades, segons el criteri BIC, està format per *vuit* components (part superior de la figura). Utilitzant les aproximacions clàssiques dels models de mixtura a la classificació paramètrica, acabaríem concluint que existeixen 8 grups diferents. Veient la dispersió de les dades, potser semblaria més raonable considerar únicament 3 o 4 grups on el primer està modelat per una mixtura de tres components (grup groc), un segon grup format també per tres components (grup blau) i finalment dues components que podríem dubtar si conformen dos grups (verd i blau cel de la classificació amb quatre components) o un únic grup (color blau cel de la classificació amb 3 components).

La problemàtica descrita fou introduïda en els treballs Lee i Cho (2004); Goldberger i Roweis (2005); Li (2005); Ray i Lindsay (2005), els quals proposen una estructura jeràrquica dins les components de la mixtura. Hennig (2010) féu una revisió de les metodologies existents i proposà alguns nous mètodes. També, a Baudry *et al.* (2010) s'introduí un nou mètode basat únicament en les probabilitats a posteriori. Més recentment s'han proposat nous mètodes tots ells consistents en construir una jerarquia dins el conjunt de components de la mixtura Pastore i Tonellato (2013); Melnykov (2016, 2013); Longford i Bartosova (2014); Melnykov *et al.* (2012). En general es segueix l'esquema vist anteriorment per la classificació paramètrica, però en lloc del pas 3 es segueix de la següent manera:

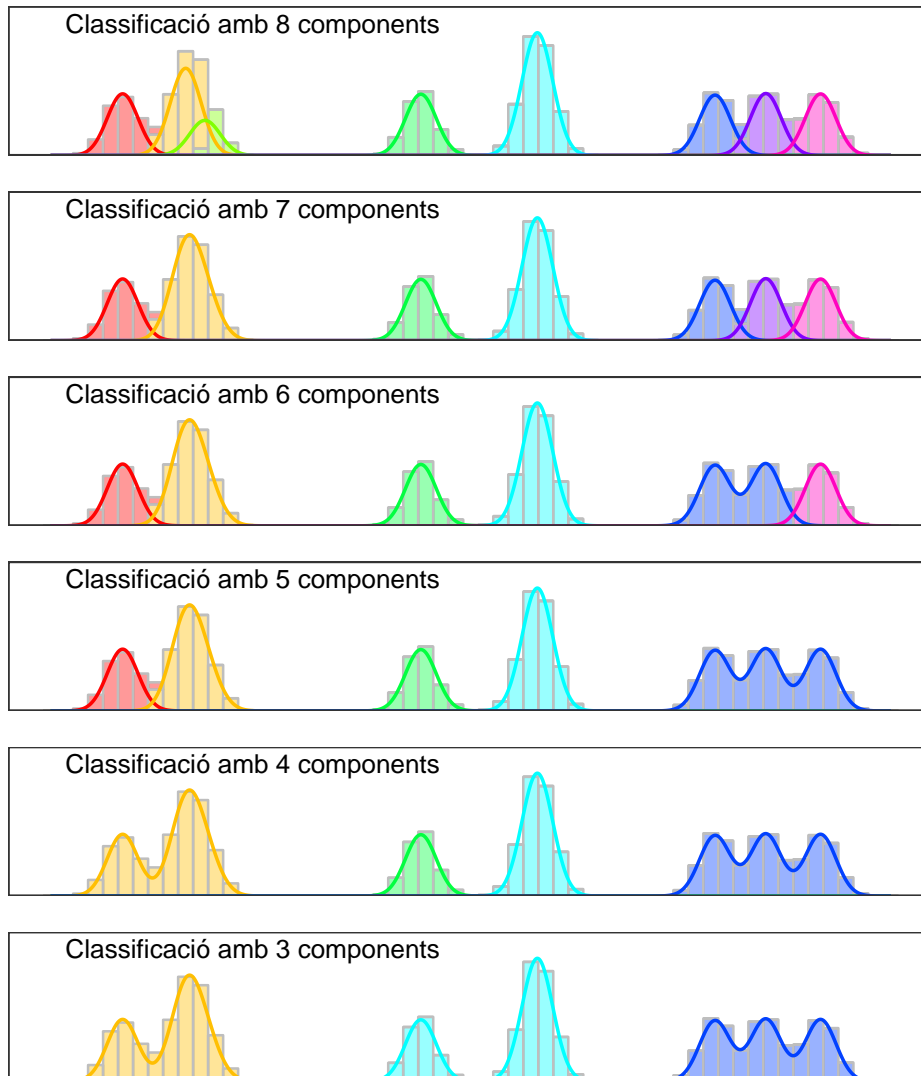


Figura 3.10: Exemple del procés de combinació de les components d'una mixtura de vuit components.

3. Si les components modelen un únic grup, salta al punt 5. Altrament, es combinen dues components  $a$  i  $b$  que estiguin modelant el mateix grup. Es deixen de considerar les components  $a$  i  $b$  separades i es consideren com una única component  $a \amalg b$ .
4. Per cada observació  $\mathbf{x}_i$  s'actualitza la probabilitat de pertànyer a la nova component  $f_{a \amalg b}$  com  $\tau_a(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) + \tau_b(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ . Es torna al punt 3.
5. Cada observació  $\mathbf{x}_i$  és assignada a la component  $j$ -èssima amb

$$j = \arg \max_{h \in \text{"Components"}} \tau_h(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}).$$

Dels diferents mètodes que podem trobar per combinar les components d'una mixtura, n'hi ha tres que es basen únicament amb el valor de les probabilitats a posteriori  $\tau_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$  vistes anteriorment. Baudry *et al.* (2010) proposa combinar de forma seqüencial aquelles dues components que minimitzin l'entropia total,

$$\text{Entropia total} = \sum_{i=1}^n \sum_{j=1}^k \tau_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$$

dels vectors de probabilitat a posterior resultant. Hennig (2010) proposa combinar de forma seqüencial aquelles dues components que maximitzin el que anomena *Probabilitats de classificació incorrecta estimades directament* (DEMP de l'anglès "Directly Estimated Misclassification Probabilities"). Finalment, Longford i Bartosova (2014) proposa combinar de forma seqüencial aquelles dues components que maximitzin la probabilitat de classificar una observació a una component sabent que s'ha generat d'una de les dues components.

Capítol 4

Articles

La Dra. Glòria Mateu Figueras, com a coautora dels articles següents:

- M.Comas-Cufí, J.A.Martín-Fernández i G.Mateu-Figueras. Log-ratio methods in mixture models for compositional data sets. *Statistics and Operations Research Transactions*. 40(2):349- 374 (2016).
- M.Comas-Cufí, J.A.Martín-Fernández i G.Mateu-Figueras. Merging the components of a finite mixture using posterior probabilities. *Statistical Modelling*. A impremta (2017).
- M.Comas-Cufí, J.A.Martín-Fernández, G.Mateu-Figueras i J.Palarea-Albaladejo. Modelling count data using the logratio-normal-multinomial distribution. Enviat a *Computational Statistics & Data Analysis* (2018).

Accepto que el Sr. Marc Comas Cufí presenti els articles esmentats com a autor principal i com a part de la seva tesi doctoral, i que aquests articles no puguin, per tant, formar part de cap altra tesi doctoral.

I perquè així consti i tingui els efectes oportuns, signo aquest document.

Signatura,



Glòria Mateu Figueras

Girona, 20 d'abril de 2018.

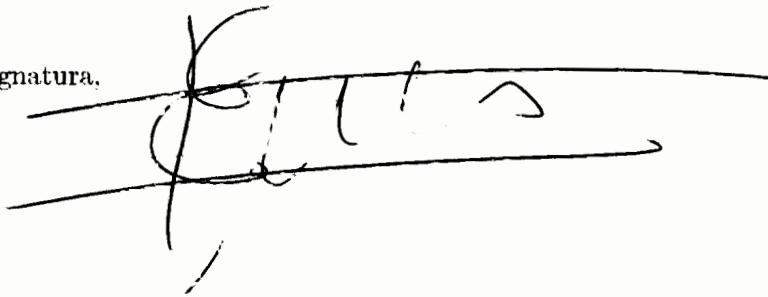
El Dr. Josep Antoni Martín Fernández, com a coautor dels articles següents:

- M.Comas-Cufí, J.A.Martín-Fernández i G.Mateu-Figueras. Log-ratio methods in mixture models for compositional data sets. *Statistics and Operations Research Transactions*. 40(2):349-374 (2016).
- M.Comas-Cufí, J.A.Martín-Fernández i G.Mateu-Figueras. Merging the components of a finite mixture using posterior probabilities. *Statistical Modelling*. A impremta (2017).
- M.Comas-Cufí, J.A.Martín-Fernández, G.Mateu-Figueras i J.Palarea-Albaladejo. Modelling count data using the logratio-normal-multinomial distribution. Enviat a *Computational Statistics & Data Analysis* (2018).

Accepto que el Sr. Marc Comas Cufí presenti els articles esmentats com a autor principal i com a part de la seva tesi doctoral, i que aquests articles no puguin, per tant, formar part de cap altra tesi doctoral.

I perquè així consti i tingui els efectes oportuns, signo aquest document.

Signatura,

A handwritten signature in black ink, appearing to read 'J. Martín Fernández', is written over two horizontal lines. The signature is stylized and somewhat cursive.

Josep Antoni Martín Fernández

Girona, 20 d'abril de 2018.





El Dr. Javier Palarea Albaladejo, com a coautor de l'article següent:

- M.Comas-Cufí, J.A.Martín-Fernández, G.Mateu-Figueras i J.Palarea-Albaladejo. Modelling count data using the logratio-normal-multinomial distribution. Enviat a *Computational Statistics & Data Analysis* (2018).

Accepto que el Sr. Marc Comas Cufí presenti l'article esmentat com a autor principal i com a part de la seva tesi doctoral, i que aquest article no pugui, per tant, formar part de cap altra tesi doctoral.

I perquè així consti i tingui els efectes oportuns, signo aquest document.

Signatura,

Javier Palarea Albaladejo

Girona, 23 d'abril de 2018.

## **4.1 Statistics and Operations Research Transactions**

Aquest primer article cobreix els objectius Obj. 1 Obj. A i Obj. B descrits a la secció 2.1. En resum es proposa la metodologia logquocient per a la construcció de mixtures de distribucions ben definides dins l'espai del Símplex. També, es fa una descripció dels desavantatges d'utilitzar altres metodologies actualment existents.

L'article ha estat publicat a la revista Statistics and Operations Research Transactions.

Volum: 40 , Número: 2 , Pàgines: 349-374 , Enviat: Febrer 2016,

Acceptat: Octubre 2016

DOI: 10.2436/20.8080.02.47

Factor d'impacte: 1.333 (Q2).



# 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

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

where  $f_1, \dots, f_k$  are pdf's of distributions with parameters  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$  respectively, and  $\pi_1, \dots, \pi_k$  are positive numbers with  $\sum_{i=1}^k \pi_i = 1$  (McLachlan and Peel, 2000). The pdfs  $f_1, \dots, f_k$  are typically called *mixture 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  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$  belonging to a unique set  $\Theta$ .

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

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  $\mathcal{S}^D$ , 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  $\mathbf{x} = (x_1, \dots, x_D)$  with all its parts strictly positive and carrying only relative information. A *D*-part composition is usually restricted to sum to a fixed constant  $\kappa$ , i.e.

$$\sum_{i=1}^D x_i = \kappa. \quad (2)$$

As a convention, it is usual to assume  $\kappa = 1$  for proportions and  $\kappa = 100$  for percentages. Because the value of  $\kappa$  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](http://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  $\mathcal{S}^D$ : *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  $\alpha$  as:

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

where  $C(\mathbf{x}) = \frac{\kappa}{\sum x_k}(x_1, \dots, x_D)$  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

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i < j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \quad (4)$$

provides  $\mathcal{S}^D$  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  $\mathcal{S}^D$  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  $\mathcal{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}\}$  is fixed, any  $D$ -part composition  $\mathbf{x}$  can be expressed as the linear combination

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

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

$$h_i = \sqrt{\frac{i}{i+1}} \ln \sqrt{\frac{\prod_{j=1}^i x_j}{x_{i+1}}}, \quad i = 1, \dots, D-1, \quad (5)$$

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

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

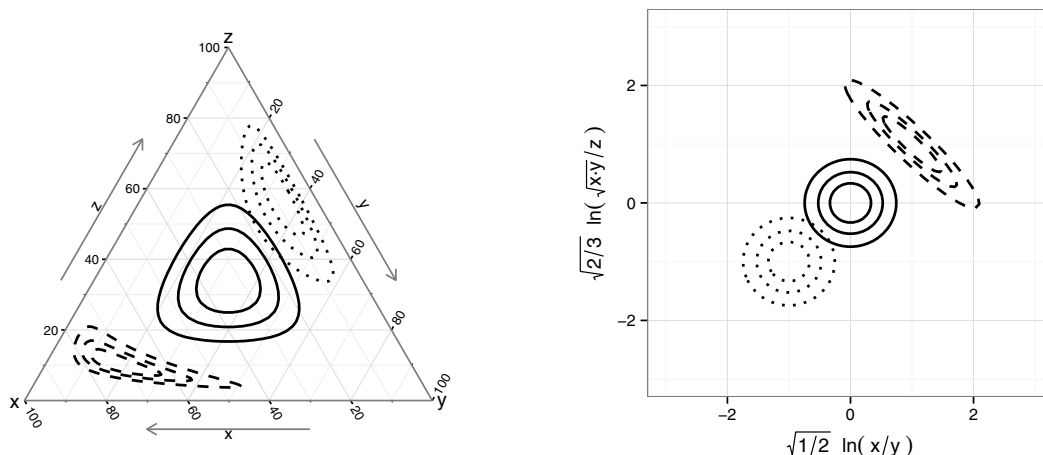
In this paper we use the coordinates in Equation 5 but any other orthonormal basis can also be considered. Determining which basis or coordinates are the most appropriate to solve a specific problem, is not straightforward. Nevertheless, the sequential binary partition introduced by Egozcue and Pawłowsky (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  $f^*(\cdot; \boldsymbol{\theta}) : \mathbb{R}^{D-1} \rightarrow \mathbb{R}^+$  be a pdf defined on real space with parameters  $\boldsymbol{\theta}$ . Then,  $f_{\mathcal{B}}(\mathbf{x}; \boldsymbol{\theta}) = f^*(\mathbf{h}_{\mathcal{B}}(\mathbf{x}); \boldsymbol{\theta})$  defines a pdf on the simplex,  $f_{\mathcal{B}}(\cdot; \boldsymbol{\theta}) : \mathcal{S}^D \rightarrow \mathbb{R}^+$ , with respect to the Aitchison measure on  $\mathcal{S}^D$ . For example, fixing an orthonormal basis  $\mathcal{B}$ , the log-ratio normal distribution with parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  is defined as

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

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  $\mathcal{S}^D$ , 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  $\mathcal{S}^3$ . 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

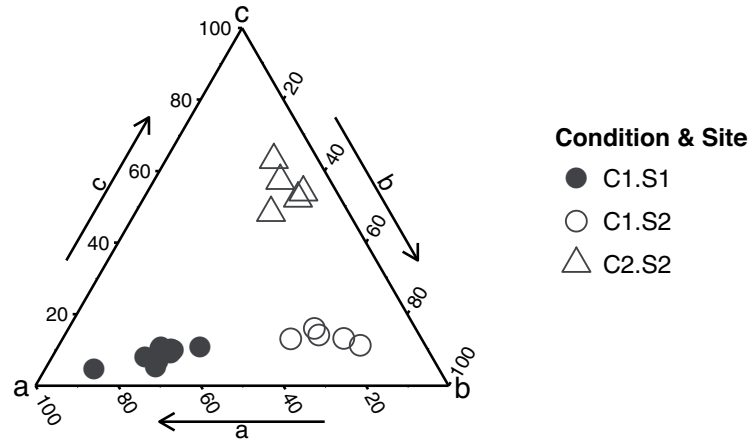
This approach assumes that  $\mathcal{S}^D$  is a subset of  $\mathbb{R}^D$  and its particular Euclidean space structure described in Section 2 is ignored. It is assumed that compositions are generated from a finite mixture distributions with pdf given by Equation 1 where  $f(\cdot; \theta_i) : \mathbb{R}^D \rightarrow \mathbb{R}^+$  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.  $P(\{\mathbf{x} \in \mathcal{S}^D | x_i < 0\}) > 0$ . 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.

<b>a</b>	<b>b</b>	<b>c</b>	$h_1$	$h_2$	<b>site</b>	<b>condition</b>
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

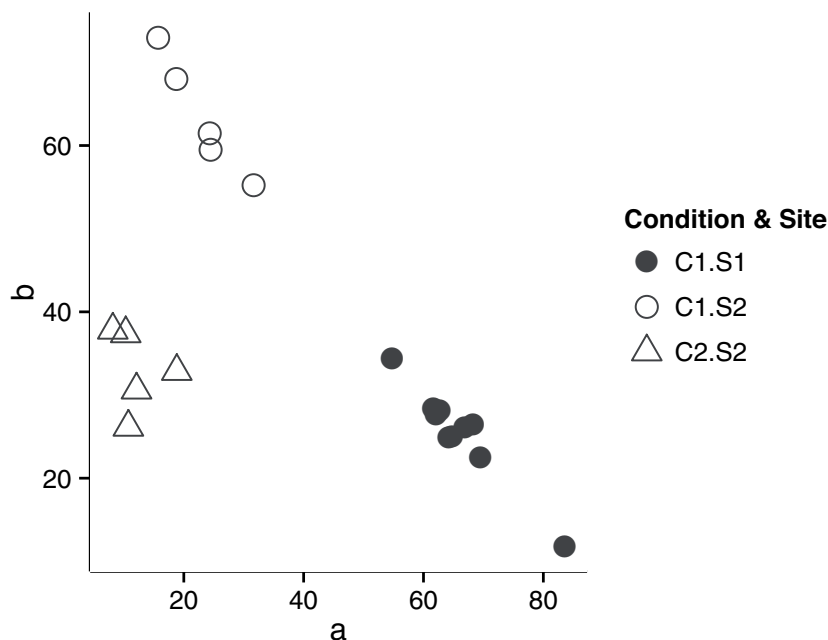
In addition, this approach defined on the real space also ignores the constant sum constraint. Therefore, a further limitation is the collinearity that appears between parts after restricting the parts to sum a constant (Equation 2). This collinearity implies that the covariance matrix is singular, and 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., Papa-georgiou 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 misleading 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  $f(\cdot; \boldsymbol{\theta}) : \mathcal{S}^D \rightarrow \mathbb{R}^+$ , is a pdf defined on the simplex. The Dirichlet distribution has been traditionally used as the probability distribution on  $\mathcal{S}^D$ . 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 greater 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  $\mathcal{S}^D$  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

$$\pi_1 f_{\mathcal{B}}(\cdot; \boldsymbol{\theta}_1) + \cdots + \pi_k f_{\mathcal{B}}(\cdot; \boldsymbol{\theta}_k) \quad (7)$$

where  $f_{\mathcal{B}}(\mathbf{x}; \boldsymbol{\theta}_i)$  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

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

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

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

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

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

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

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

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

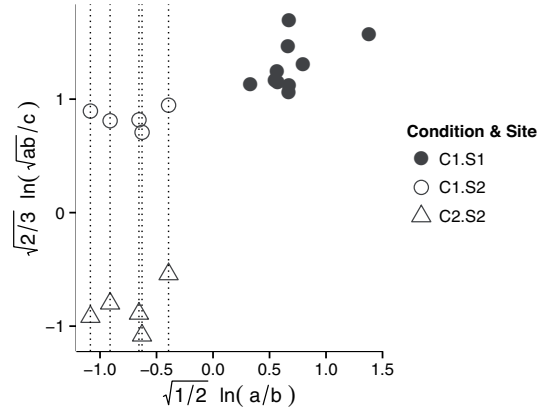
$$= \arg \max_{\pi_1, \dots, \pi_k, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k} \prod_{i=1}^n \sum_{j=1}^k \pi_j f^*(\mathbf{h}_{\mathcal{B}}(\mathbf{x}_i); \boldsymbol{\theta}_j). \tag{11}$$

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  $\mathcal{B}$  of  $\mathcal{S}^3$  is selected, for example

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

and the compositions of  $\mathbf{X}$  are expressed in terms of their coordinates  $\mathbf{X}^T$  ( $h_1 = \sqrt{1/2} \ln(a/b)$  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  $C_2$ .

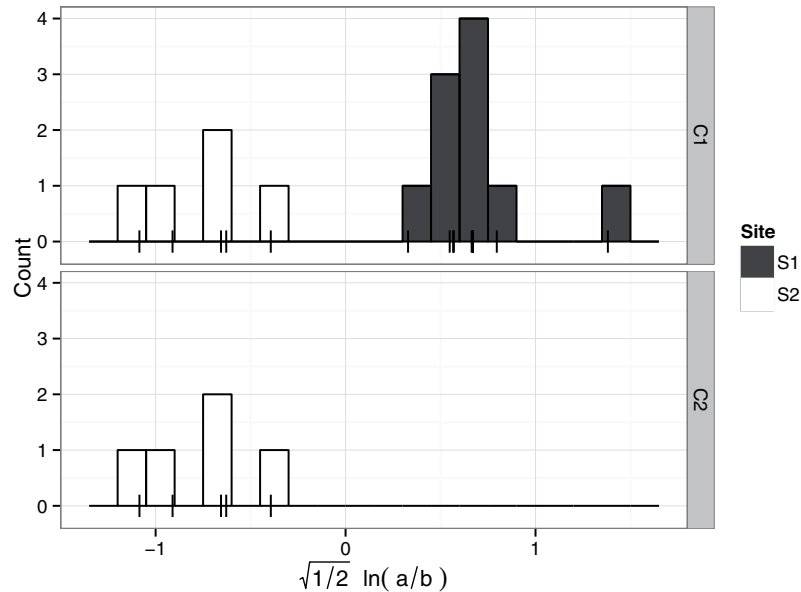
In this case the interpretations based only in terms of parts  $a$  and  $b$  will not be misleading. In fact, if the analyst also decides to remove part  $c$ , a basis  $\mathcal{B}'$  of  $\mathcal{S}^2$  is selected as:



**Figure 4:** Scatterplot of log-ratio coordinates for the 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  $C_2$ .

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

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  $\mathbf{X}$ . Two stratas correspond to weather conditions.

In Equations 9 and 10, we fit the mixture using the coordinates  $\mathbf{h}_{\mathcal{B}}(\mathbf{x})$  with respect to a specific basis  $\mathcal{B}$  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  $\mathcal{B}_1$  and  $\mathcal{B}_2$  be two basis on  $\mathcal{S}^D$ . Let  $\Theta$  be a parameter space for a probability density function  $f^* : \mathbb{R}^{D-1} \rightarrow \mathbb{R}^+$ . A probability density function  $f^*$  is basis invariant if for any two different basis  $\mathcal{B}_1, \mathcal{B}_2$ , for any parameters  $\theta_1 \in \Theta$ , there are parameters  $\theta_2 \in \Theta$  such that

$$f^*(\mathbf{h}_{\mathcal{B}_1}(\mathbf{x}); \theta_1) = f^*(\mathbf{h}_{\mathcal{B}_2}(\mathbf{x}); \theta_2).$$

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 bivariate 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.

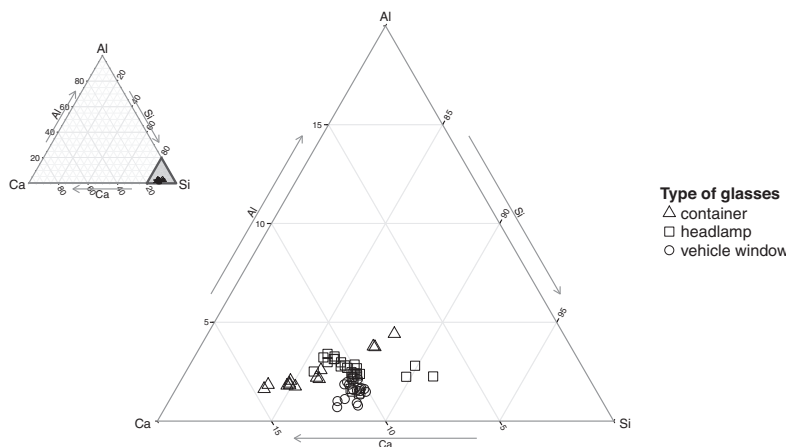


Figure 6: Reduced Forensic Glass data set in ternary diagram: Calcium (Ca), Silica (Si) and Aluminium (Al) chemical elements. Three groups of glass: containers (circles), headlamps (triangles) and vehicle windows (squares). The large ternary diagram is a zoom of the shadow area seen in the smaller initial 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 ( $\text{BIC} = -763.4$ ) 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

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

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

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

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 where 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.

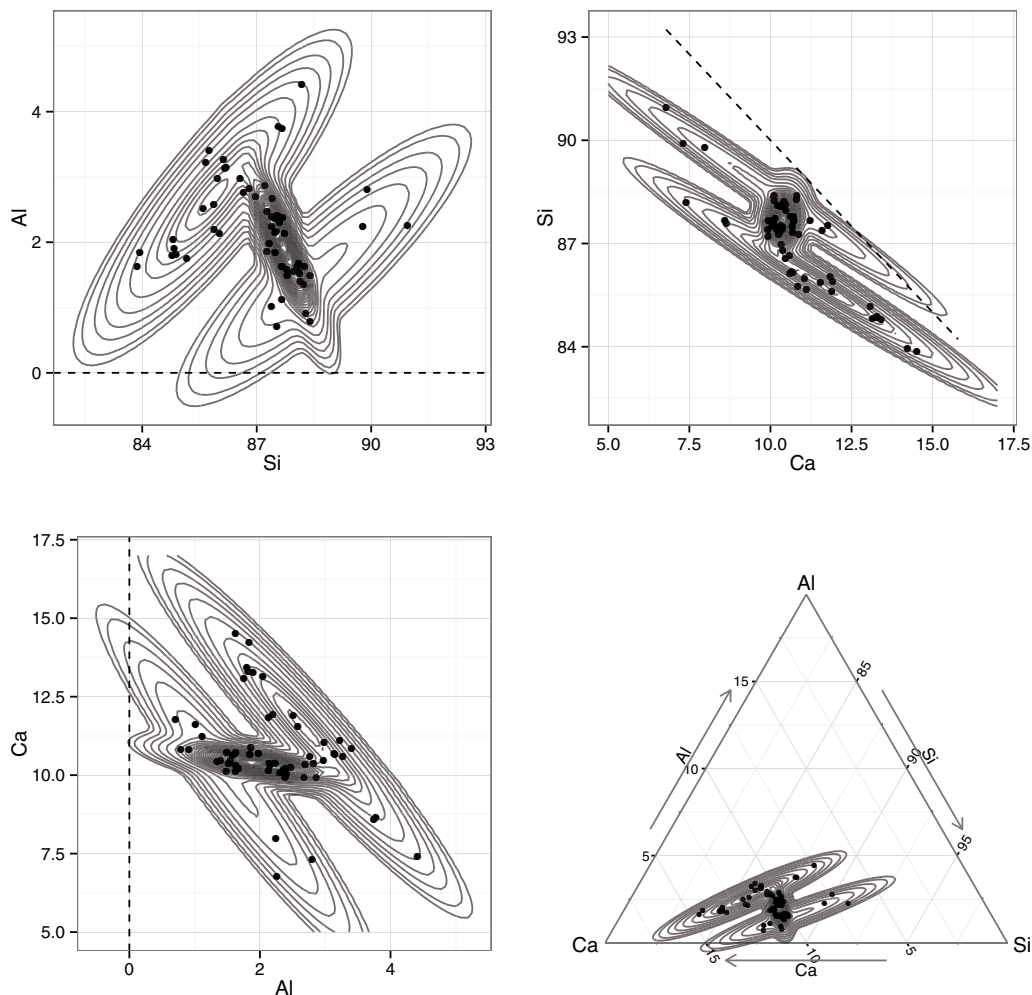
Despite the fact that in Gaussian mixtures the maximum likelihood function is invariant whatever part is removed, we stated that in practice the numerical algorithm gets stuck in a local optimum. That is, the invariance of the results is not guaranteed, and different mixtures may be obtained depending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

A Dirichlet probability distribution is specified by the parameters  $\boldsymbol{\alpha} = (\alpha^1, \dots, \alpha^D)$ . Therefore, to fit a mixture of  $K$  Dirichlet distributions the parameters  $\pi_1, \dots, \pi_K$  and  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K$  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 ( $\text{BIC} = -732.9$ ) was  $\pi_1 f(\cdot; \boldsymbol{\alpha}_1) + \pi_2 f(\cdot; \boldsymbol{\alpha}_2) + \pi_3 f(\cdot; \boldsymbol{\alpha}_3)$  with estimates

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

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

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

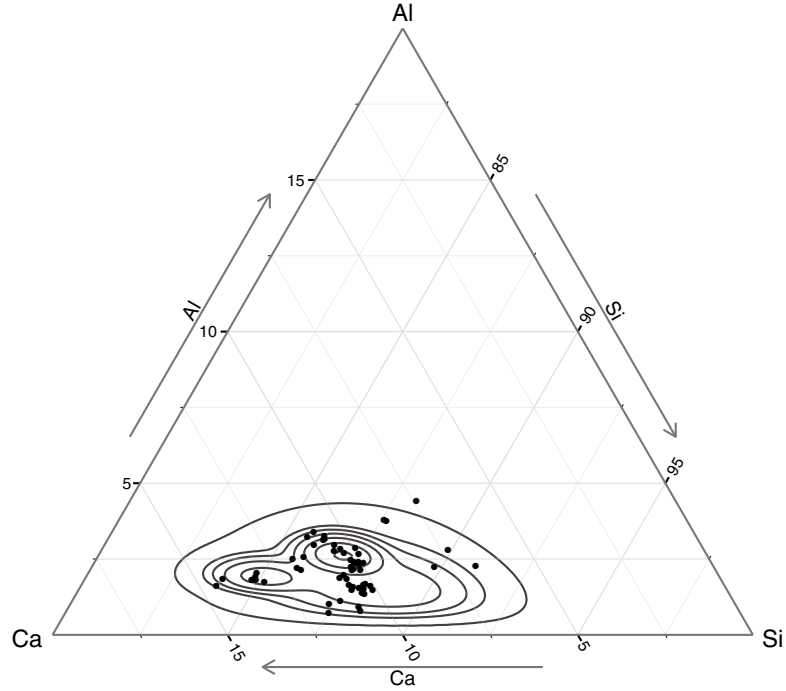


**Figure 7:** Reduced Forensic Glass data set. On the top-left, top-right and bottom-left isodensity curves for mixtures of Gaussian distributions in  $R^2$  after removing the Ca, the Al and the Si part respectively. On bottom-right the isodensity curves transformed into the simplex.

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

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

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

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

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

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

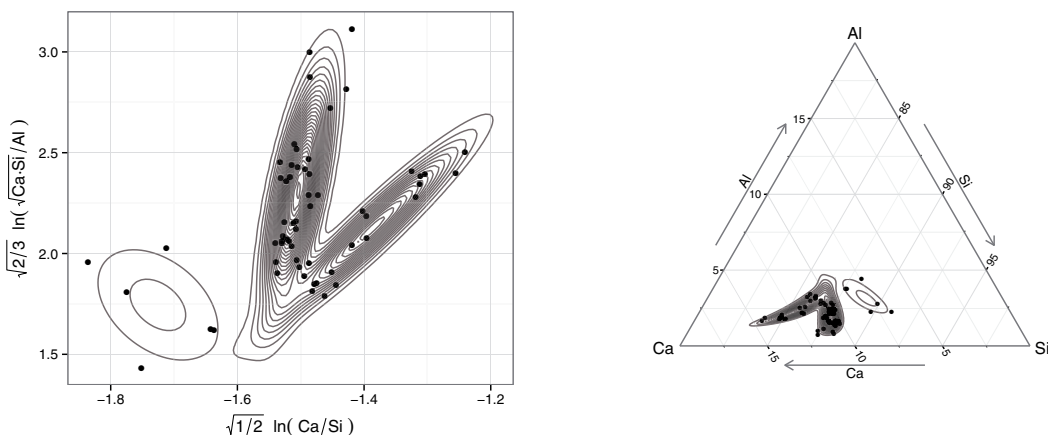


Figure 9: Log-ratio Gaussian mixtures for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.

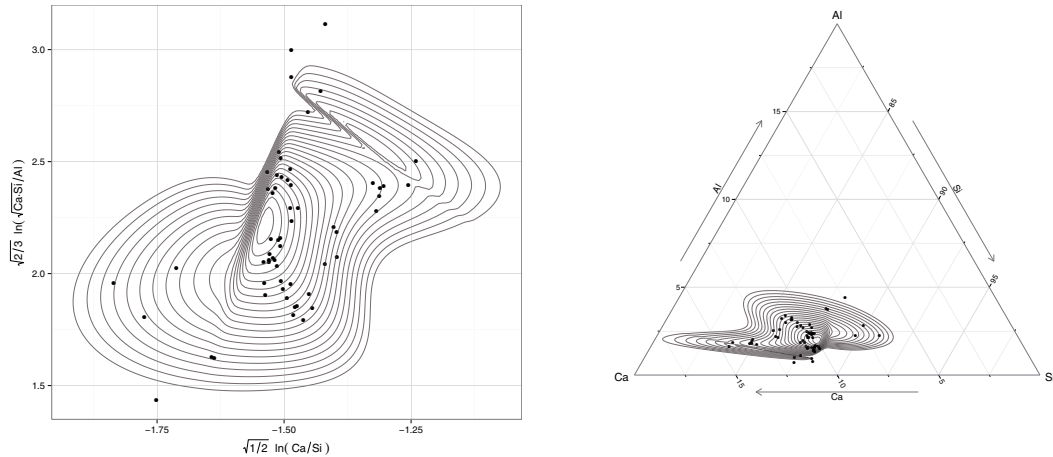
the space of coordinates (left) and in the ternary diagram (right). Looking at the coordinate space, we see that this mixture can model elliptical forms of variability and consequently, on the simplex the estimated mixture is able to model those typical arc shaped forms (Figure 9 (right)). Because multivariate 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  $\mu_i$  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  $\Sigma_i$ , 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



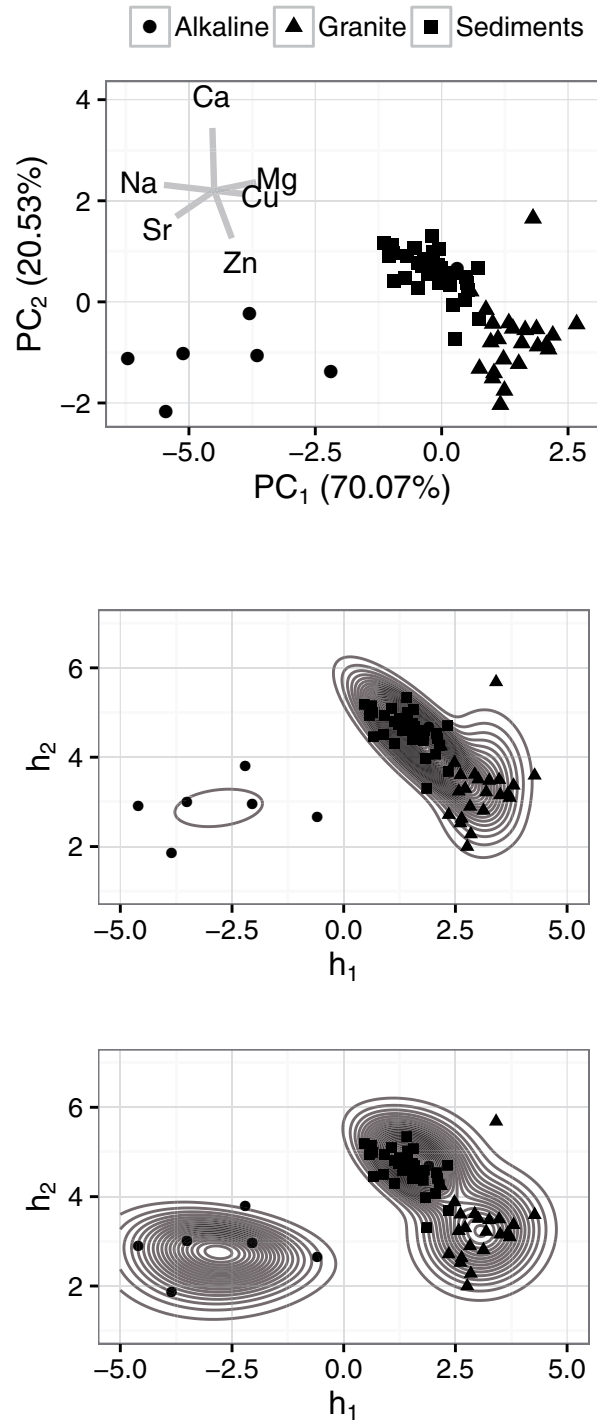
**Figure 10:** Log-ratio skew normal mixture adjusted for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.

Silica elements. Due to  $\text{var}(\ln(\text{Ca}/\text{Si}))$  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  $\ln(\text{Si}/\text{Al})$ .

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) ( $\text{BIC} = -62.3$ ). 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  $\mathcal{B}$  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 significantly 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 latent classes. *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*, *Wiley 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 Martín-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.

## 4.2 Statistical Modelling

El segon article està relacionat amb tots els objectius descrits a la secció 2.1. En aquest article es proposa un enfocament general que engloba tots els mètodes existents basats en probabilitats a posterior per la combinació de les components d'una mixtura. A més, es proposen un seguit de mètodes que nous basats en mesures en el Símplex que respecten l'escala i són composicionalment coherents.

L'article ha estat publicat a la revista *Statistical Modelling*.

Enviat: Agost 2016, Acceptat: Agost 2017

DOI: 10.1177/1471082X17735919

Factor d'impacte: 0.932 (Q2).



## Merging the components of a finite mixture using posterior probabilities

Marc Comas-Cufí<sup>1</sup>, Josep A. Martín-Fernández<sup>1</sup> and Glòria Mateu-Figueras<sup>1</sup>

<sup>1</sup>Department of Computer Science, Applied Mathematics and Statistics, Polytechnic School, University of Girona, Spain.

**Abstract:** Methods in parametric cluster analysis commonly assume data can be modelled by means of a finite mixture of distributions. However, associating each mixture component to one cluster is frequently misleading because different mixture components can overlap, and then, associated clusters can overlap too suggesting a unique cluster. A number of approaches have already been proposed to construct the clusters by merging components using the posterior probabilities. This article presents a generic approach for building a hierarchy of mixture components that integrates and generalizes some techniques proposed earlier in the literature. Using this proposal, two new techniques based on the log-ratio of posterior probabilities are introduced. Moreover, to decide the final number of clusters, two new methods are presented. Simulated and real datasets are used to illustrate this methodology.

**Key words:** hierarchical clustering, log-ratio, merging components, mixture model, model-based clustering, simplex

Received August 2016; revised March 2017; accepted August 2017

### 1 Introduction

A common approach in parametric cluster analysis assumes data can be modelled by means of a ‘finite mixture of distributions’ (Fraley and Raftery, 2002; Punzo, 2014; Comas-Cufí et al., 2016), also called a ‘finite mixture model’ (FMM). An FMM is a probability distribution whose probability density function (pdf) can be expressed as a convex combination of pdf from other distributions with same domain  $\mathbb{X}$ . More precisely, the pdf  $f$  of an FMM can be expressed as

$$f(\cdot; \pi_1, \dots, \pi_K, \Theta) = \pi_1 f_1(\cdot; \theta_1) + \dots + \pi_K f_K(\cdot; \theta_K), \quad (1.1)$$

where  $\Theta = \{\theta_1, \dots, \theta_K\}$ ,  $\theta_j$  are the parameters of pdf  $f_j$ ,  $1 \leq j \leq K$ , and  $\pi_j$  is the ‘weight’ of component  $f_j$ . Restriction  $\sum_{\ell=1}^K \pi_\ell = 1$  guarantees that  $\int_{\mathbb{X}} f = 1$ . Originally, the clustering algorithm based on an FMM follows two steps:

---

Address for correspondence: Josep A. Martín-Fernández, Department of Computer Science, Applied Mathematics and Statistics, University of Girona, P-IV, Campus Montilivi, E-17003 Girona, Spain.  
E-mail: josepantoni.martin@udg.edu



2 *Marc Comas-Cufí et al.*

1. to calculate estimates  $\hat{\pi}_1, \dots, \hat{\pi}_K$  and  $\hat{\Theta}$  of parameters  $\pi_1, \dots, \pi_K$  and  $\Theta$ , using a sample  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and
2. to classify each observation  $\mathbf{x}_i \in \mathbf{X}$  to a cluster  $c$ ,  $1 \leq c \leq K$ , according to the criterium of maximizing the posterior probability

$$\hat{\tau}_{ij} = \frac{\hat{\pi}_j f_j(\mathbf{x}_i; \hat{\theta}_j)}{\sum_{\ell=1}^K \hat{\pi}_\ell f_\ell(\mathbf{x}_i; \hat{\theta}_\ell)},$$

that is, one observation  $\mathbf{x}_i$  is classified to cluster  $c$ , if

$$c = \arg \max_{j=1, \dots, K} \hat{\tau}_{ij}. \quad (1.2)$$

Note that, in this process, the number of clusters is equal to the number of mixture components. The different approaches to decide the number of components  $K$  are reviewed in McLachlan and Rathnayake (2014). In our work, we use the Bayesian Information Criterion (BIC) because under certain regularity conditions, it estimates consistently the number of mixture components (Keribin, 1998, 2000). In addition, BIC is effective as a model selection criteria on a practical level (Fraley and Raftery, 1998).

Because a cluster is formed by similar observations (Melnykov, 2016), different components can be modelling one single cluster. Lee and Cho (2004), Hennig (2010), Baudry et al. (2010), Melnykov (2013), Pastore and Tonellato (2013) and Melnykov (2016) propose separating the concepts of cluster and mixture component. The authors show that associating each mixture component to one cluster can be misleading because different mixture components are frequently so overlapped that they can in fact be modelling a unique cluster. In other words, one cluster could be modelled by the distribution resulting from the merging of two or more mixture components. According to this approach, the FMM clustering algorithm is completed with the following third step:

3. to analyse which of the  $K$  mixture components should be merged to model  $k$  clusters,  $k \leq K$ .

The crux of this new step is how to decide which components have to be merged. Importantly, the underlying finite mixture does not change after two components have been merged. Therefore, for a given sample, the likelihood function remains invariant. This fact makes it impossible to decide which components are modelling a single cluster in terms of the likelihood or through the BIC criteria, and then, there is a subjective component in this decision (Hennig, 2010).

To guide the decision of merging components, different approaches have been proposed. Some methods are based on the modality of the resulting distribution (Ray and Lindsay, 2005), other methods are based on the parameters of the FMM (Bhattacharyya distance method in Hennig (2010) or the DEMP+ in Melnykov

*Merging the components of a finite mixture using posterior probabilities* 3

(2016)), and others are based on the estimated posterior probabilities obtained after fitting the FMM (an entropy difference approach in Baudry et al. (2010), the DEMP approach by Hennig (2010) or Longford and Bartošová (2014)).

Our approach is based on the posterior probabilities which are obtained after adjusting an FMM. We introduce a generic expression that integrates and generalizes methods given by Baudry et al. (2010), Hennig (2010) and Longford and Bartošová (2014). In addition, with this new approach, the analyst can define their own method, for example, based on log-ratio transformations of posterior probabilities (Aitchison, 1986). Using this generic approach for merging components, one can also build a hierarchy over the set of mixture components. On the first level of this hierarchy, we consider  $K$  clusters where each cluster is modelled by one component, and on the second level, we have  $K - 1$  clusters where one cluster is modelled by two components. On the successive levels, subsequent clusters are modelled by merging different components. The final level contains one cluster (the original sample) modelled by the entire mixture.

The article is organized as follows. The definitions and notation required throughout this article are given in Section 2. Section 3 presents the general merging criteria, and shows that the most important techniques from literature can be condensed into this new approach. Using this proposal, Section 4 presents a new family of techniques based on the log-ratio methodology. In Section 5, two heuristic methods to decide the final number of clusters are proposed. Section 6 includes two examples to illustrate the algorithm with different types of mixture distributions and one simulated example considering a range of situations. To conclude, final remarks are made in Section 7. The programming of the data analyses discussed in this work has been conducted using the open-source R statistical environment (R Development Core Team, 2015). The corresponding computer routines implementing the methods and the datasets can be obtained from the website <http://www.compositionaldata.com> and from the R packages ‘mixpack’ and ‘zCompositions’ (R Development Core Team, 2015).

## 2 Definitions and notation

Let  $\mathcal{I}^K = \{1, \dots, K\}$  be a set of natural numbers to indicate the components of an FMM. A ‘partition’ of  $\mathcal{I}^K$  of size  $k$ , denoted  $\mathcal{P}_k$ , is a collection of  $k$  subsets  $I_1, \dots, I_k$  of  $\mathcal{I}^K$ , called parts, such that  $\bigcup_{j=1}^k I_j = \mathcal{I}^K$ , and for any two parts  $I_a, I_b \in \mathcal{P}_k$  with  $a \neq b$ ,  $I_a \cap I_b = \emptyset$  holds. For example, given  $\mathcal{I}^6 = \{1, 2, 3, 4, 5, 6\}$ , we could form  $\mathcal{P}_4 = \{\{1, 2\}, \{3\}, \{4\}, \{5, 6\}\}$  or  $\mathcal{P}_4 = \{\{1, 2, 3\}, \{4\}, \{5\}, \{6\}\}$  (see Section 6.2).

Given a partition  $\mathcal{P}_k$ , the pdf  $f$  of an FMM (Equation 1.1) can be written as

$$f = \pi_{I_1} f_{I_1}(\cdot; \Theta) + \dots + \pi_{I_k} f_{I_k}(\cdot; \Theta), \quad (2.1)$$

where  $f_{I_j}(\cdot; \Theta) = \sum_{\ell \in I_j} \frac{\pi_\ell}{\pi_{I_j}} f_j(\cdot; \theta_j)$  and  $\pi_{I_j} = \sum_{\ell \in I_j} \pi_\ell$ . Note that using this notation, each  $f_{I_j}(\cdot; \Theta)$  is also an FMM. Because each part  $I_j$  defines a single component  $f_{I_j}$ , when

## 4 Marc Comas-Cufí et al.

there is no confusion, we use  $I_j$  referring either to the part  $I_j$  or to the component  $f_{I_j}$ . Note that, given  $k$  components of an FMM  $f$ , there is  $B_k$  different ways to express the mixture  $f$  in terms of a partition.<sup>1</sup>

A *hierarchical sequence of partitions of  $\mathcal{I}^K$*  is a sequence  $\mathcal{P}_1, \dots, \mathcal{P}_K$ , verifying that

- $\mathcal{P}_1$  is the one-part partition  $\mathcal{P}_1 = \{\mathcal{I}^K\}$ ;
- if a part is  $I_j \in \mathcal{P}_{k-1}$ , then either there is a part  $I_a \in \mathcal{P}_k$  with  $I_j = I_a$  or there are two parts  $I_a, I_b \in \mathcal{P}_k$  with  $I_j = I_a \cup I_b$ ; and
- $\mathcal{P}_K = \{\{1\}, \{2\}, \dots, \{K\}\}$ .

One can extend Equation (1.2) in terms of partitions. Indeed, let  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  be a sample formed by observations of  $\mathbb{X}$ . Given a partition  $\mathcal{P}_k = \{I_1, \dots, I_k\}$ , we define the posterior probability of  $\mathbf{x}_i$  being classified to part  $I_j \in \mathcal{P}_k$  as

$$\hat{\tau}_{iI_j} = \frac{\hat{\pi}_{I_j} f_{I_j}(\mathbf{x}_i; \hat{\Theta})}{\sum_{\ell=1}^k \hat{\pi}_{I_\ell} f_{I_\ell}(\mathbf{x}_i; \hat{\Theta})}$$

where  $\hat{\pi}_{I_j} = \sum_{\ell \in I_j} \hat{\pi}_\ell$ . Then, the posterior probability vector associated to observation  $\mathbf{x}_i$  is

$$\hat{\boldsymbol{\tau}}_{i\mathcal{P}_k} = (\hat{\tau}_{iI_1}, \dots, \hat{\tau}_{iI_k}). \quad (2.2)$$

The posterior probability vector  $\hat{\boldsymbol{\tau}}_{i\mathcal{P}_k}$  denotes the conditional probability that  $\mathbf{x}_i$  arises from mixture components  $f_{I_1}, \dots, f_{I_k}$ . Since  $\mathcal{P}_k$  is a partition,  $\sum_{j=1}^k \hat{\tau}_{iI_j} = 1$  holds for  $1 \leq i \leq n$ . Similarly to Equation (1.2), the posterior probability vectors  $\hat{\boldsymbol{\tau}}_{i\mathcal{P}_k}$  can be used to classify  $\mathbf{x}_i \in \mathbf{X}$  to the cluster  $c$ ,  $1 \leq c \leq k$ , if

$$c = \arg \max_{j=1}^k \{\hat{\tau}_{iI_j}\}. \quad (2.3)$$

Let  $\hat{\mathbf{T}}_{\mathcal{P}_k}$  be the matrix with  $n$  rows and  $k$  columns formed by the  $n$  vectors of posterior probabilities  $\hat{\boldsymbol{\tau}}_{i\mathcal{P}_k}$  associated to partition  $\mathcal{P}_k$ . Accordingly,  $\hat{\mathbf{T}}_{\mathcal{P}_K}$  is the initial matrix with the posterior probabilities when one component is modelling one cluster; and  $\hat{\mathbf{T}}_{\mathcal{P}_1}$ , the final matrix, is formed only by the column  $\mathbf{1} = (1, \dots, 1)$ . Importantly, with Equation (2.2), any matrix  $\hat{\mathbf{T}}_{\mathcal{P}_k}$  can be obtained from matrix  $\hat{\mathbf{T}}_{\mathcal{P}_K}$ , respectively, aggregating the corresponding columns of each of the parts  $I_1, \dots, I_k$ .

Hereinafter to simplify notation, we denote the estimation  $\hat{\tau}_{iI_j}$  as  $\tau_{iI_j}$ . Similarly, we write  $\boldsymbol{\tau}_{i\mathcal{P}_k}$  and  $\mathbf{T}_{\mathcal{P}_k}$ , respectively, instead of  $\hat{\boldsymbol{\tau}}_{i\mathcal{P}_k}$  and  $\hat{\mathbf{T}}_{\mathcal{P}_k}$ .

<sup>1</sup> $B_k$  is the  $k$ th Bell number defined recursively as  $B_0 = 1$  and  $B_{k+1} = \sum_{i=0}^k \binom{k}{i} B_i$ .

### 3 Generic approach

#### 3.1 General merging criteria

Let  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  be a sample defined in a domain  $\mathbb{X}$ , and let  $f$  be an FMM with  $k$  components defined on  $\mathbb{X}$  (Equation (1.1)). Given a partition  $\mathcal{P}_k = \{I_1, \dots, I_k\}$ , let  $\boldsymbol{\tau}_{i\mathcal{P}_k} = (\tau_{iI_1}, \dots, \tau_{iI_k})$  be the posterior probability vector associated to observation  $\mathbf{x}_i$ ,  $1 \leq i \leq n$ .

For a partition  $\mathcal{P}_k$  and matrix of posterior probabilities  $\mathbf{T}_{\mathcal{P}_k}$ , we propose merging the parts  $I_a, I_b \in \mathcal{P}_k$ , maximizing the weighted mean

$$S_{\omega, \lambda}(\mathbf{T}_{\mathcal{P}_k}, I_a, I_b) = \frac{\sum_{i=1}^n \omega(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) \lambda(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)}{\sum_{i=1}^n \omega(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)}, \quad (3.1)$$

where  $\lambda(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)$  is a real valued function and  $\omega(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)$  is a non-negative function. We refer to this maximum as the  $S$ -value. Function  $\lambda(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)$  has the role of a utility function. It measures our preferences for considering components  $f_{I_a}$  and  $f_{I_b}$  as a single component, and therefore, to model two clusters by means of the corresponding merged component as one single cluster. Function  $\omega(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b)$  is a weight function. It permits the influence that each posterior probability vector  $\boldsymbol{\tau}_{i\mathcal{P}_k}$  has in Equation (3.1) for parts  $I_a$  and  $I_b$  to be modified. The behaviour of function  $S_{\omega, \lambda}$  is wholly determined by the choice of functions  $\omega$  and  $\lambda$ . Importantly, function  $S_{\omega, \lambda}$  has the same codomain as function  $\lambda$  has.

Starting from partition  $\mathcal{P}_K = \{\{1\}, \dots, \{K\}\}$ , where the number of clusters is equal to the number of components, two of these parts are merged according to the  $S$ -value (Equation (3.1)). Iteratively repeating this process, the algorithm builds an agglomerative hierarchical sequence of partitions until partition  $\mathcal{P}_1$  is obtained. Remarkably, by the definition of functions  $\omega$  and  $\lambda$ , the process only depends on the posterior probability vectors  $\mathbf{T}_{\mathcal{P}_k}$ .

For the initial partition  $\mathcal{P}_K$ , we need to evaluate  $K^2 - K$  times function  $S_{\omega, \lambda}$  to obtain the corresponding  $S$ -value. The process is repeated from partition  $\mathcal{P}_{K-1}$  to  $\mathcal{P}_1$ , evaluating the function  $S_{\omega, \lambda}$  a maximum of  $\frac{K^3 - K}{3}$  times. These quantities are reduced by half when the function  $S_{\omega, \lambda}$  is symmetric with regards to  $I_a$  and  $I_b$ .

#### 3.2 Minimizing the final entropy

Baudry et al. (2010) propose an algorithm to build a hierarchical sequence of partitions based on the concept of entropy. The Shannon entropy of a posterior probability vector  $\boldsymbol{\tau}_{i\mathcal{P}_k} = (\tau_{iI_1}, \dots, \tau_{iI_k})$  is

$$\text{Ent}(\boldsymbol{\tau}_{i\mathcal{P}_k}) = - \sum_{j=1}^k \tau_{iI_j} \log(\tau_{iI_j}).$$

## 6 Marc Comas-Cufí et al.

The entropy can be interpreted as a measure of similarity between a probability vector  $\boldsymbol{\tau}_{i\mathcal{P}_k}$  and the probability vector  $\boldsymbol{\tau}_k^0 = \left(\frac{1}{k}, \dots, \frac{1}{k}\right)$ , taking the maximum value  $-\log\left(\frac{1}{k}\right)$ , when  $\boldsymbol{\tau}_{i\mathcal{P}_k} = \boldsymbol{\tau}_k^0$ . Given a partition  $\mathcal{P}_k = \{I_1, \dots, I_k\}$ , the algorithm iteratively merges the two mixture components, optimizing the overall entropy. Let  $\mathcal{P}_{k-1}^{I_a \cup I_b}$  be the partition obtained after merging components  $I_a$  and  $I_b$  from  $\mathcal{P}_k$ . The parts  $I_a$  and  $I_b$  merged minimize expression

$$\sum_{i=1}^n \text{Ent}(\boldsymbol{\tau}_{i\mathcal{P}_{k-1}^{I_a \cup I_b}}).$$

According to Baudry et al. (2010), minimizing the previous expression is equivalent to maximizing the loss of entropy, that is, maximizing the sum

$$\sum_{i=1}^n \left\{ \text{Ent}(\boldsymbol{\tau}_{i\mathcal{P}_k}) - \text{Ent}(\boldsymbol{\tau}_{i\mathcal{P}_{k-1}^{I_a \cup I_b}}) \right\},$$

which can be written only in terms of  $\tau_{iI_a}$  and  $\tau_{iI_b}$  as

$$\sum_{i=1}^n \left\{ (\tau_{iI_a} + \tau_{iI_b}) \log(\tau_{iI_a} + \tau_{iI_b}) - [\tau_{iI_a} \log(\tau_{iI_a}) + \tau_{iI_b} \log(\tau_{iI_b})] \right\}. \quad (3.2)$$

Note that using our general merging criteria, we can define function  $\lambda_{\Delta\text{Ent}}$  as

$$\lambda_{\Delta\text{Ent}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = (\tau_{iI_a} + \tau_{iI_b}) \log(\tau_{iI_a} + \tau_{iI_b}) - [\tau_{iI_a} \log(\tau_{iI_a}) + \tau_{iI_b} \log(\tau_{iI_b})],$$

and function  $\omega_{\text{cnst}}$  as a constant, for example,

$$\omega_{\text{cnst}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = 1.$$

When assuming that  $\omega_{\text{cnst}}$  is constant, we are considering each observation as being equally important to compute the  $S$ -value. That is, the weighting is the same regardless of the parts  $I_a$  and  $I_b$  we are willing to merge. In this case, the  $S$ -values (Equation (3.1)) take the form that is, the average of loss of entropy.

$$S_{\omega_{\text{cnst}}, \lambda_{\Delta\text{Ent}}}(\mathbf{T}_{\mathcal{P}_k}, I_a, I_b) = \frac{1}{n} \sum_{i=1}^n \left\{ (\tau_{iI_a} + \tau_{iI_b}) \log(\tau_{iI_a} + \tau_{iI_b}) - [\tau_{iI_a} \log(\tau_{iI_a}) + \tau_{iI_b} \log(\tau_{iI_b})] \right\},$$

Note that in this approach function,  $S_{\omega_{\text{cnst}}, \lambda_{\Delta\text{Ent}}}$  is symmetric with respect to  $I_a$  and  $I_b$ . Therefore, for partition  $\mathcal{P}_K$ , we only need to evaluate function  $S_{\omega_{\text{cnst}}, \lambda_{\Delta\text{Ent}}}$  for  $\frac{K^2 - K}{2}$

*Merging the components of a finite mixture using posterior probabilities* 7

times. Note also that for partition  $\mathcal{P}_{K-1}$ , we only need to update the value of  $S_{\omega_{\text{cnst}}, \lambda_{\Delta \text{Ent}}}$  for  $K - 2$  different values. Finally, to calculate the  $S$ -values in the whole process, we only need to evaluate the function  $S_{\omega_{\text{cnst}}, \lambda_{\Delta \text{Ent}}}$   $(K - 1)^2$  times.

### 3.3 Maximizing the misclassification probability

Hennig (2010) proposes merging the components  $I_a$  and  $I_b$  from  $\mathcal{P}_k$  that maximize *the probability of classifying to component  $I_b$ , an observation generated from component  $f_{I_a}$* . To estimate this probability, Hennig (2010) introduces a consistent estimator, the Directly Estimated Misclassification Probabilities (DEMP), defined as

$$\frac{\frac{1}{n} \sum_{i=1}^n \tau_{iI_a} \mathbb{1}(\forall j \tau_{iI_b} \geq \tau_{iI_j})}{\hat{\pi}_{I_a}},$$

where  $\mathbb{1}(\cdot)$  is the indicator function. Because  $\hat{\pi}_{I_a} = \frac{1}{n} \sum_{i=1}^n \tau_{iI_a}$ , the estimator DEMP can be written in terms of posterior probability as

$$\frac{\sum_{i=1}^n \tau_{iI_a} \mathbb{1}(\forall j \tau_{iI_b} \geq \tau_{iI_j})}{\sum_{i=1}^n \tau_{iI_a}}. \tag{3.3}$$

Note that when parts  $I_a$  and  $I_b$  overlap, Equation (3.3) takes higher values because the posterior probability  $\tau_{iI_b}$  is higher for observations  $i$  generated by component  $f_{I_a}$ .

Using our general merging criteria, the estimator DEMP (Equation (3.3)) is equivalent to set functions  $\omega$  and  $\lambda$  as

$$\lambda_{\text{DEMP}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \mathbb{1}(\forall j \tau_{iI_b} \geq \tau_{iI_j}) \tag{3.4}$$

and

$$\omega_{\text{prop}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \tau_{iI_a}.$$

In this approach, function  $\lambda_{\text{DEMP}}$  is giving preference to observations  $\mathbf{x}_i$  classified to part  $I_b$ . Moreover, the  $\omega_{\text{prop}}$  function is weighing higher than those observations with high posterior probability  $\tau_{iI_a}$ . This approach permits us to calculate the  $S$ -value from  $I_a$  to  $I_b$ , measuring our preference to merge  $I_a$  into  $I_b$ . Note that in this case, function  $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}$  is not symmetric, that is,  $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}(\mathbf{T}_{\mathcal{P}_k}, I_a, I_b) \neq S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}(\mathbf{T}_{\mathcal{P}_k}, I_b, I_a)$ . Therefore, to calculate the  $S$ -values, we need to evaluate the function  $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}$   $2(K - 1)^2$  times. In addition, this lack of symmetry indicates that the merging process is extended to a type of absorption process of one component by the other.

Longford and Bartošová (2014) propose a variation of the approach introduced in Hennig (2010). Rather than considering function  $\lambda_{\text{DEMP}}$  given by Equation (3.4),

8 *Marc Comas-Cufí et al.*

in this case, the preference for merging  $I_a$  into  $I_b$  is given by

$$\lambda_{\text{DEMP}_m}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \frac{\tau_{iI_b}}{\tau_{iI_a} + \tau_{iI_b}}. \quad (3.5)$$

When a method uses function  $\lambda_{\text{DEMP}_m}$ , we call it a DEMP-modified approach. Similarly, to function  $\lambda_{\text{DEMP}}$ , function  $\lambda_{\text{DEMP}_m}$  gives high scores to observations with high posterior probability  $\tau_{iI_b}$ . In this case,  $\lambda_{\text{DEMP}_m}$  is the probability of being generated by component  $f_{I_b}$  conditioned being generated by either  $f_{I_a}$  or  $f_{I_b}$ . Again, function  $\lambda_{\text{DEMP}_m}$  causes function  $S_{\omega, \lambda_{\text{DEMP}_m}}$  not to be symmetric. The function  $\lambda_{\text{DEMP}_m}$ , having the codomain  $[0, 1]$ , increases when the probability  $\tau_{iI_b}$  increases, and decreases when the probability  $\tau_{iI_a}$  is increased.

#### 4 Other criteria: The Log-ratio approach

The generic expression of  $S_{\omega, \lambda}(\mathbf{T}_{\mathcal{P}_k}, I_a, I_b)$  (Equation (3.1)) permits to extend the criteria for merging components to expressions defined by the analyst. For example, in Hennig (2010) and Longford and Bartošová (2014), observations with high  $\tau_{I_b}$  are preferred. In this context, it might be reasonable to consider

$$\lambda_{\text{prop}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \tau_{iI_b}.$$

In this case, function  $\lambda_{\text{prop}}$  combined with  $\omega_{\text{prop}}$  results in an algorithm with an easily computable function  $S_{\omega_{\text{prop}}, \lambda_{\text{prop}}}$ . As an alternative to function  $\omega_{\text{prop}}$ , we can define

$$\omega_{\text{dich}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \mathbb{1}(\forall j \tau_{iI_a} \geq \tau_{iI_j}).$$

This function gives weight one to observations classified to part  $I_a$  and zero to the others, which, in turn, results in an average of  $\lambda$  values for observations classified to part  $I_a$ .

Although there are many possible functions  $\omega$  and  $\lambda$ , we present two alternatives to function  $\lambda$  from a well-founded background. Aitchison (1986) introduces the main elements for the statistical analysis of samples defined in the simplex space  $\mathcal{S}^K$ , that is,  $\mathcal{S}^K = \{(x_1, \dots, x_K) \mid x_i > 0 \text{ and } \sum_{i=1}^K x_i = 1\}$ . Martín-Fernández and Thió-Henestrosa (2016) present a summary of last advances on this topic. The central idea of this methodology is that only the ratios between variables are of interest. Here, we take advantage that posterior probability vector- $\boldsymbol{\tau}_{i\mathcal{P}_k}$  is defined in  $\mathcal{S}^K$  to define two new functions  $\lambda$ . The first function is motivated by the entropy concept introduced by Baudry et al. (2010), while the second one is based on the definition given by Longford and Bartošová (2014).

Following Baudry et al. (2010), the closer the posterior probability vector  $\boldsymbol{\tau}_{i\mathcal{P}_{k-1}^{I_a \cup I_b}}$  is to the vector of the uniform distribution  $\boldsymbol{\tau}_k^0$ , the higher is our preference to merge components  $I_a$  and  $I_b$  (the lower is the entropy after merging  $I_a$  and  $I_b$ ). To

*Merging the components of a finite mixture using posterior probabilities* 9

measure our preference for merging  $I_a$  and  $I_b$ , we propose calculating the similarity between the vector  $\boldsymbol{\tau}_{(I_a, I_b)} = \left( \frac{\tau_{iI_a}}{\tau_{iI_a} + \tau_{iI_b}}, \frac{\tau_{iI_b}}{\tau_{iI_a} + \tau_{iI_b}} \right)$  and the vector  $\boldsymbol{\tau}_2^0 = \left( \frac{1}{2}, \frac{1}{2} \right)$ . That is, we restrict only on the probabilities associated to the components to be merged. According to Frey and Dueck (2007), we can use an appropriate distance to define the similarity measure. Indeed, using the squared Aitchison distance (Palarea-Albaladejo et al., 2012) given by

$$d_{\mathcal{A}}^2(\boldsymbol{\tau}_{(I_a, I_b)}, \boldsymbol{\tau}_2^0) = \log^2 \left( \frac{\tau_{iI_b}}{\tau_{iI_a}} \right),$$

we can define the similarity function

$$\lambda_{\text{dist}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = -d_{\mathcal{A}}^2(\boldsymbol{\tau}_{(I_a, I_b)}, \boldsymbol{\tau}_2^0) = -\log^2 \left( \frac{\tau_{iI_b}}{\tau_{iI_a}} \right).$$

Similarly, to function  $\lambda_{\Delta\text{Ent}}$ , the function  $\lambda_{\text{dist}}$  measures how close  $\boldsymbol{\tau}_{(I_a, I_b)}$  is to  $\boldsymbol{\tau}_2^0$ . However, the codomain of function  $\lambda_{\Delta\text{Ent}}$  also implicitly depends on the posterior probabilities for the other parts different of those to be merged. Figure 1 shows these functions for a partition with three components  $I_a$ ,  $I_b$  and  $I_c$ . For different values of  $\tau_{iI_c}$ , the curves represent the effect of  $\tau_{iI_a}$  and  $\tau_{iI_b}$  into functions  $\lambda$ . We see that both functions  $\lambda$  take their maximum for  $\tau_{iI_a} = \tau_{iI_b} = \frac{1 - \tau_{iI_c}}{2}$ . However, whereas the maximum value of function  $\lambda_{\text{dist}}$  is zero regardless the value of  $\tau_{iI_c}$  (Figure 1 (right)), the codomain of function  $\lambda_{\Delta\text{Ent}}$  depends on  $\tau_{iI_c}$  (Figure 1 (left)). This fact suggests that the selection of the best components to be merged using  $\lambda_{\Delta\text{Ent}}$ , based on the corresponding  $S$ -value, can be affected by the size of the clusters. This effect agrees with the performance described by Baudry et al. (2010), when the authors apply their own criteria.

Following the function  $\lambda_{\text{DEMP}_m}$ , introduced by Longford and Bartošová (2014) (Equation (3.5)), we can define another function  $\lambda$  using log-ratios. We propose measuring the relative difference between  $\tau_{iI_b}$  and  $\tau_{iI_a}$  with the log-ratio

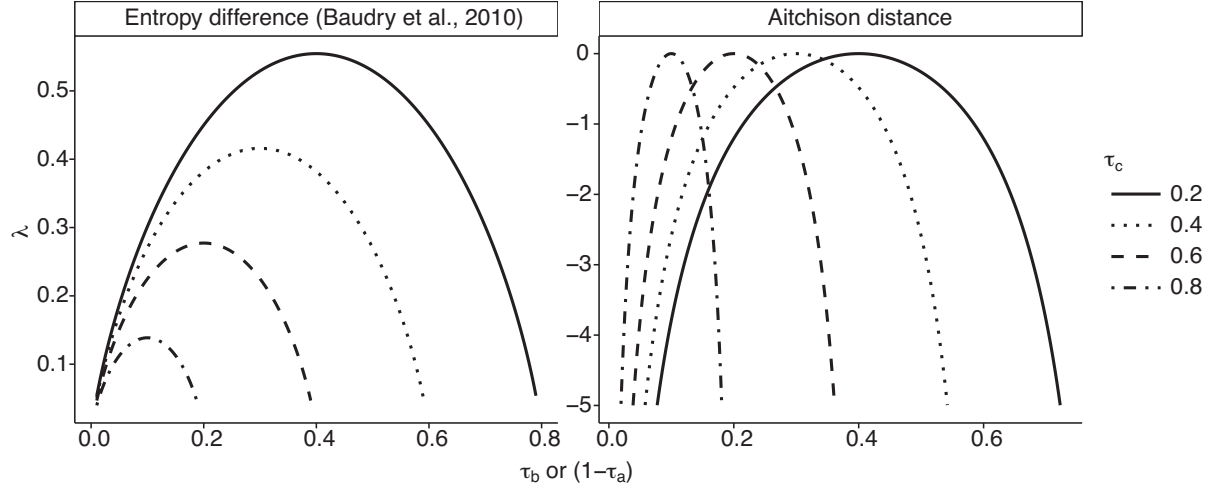
$$\lambda_{\text{log}}(\boldsymbol{\tau}_{i\mathcal{P}_k}, I_a, I_b) = \log \left( \frac{\tau_{iI_b}}{\tau_{iI_a}} \right),$$

which increases when the probability  $\tau_{iI_b}$  increases, and decreases when the probability  $\tau_{iI_a}$  increases.

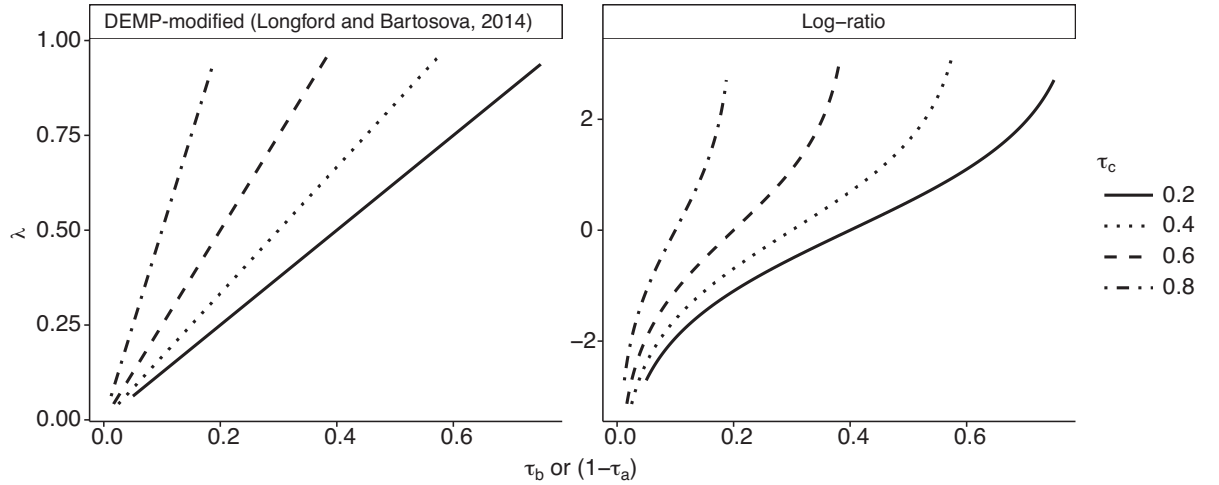
Figure 2 shows the behaviour of functions  $\lambda_{\text{DEMP}_m}$  and  $\lambda_{\text{log}}$  for posterior probability vectors  $(\tau_{iI_a}, \tau_{iI_b}, \tau_{iI_c})$ , when  $\tau_{iI_c} \in \{0.2, 0.4, 0.6, 0.8\}$ . Note that the maximum value of both functions is not affected by the value  $\tau_{iI_c}$ . However, whereas the codomain of function  $\lambda_{\text{DEMP}_m}$  is the interval  $[0, 1]$ , the codomain of function  $\lambda_{\text{log}}$  is the real space. We can see that the highest values of function  $\lambda_{\text{log}}$  occur for observations with high  $\tau_{iI_b}$  relative to  $\tau_{iI_a}$ . Because the comparison is relative, observations not related



10 *Marc Comas-Cufí et al.*



**Figure 1** Function  $\lambda$  for posterior probability vectors  $(\tau_{il_a}, \tau_{il_b}, \tau_{il_c})$  with  $\tau_{il_c} \in \{0.2, 0.4, 0.6, 0.8\}$ : (left)  $\lambda_{\text{Ent}}$ ; (right)  $\lambda_{\text{dist}}$



**Figure 2** Function  $\lambda$  for posterior probability vectors  $(\tau_{il_a}, \tau_{il_b}, \tau_{il_c})$  with  $\tau_{il_c} \in \{0.2, 0.4, 0.6, 0.8\}$ : (left)  $\lambda_{\text{DEMP}_m}$ ; (right)  $\lambda_{\text{log}}$

to parts  $I_a$  or  $I_b$  are able to play an important role in the final  $S$ -value. Therefore, the selection of function  $\omega(\tau_{i\mathcal{P}_k}, I_a, I_b)$  is especially important. To weight higher those observations related to part  $I_a$ , a reasonable selection for function  $\omega$  would be  $\omega_{\text{prop}}$  or  $\omega_{\text{dich}}$ . In this context, because we need to weight higher those observations related with part  $I_a$ , function  $\omega_{\text{cnst}}$  makes no sense.

As  $\lambda_{\text{dist}}$  and  $\lambda_{\text{log}}$  use log-ratios, they take into account the geometric properties of the simplex space (Aitchison, 2002). When working with log-ratios between the components of a posterior probability vector, ‘subcompositional coherence’ holds (Aitchison, 1986). This property guarantees that any statistical inference obtained

*Merging the components of a finite mixture using posterior probabilities* 11

using only partial information is coherent with results obtained using complete information. Formally, in our context subcompositional coherence can be defined as:

**Definition 4.1.** Let  $\mathcal{P}_1, \dots, \mathcal{P}_K$  be a hierarchical sequence of partitions obtained from posterior probability matrix  $\mathbf{T}_{\mathcal{P}_K}$ , using a merging approach  $M$ . Let  $I = \{j_1, \dots, j_k\}$  be an element (a part) of  $\mathcal{P}_k$ , for some  $k, 1 \leq k \leq K$ . A method  $M$  is ‘subcompositional coherent’, if the hierarchy subsequence of partitions obtained using only posterior probability vectors  $\{\boldsymbol{\tau}_{\dot{j}_1}, \dots, \boldsymbol{\tau}_{\dot{j}_k}\}_{1 \leq i \leq n}$  is contained in the original hierarchy.

Another interesting feature of the log-ratio approach is the ‘scale invariance’ property (Aitchison, 1986), formally

$$S_{\omega, \lambda}(\boldsymbol{\tau}_{\mathcal{P}_k}, I_a, I_b) = S_{\omega, \lambda}(\kappa \cdot \boldsymbol{\tau}_{\mathcal{P}_k}, I_a, I_b) \text{ for } \kappa > 0.$$

This property suggests that the log-ratio approach is not restricted only to posterior probability vectors. On the contrary, it can be applied to any other kind of vector giving relative information between mixture components. That is, the two methods considered here are suitable to be applied in more general scenarios such as using vectors of weights of parts (e.g., weights in fuzzy clustering).

Remarkably, when  $\omega_{\text{prop}}$  and  $\lambda_{\text{log}}$  are considered, the  $S$ -value (Equation (3.1)) results in

$$S_{\omega_{\text{prop}}, \lambda_{\text{log}}}(\boldsymbol{\tau}_{\mathcal{P}_k}, I_a, I_b) = \frac{\sum_{i=1}^n \tau_{iI_a} \log\left(\frac{\tau_{iI_b}}{\tau_{iI_a}}\right)}{\sum_{i=1}^n \tau_{iI_a}}.$$

For a fixed component  $I_a$ , the denominator is constant, and we only need to maximize the numerator. The expression in the numerator has the essence of the Kullback–Leibler divergence (in negative sign), comparing the distribution of classifying observations to  $I_a$  against the distributions of classifying the same observations to  $I_b$ .

## 5 Deciding the number of clusters

Given a finite mixture adjusted with BIC criteria, we have presented a generic approach to build a hierarchical sequence of partitions. One of the main difficulties is deciding the final number of clusters. For any partition  $\{I_1, \dots, I_k\}$ , the likelihood function of mixture  $f = \pi_{I_1}f_{I_1} + \dots + \pi_{I_k}f_{I_k}$  is always the same. In other words, from a frequentist perspective, it is not possible to decide which of the different ways of modelling a cluster with different components is the best (Hennig, 2010). Therefore, we need to use heuristic methods to decide the final number of clusters.

12 *Marc Comas-Cufí et al.*

## 5.1 Using $S$ -values

A first option when deciding the number of clusters is the  $S$ -values. In the case of  $\omega_{\text{cnst}}$  and  $\lambda_{\Delta\text{Entr}}$ , Baudry et al. (2010) propose visualizing the  $S$ -values and apply the elbow rule. For  $\omega_{\text{prop}}$  and  $\lambda_{\text{DEMP}}$ , Hennig (2010) propose setting an arbitrary threshold and stop merging when the  $S$ -value is lower than the fixed threshold. Although there is no rule of thumb to define a method for the general merging criteria, for the particular proposals described in Sections 3 and 4, the  $S$ -values can be a useful tool in deciding the number of clusters. Indeed, from the definition of  $\lambda$  values as a utility function, a small value of its weighted average (the  $S$ -value) might suggest stopping the merging process. Hennig (2010) introduces a criteria to decide when an  $S$ -value is small, setting a threshold using simulation in borderline situations.

To better interpret the  $S$ -values, we can normalize them to the interval  $[0, 1]$ , using a monotone function. Two options are feasible: scaling function  $\lambda$  or scaling function  $S_{\omega, \lambda}$ . Note that any function  $\phi$  that scales function  $\lambda$  also scales function  $S_{\omega, \lambda}$ . We denote these two scaling options by  $S_{\omega, \phi \circ \lambda}$  and  $\phi \circ S_{\omega, \lambda}$ . Importantly, the first approach  $S_{\omega, \phi \circ \lambda}$  is modifying function  $\lambda$ , and therefore, it can be considered as a new method in itself. It is worth mentioning that for  $\phi_{\log}(x) = \frac{e^x}{1 + e^x}$ , the scaling  $S_{\omega, \phi_{\log} \circ \lambda_{\log}}$  reduces to  $S_{\omega, \lambda_{\text{DEMP}_m}}$ . That is, the function  $\lambda_{\text{DEMP}_m}$  (Longford and Bartošová, 2014) is a normalized version of the function  $\lambda_{\log}$ .

From their definition, the functions  $\lambda_{\Delta\text{Entr}}$ ,  $\lambda_{\text{dist}}$  and  $\lambda_{\log}$  are not scaled into the interval  $[0, 1]$ . For scaling these functions, we propose  $\phi_{\Delta\text{Entr}}(x) = -x/\log\left(\frac{1}{k}\right)$ ,  $\phi_{\text{dist}}(x) = 1 - e^{-x}$  and  $\phi_{\log}(x) = \frac{e^x}{1 + e^x}$ , respectively. Despite these functions not being the only possibilities for the scaling, we selected them because of their reasonable performance in our experiments. The corresponding scaled  $S$ -value functions are:  $\phi_{\Delta\text{Entr}} \circ S_{\omega, \lambda_{\Delta\text{Entr}}}$  or  $S_{\omega, \phi_{\Delta\text{Entr}} \circ \lambda_{\Delta\text{Entr}}}$ ,  $\phi_{\text{dist}} \circ S_{\omega, \lambda_{\text{dist}}}$  or  $S_{\omega, \phi_{\text{dist}} \circ \lambda_{\text{dist}}}$  and  $\phi_{\log} \circ S_{\omega, \lambda_{\log}}$  or  $S_{\omega, \phi_{\log} \circ \lambda_{\log}}$ . Once the  $S$ -values are normalized, one reasonable rule of thumb is to stop the merging process when these values are close to zero.

## 5.2 Using the location of the posterior probabilities

For a sample with  $k$  well-separated clusters of observations, when each cluster is modelled by a different probability distribution, the posterior probability vectors of the elements of each cluster are located close to a vertex of the simplex  $\mathcal{S}^k$ . For example, in Figure 3, we have a sample generated following a mixture of four Gaussian distributions. In the first scenario (top left), each component models one different cluster, while in the second scenario (bottom left), we are modelling three clusters where the two components centred at  $(10, 10\sqrt{3})$  are modelling one cluster together. When we consider the sample with four clusters, the posterior probability vectors can be represented in a quaternary diagram (top right). In the quaternary

*Merging the components of a finite mixture using posterior probabilities* 13

diagram, we see the posterior probability vectors of parts  $I_1$  and  $I_3$  are respectively located close to vertices  $\tau_1$  and  $\tau_3$  of  $\mathcal{S}^4$ . In contrast, the posterior probability vectors of individuals  $I_2$  and  $I_4$  are not well separated, and they expand through the edge running from  $\tau_2$  to  $\tau_4$ . The situation changes if we consider the three cluster sample (bottom left). In this case, we can represent the posterior probability vectors in a ternary diagram (bottom right) to discern that the posterior probability vectors of all three parts are located in three different vertices.

Therefore, we can use the geometric structure of  $\mathcal{S}^k$  (Pawlowsky-Glahn and Egozcue, 2001) to define the number of clusters based on how close the posterior probability vectors are to the vertices and how *well* separated the posterior probability vectors associated to each cluster are.

Let  $I_1, \dots, I_k$  be the parts defining the  $k$  clusters. The posterior probability vectors,  $\tau_i$ , of observations  $x_i$  assigned to  $I_a$  are close to the corresponding vertex, if  $\tau_{iI_a}$  is close to one. In addition,  $\tau_{iI_b}$ , for  $b \neq a$ , is close to zero. Consequently, the log-ratio  $\log(\tau_{iI_a}/\tau_{iI_b})$  should take larger values for all observations  $x_i$  assigned to cluster  $I_a$ . To decide how close  $\tau_i$  is to the vertex of the simplex, we calculate

$$v_{i,a} = \min_{b, b \neq a} \log \left( \frac{\tau_{iI_a}}{\tau_{iI_b}} \right).$$

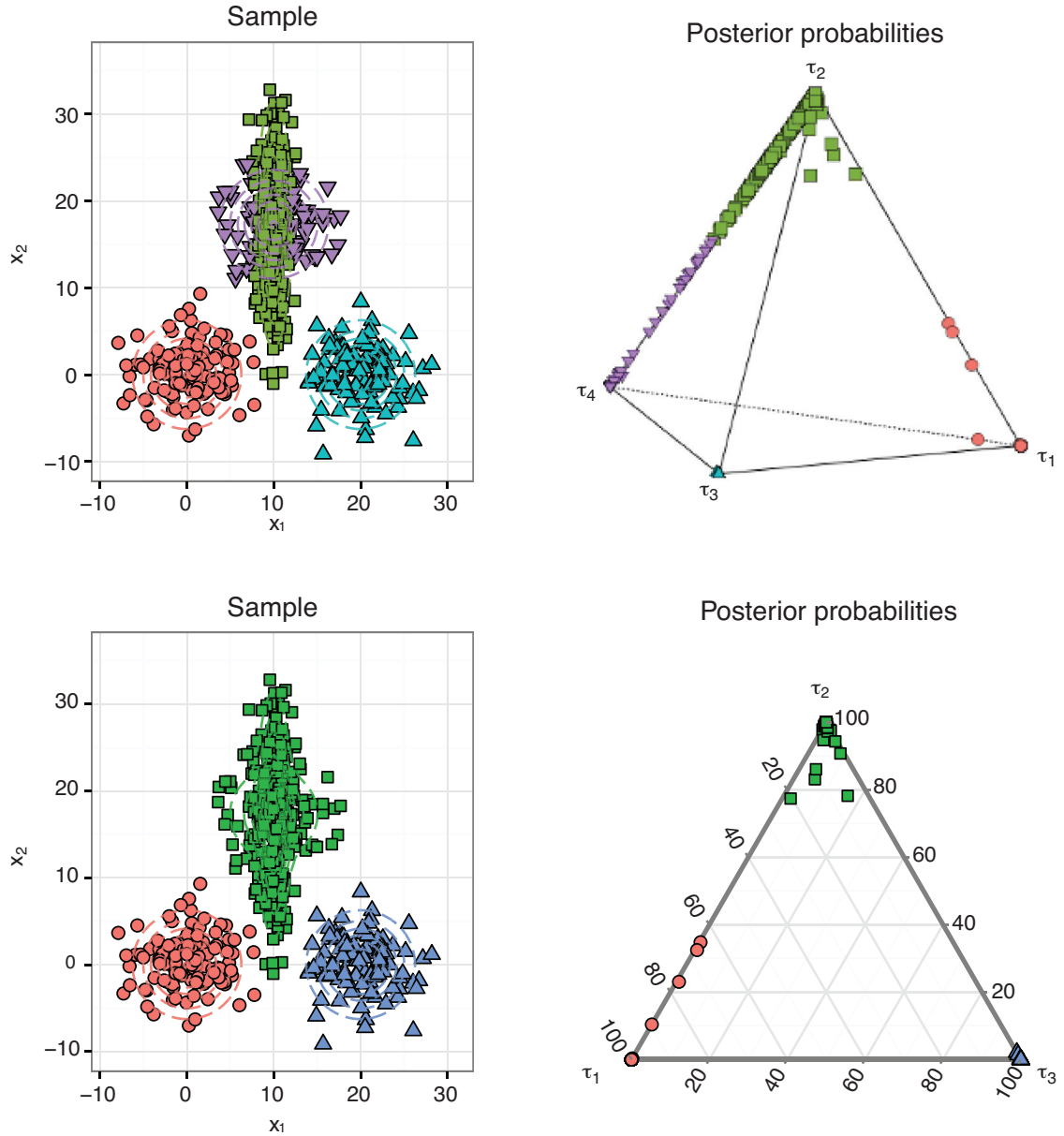
Because  $\tau_{iI_a}$  is the highest posterior probability (observation  $x_i$  is assigned to  $I_a$ ),  $v_{i,a}$  is positive. Importantly, the lower the value  $v_{i,a}$ , the farther is  $\tau_i$  from vertex  $I_a$ . For example, when considering four clusters (Figure 3 (top left)), the averages of  $v_{i,j}$  for observations assigned to clusters centred at  $(0, 0)$  and  $(20, 0)$  are 15.83 and 14.46, respectively. In contrast, the average of  $v_{i,j}$  for observations assigned to clusters centred at  $(10, 10\sqrt{3})$  is significantly lower (2.15 and 4.00). When three clusters are considered (Figure 3 (bottom left)), the averages of  $v_{i,j}$  for observations assigned to clusters centred at  $(0, 0)$ ,  $(20, 0)$  and  $(10, 10\sqrt{3})$  are 15.82, 14.46 and 19.48, respectively.

Therefore, using this measure to identify if there are two clusters not well separated, we calculate the following index:

$$\mathcal{V} = \min_{j \in \{1, \dots, k\}} v_{i,j}.$$

In the first scenario (Figure 3 (top right)),  $\mathcal{V} = 2.15$  and, in the second scenario, (Figure 3 (bottom right)),  $\mathcal{V} = 14.46$ .

A different approach to identify if the posterior probability  $\tau_i$  are close to the vertices is to identify if the  $\tau_i$  associated to parts are forming a cluster by themselves. For example, in Figure 3 (top right), the posterior probability vectors associated to  $\tau_2$  and  $\tau_4$  are not forming two well-separated clusters. To measure this separation, we propose using indices that can be calculated using distances between individuals. For this purpose, we used the Calinski–Harabasz (G1) and the Goodman–Kruskal (G2) indices (Milligan, 1985), using the Aitchison distance in  $\mathcal{S}^k$ . For example, the indices

14 *Marc Comas-Cufí et al.*

**Figure 3** Top left: Sample generated from a mixture of four Gaussian distributions where each component is considered as one cluster. Top right: Posterior probability vectors of the sample represented in a quaternary diagram. Bottom left: Same sample where the four components are modelling three clusters. Bottom right: Posterior probability vectors of the sample represented in a ternary diagram

G1 and G2, respectively, take the values 350.28 and 0.68 for four-clusters solution in the Figure 3 (top right), whereas for three-clusters solution, the corresponding values are 759.49 and 0.85 (bottom right). These differences indicate that the structure with three clusters seems to be more adequate for the sample.

## 6 Examples

### 6.1 Merging components in a mixture of Gaussian distributions

Consider the bivariate Gaussian mixture of six components (Baudry et al., 2010)

$$f = \sum_{j=1}^6 \pi_j \phi(\cdot; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

with the parameters shown in the Table 1.

**Table 1** Parameters defining a two dimensional Gaussian mixture with six components. The parameters  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$  are expressed in terms of the univariate means  $\mu_{jx_1}$ ,  $\mu_{jx_2}$  and the univariate variances  $\sigma_{jx_1}^2$ ,  $\sigma_{jx_2}^2$ . The correlation  $\rho_{jx_1 x_2}$  between  $x_1$  and  $x_2$  was fixed at zero

$j$	$\pi_j$	$\mu_{jx_1}$	$\mu_{jx_2}$	$\sigma_{jx_1}^2$	$\sigma_{jx_2}^2$
1	1/6	0	0	50	5
2	1/6	0	40	5	50
3	1/6	40	40	5	50
4	1/6	0	0	5	50
5	1/6	40	0	50	5
6	1/6	40	40	50	5

Let the parameters of  $f$  be known. Figure 4 shows the isodensity curves of the estimated FMM for a random sample  $\mathbf{X}$ . We want to cluster sample  $\mathbf{X}$ .

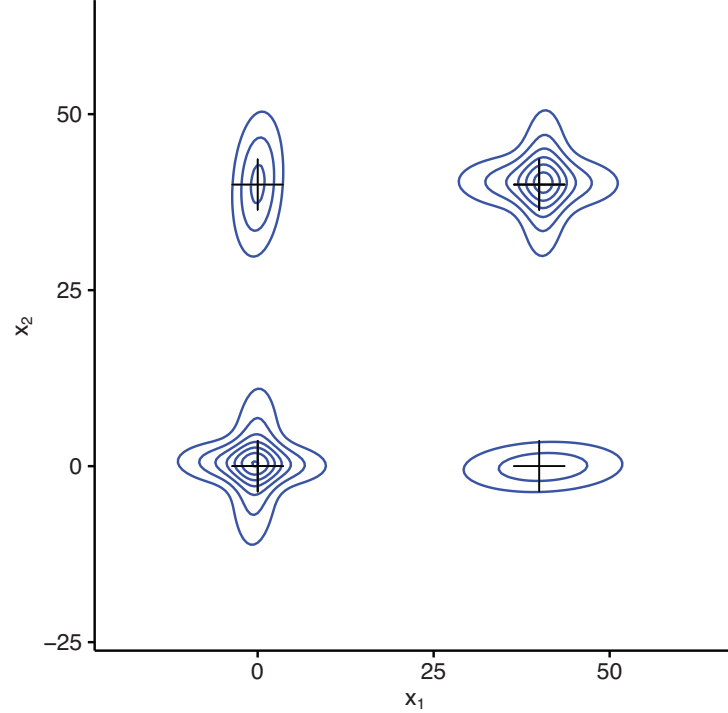
The initial partition  $\mathcal{P}_6 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$  by Equation (2.3) yields to six clusters, where each component is associated to one cluster. In Figure 5, we have separated the observations with respect to the cluster they were assigned to. The plot also includes the isodensity curves for the density modelling each cluster; in this case each cluster is modelled with a Gaussian distribution.

Using the posterior probability  $\mathbf{T}_{\mathcal{P}_k}$  and functions  $\omega_{\text{cnst}}$  and  $\lambda_{\Delta\text{Ent}}$ , we obtained the hierarchical sequence of partitions given by

$$\begin{aligned}
 \mathcal{P}_6 &= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}, \\
 \mathcal{P}_5 &= \{\{1\}, \{2\}, \{3, 6\}, \{4\}, \{5\}\}, \\
 \mathcal{P}_4 &= \{\{1, 4\}, \{2\}, \{3, 6\}, \{5\}\}, \\
 \mathcal{P}_3 &= \{\{1, 2, 4\}, \{3, 6\}, \{5\}\}, \\
 \mathcal{P}_2 &= \{\{1, 2, 4, 5\}, \{3, 6\}\}, \\
 \mathcal{P}_1 &= \{\{1, 2, 3, 4, 5, 6\}\}.
 \end{aligned} \tag{6.1}$$

In partition  $\mathcal{P}_4 = \{\{1, 4\}, \{2\}, \{3, 6\}, \{5\}\}$ , the part  $\{1, 4\}$  defines a single cluster, as does part  $\{3, 6\}$ . For this partition, using Equation (2.3) each observation  $\mathbf{x}_i$  is classified into one component. Figure 6 shows the observations separated with respect to the cluster they were classified into, together with the isodensity curves defined by

## 16 Marc Comas-Cufí et al.



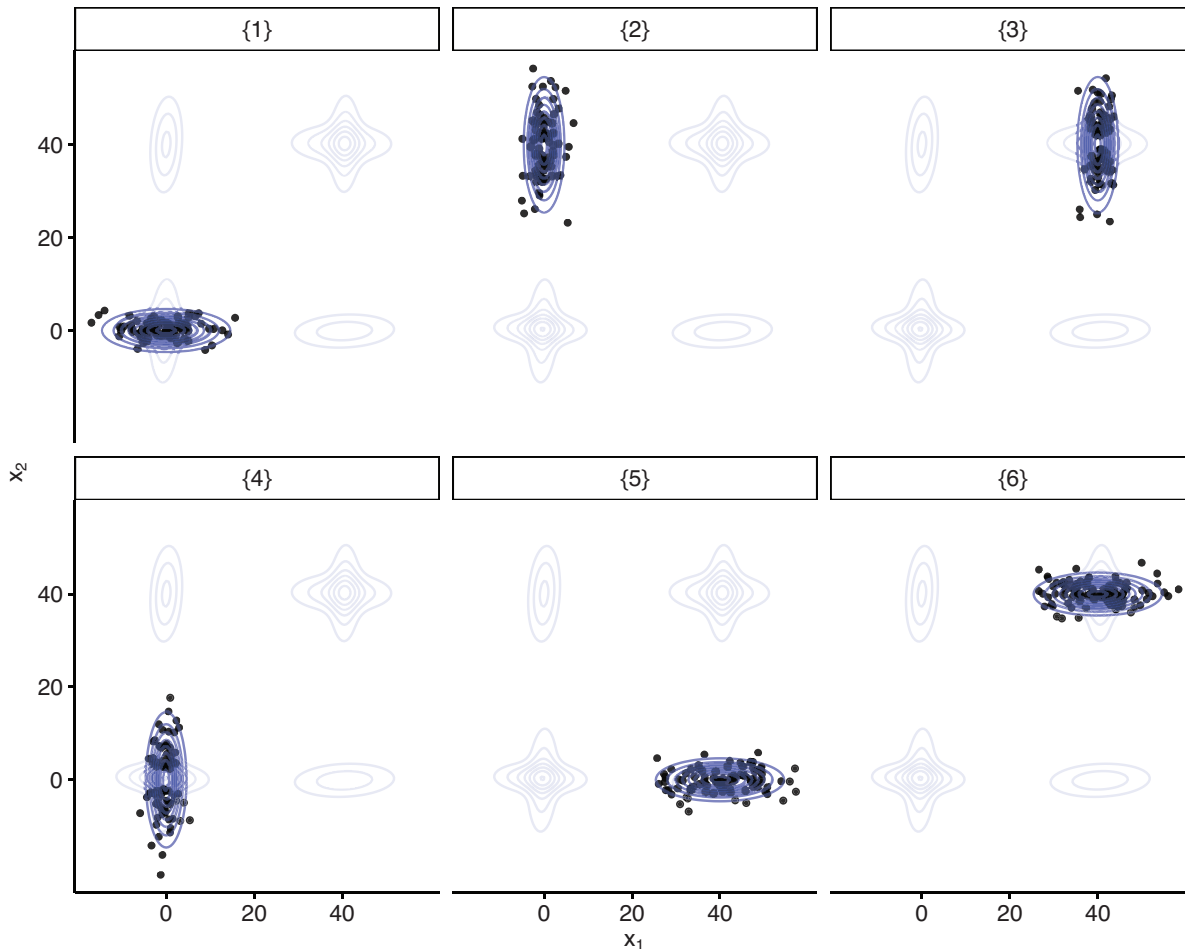
**Figure 4** Density of Gaussian mixture of six components. Each component's sample mean is represented by '+'

each component. In this case, clusters labelled  $\{1, 4\}$  and  $\{3, 6\}$  are modelled by a mixture of two components. With partition  $\mathcal{P}_4$ , the clusters are modelled by FMMs

- $f_{\{1,4\}} = \frac{1}{2}\phi(\cdot; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{2}\phi(\cdot; \boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4)$ ;
- $f_{\{2\}} = \phi(\cdot; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ ;
- $f_{\{3,6\}} = \frac{1}{2}\phi(\cdot; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) + \frac{1}{2}\phi(\cdot; \boldsymbol{\mu}_6, \boldsymbol{\Sigma}_6)$ ; and
- $f_{\{5\}} = \phi(\cdot; \boldsymbol{\mu}_5, \boldsymbol{\Sigma}_5)$ .

When  $\omega_{prop}$  is combined with  $\lambda_{DEMP_m}$ ,  $\lambda_{dist}$ ,  $\lambda_{log}$  and  $\lambda_{prop}$ , we obtained the same hierarchical partition (Equation (6.1)). Using  $\omega_{prop}$  and  $\lambda_{DEMP}$ , the hierarchical sequence of partitions obtained only differs from the previous one in partition  $\mathcal{P}_5$ , which is now  $\mathcal{P}_5 = \{\{1, 4\}, \{2\}, \{3\}, \{5\}, \{6\}\}$ .

To decide the number of clusters, we plot the  $S$ -values using different approaches (Figure 7 (top)). Using the Aitchison distance between posterior probabilities, Figure 7 (bottom) shows the Calinski–Harabasz (G1) and the Goodman–Kruskal (G2) indices (Milligan, 1985), and the closeness to the vertex index. We see that after merging components to model four clusters or less, the  $S$ -values are close to zero for all



**Figure 5** Sample separated into six clusters where each cluster is represented by a single component

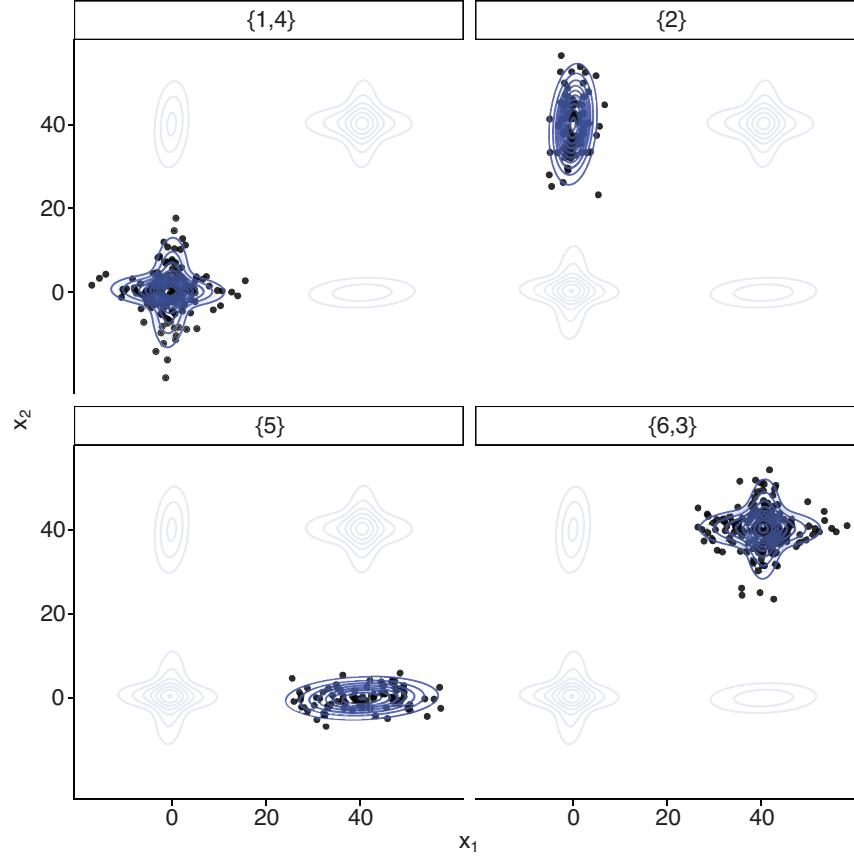
approaches. This indicates that the classifications defined with four, three or two clusters are well separated. Both  $G1$  and  $G2$  have a local maximum in four clusters, indicating that with those methods to separate the data into four clusters is preferred. The closeness to the vertex criteria increases after merging into four clusters, indicating that the clusters are close to a vertex. Combining the information obtained with the  $S$ -values, the indices  $G1$  and  $G2$ , and the closeness criteria, we then classify our data into four clusters using partition  $\mathcal{P}_4$  as shown in Figure 6.

## 6.2 Merging components in a mixture of multinomial distributions

Merging approaches presented in this article rely on the vector of posterior probabilities, which can be calculated from any FMM. Therefore, the merging generic approach introduced in Section 3 can be used for any family of FMM, for example, a finite mixture of multinomial distributions.



18 *Marc Comas-Cufí et al.*

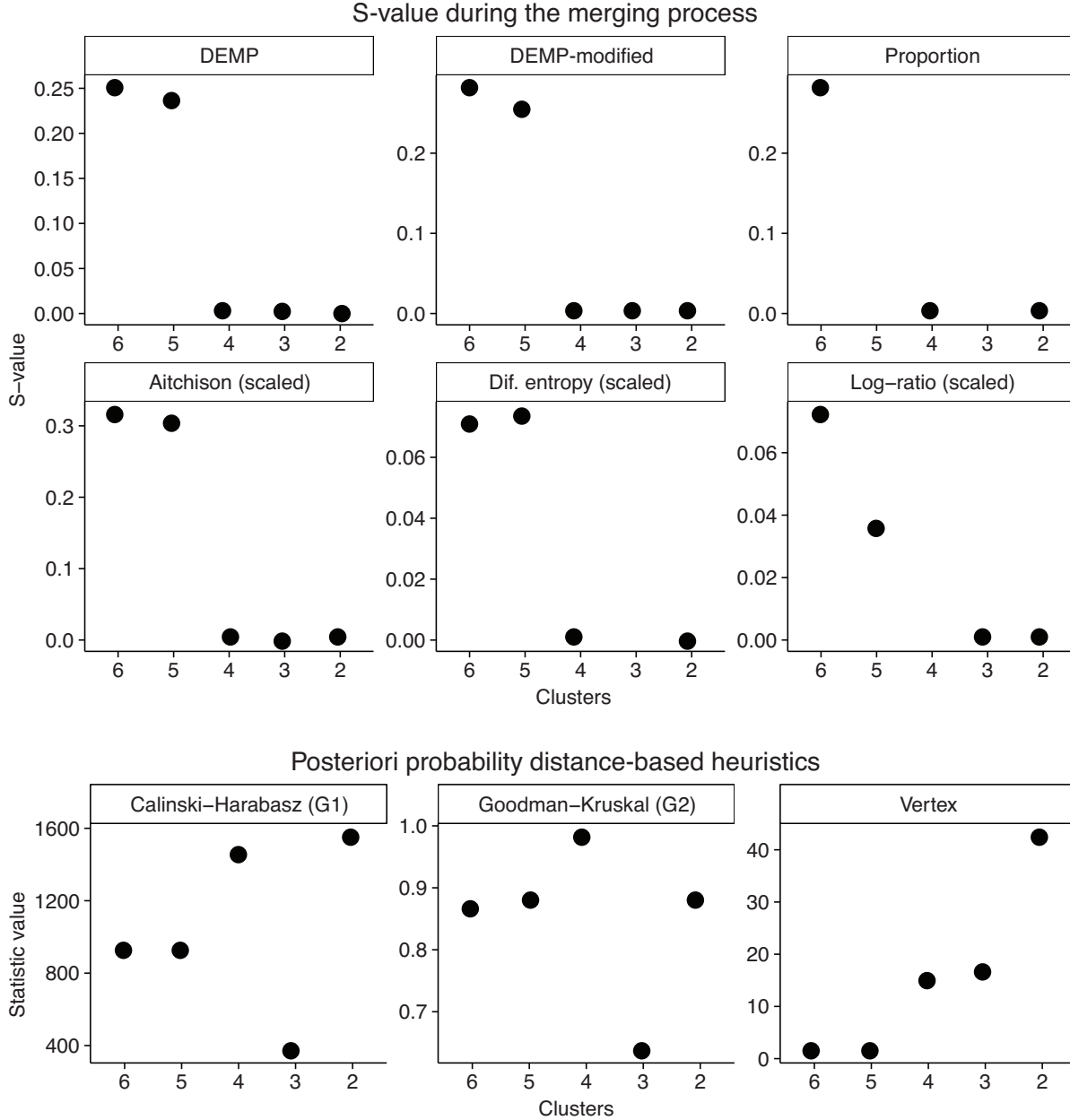


**Figure 6** Sample separated in four clusters where two clusters are represented by a mixture of two Gaussian components

Pigs dataset can be obtained from the ‘zCompositions’ (Palarea-Albaladejo and Martín-Fernández, 2015) R package. The dataset contains count data of behavioural observations of a group of 29 sows. The sows were recorded over a five-minute period in different moments, and their activity was subsequently registered. Six locations were considered for each pig: straw bed (BED), half in the straw bed (HALF.BED), dunging passage (PASSAGE), half in the dunging passage (HALF.PASS), feeder (FEEDER) and half in the feeder (HALF.FEED).

We used ‘mixtools’ (Benaglia et al., 2009) to fit a multinomial mixture. Six components were identified as optimum according to the BIC criteria. Table 2 shows the parameters of each of the components. Using these parameters, we can compute the posterior probability matrix  $\mathbf{T}_{\mathcal{P}_6}$ . Each observation is classified into one cluster following Equation (1.2) or Equation (2.3) with partition  $\mathcal{P}_6 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$ . In Figure 8, we can see the bar plot for the observations classified into the same cluster.

Using  $\omega_{\text{dich}}$  or  $\omega_{\text{prop}}$  with  $\lambda_{\text{dist}}$  or  $\lambda_{\text{log}}$ , we obtained the hierarchical structure partition given by



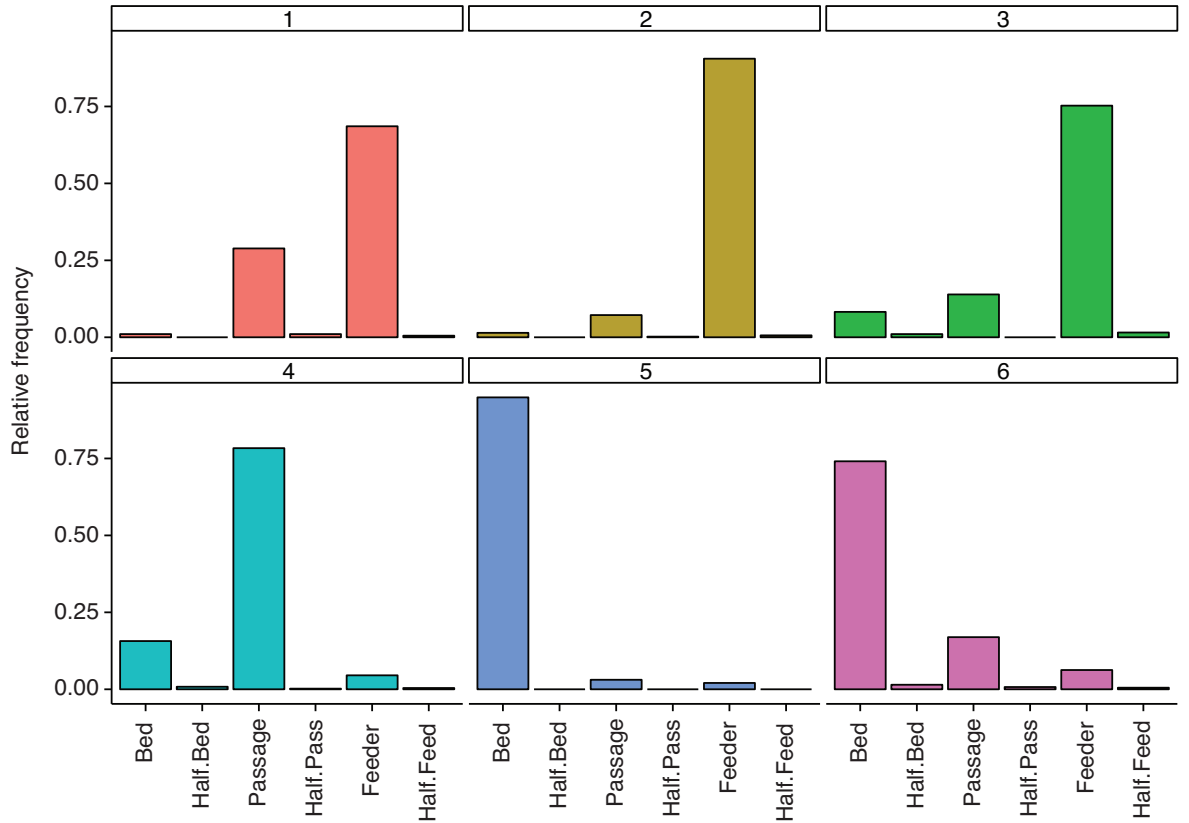
**Figure 7** Different criterias to decide the number of clusters for the sample generated with a finite mixture of six Gaussian components: (Top) S-values of the different approaches. By rows: DEMP approach (Hennig, 2010) ( $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}$ ), DEMP-modified approach (Longford and Bartošová, 2014) ( $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}_m}}$ ), posterior probability approach ( $S_{\omega_{\text{prop}}, \lambda_{\text{prop}}}$ ), Aitchison distance with proportional weighing scaled ( $\phi_{\text{dist}} \circ S_{\omega_{\text{prop}}, \lambda_{\text{dist}}}$ ), difference of entropies approach scaled ( $\phi_{\Delta \text{Ent}} \circ S_{\omega_{\text{cnst}}, \lambda_{\Delta \text{Ent}}}$ ) and log-ratio approach with proportional weighing scaled ( $\phi_{\text{log}} \circ S_{\omega_{\text{prop}}, \lambda_{\text{log}}}$ ). (Bottom) Distance-based criteria for the posterior probability vectors: Calinski-Harabasz (G1), Goodman-Kruskal (G2) and closeness to the vertex (Section 5.2)

20 *Marc Comas-Cufí et al.*

**Table 2** Parameters of a finite mixture of multinomial distributions adjusted to the Pigs dataset. For component  $j$ , the mixing proportions are denoted by  $\pi_j$  and the multinomial probabilities by  $(\theta_{j1}, \dots, \theta_{j6})$

Comp.	$\pi_j$	$\theta_{j1}$	$\theta_{j2}$	$\theta_{j3}$	$\theta_{j4}$	$\theta_{j5}$	$\theta_{j6}$
1	0.0695	0.0103	0.0000	0.2874	0.0103	0.6867	0.0052
2	0.1710	0.0144	0.0000	0.0717	0.0020	0.9057	0.0062
3	0.0699	0.0817	0.0102	0.1390	0.0000	0.7538	0.0154
4	0.1724	0.1567	0.0082	0.7835	0.0021	0.0454	0.0041
5	0.0345	0.9485	0.0000	0.0309	0.0000	0.0206	0.0000
6	0.4828	0.7408	0.0147	0.1694	0.0074	0.0626	0.0052

$$\begin{aligned}
 \mathcal{P}_6 &= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}, \\
 \mathcal{P}_5 &= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}\}, \\
 \mathcal{P}_4 &= \{\{1, 2\}, \{3\}, \{4\}, \{5, 6\}\}, \\
 \mathcal{P}_3 &= \{\{1, 2, 3\}, \{4\}, \{5, 6\}\}, \\
 \mathcal{P}_2 &= \{\{1, 2, 3\}, \{4, 5, 6\}\}, \\
 \mathcal{P}_1 &= \{\{1, 2, 3, 4, 5, 6\}\}.
 \end{aligned} \tag{6.2}$$



**Figure 8** Components after adjusting a six mixture of multinomial distributions. For each cluster, the relative amount of time seen in each location is shown

*Merging the components of a finite mixture using posterior probabilities* 21

Other criteria such as  $S_{\omega_{\text{csnt}}, \lambda_{\Delta \text{Ent}}}$ ,  $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}$  and  $S_{\omega_{\text{prop}}, \lambda_{\text{prop}}}$  differed only in partitions  $\mathcal{P}_5$  and  $\mathcal{P}_4$ . That is, these methods preferred to first merge the part  $\{1, 2, 3\}$  obtaining partitions  $\mathcal{P}_5 = \{\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6\}\}$  and  $\mathcal{P}_4 = \{\{1, 2, 3\}, \{4\}, \{5\}, \{6\}\}$ .

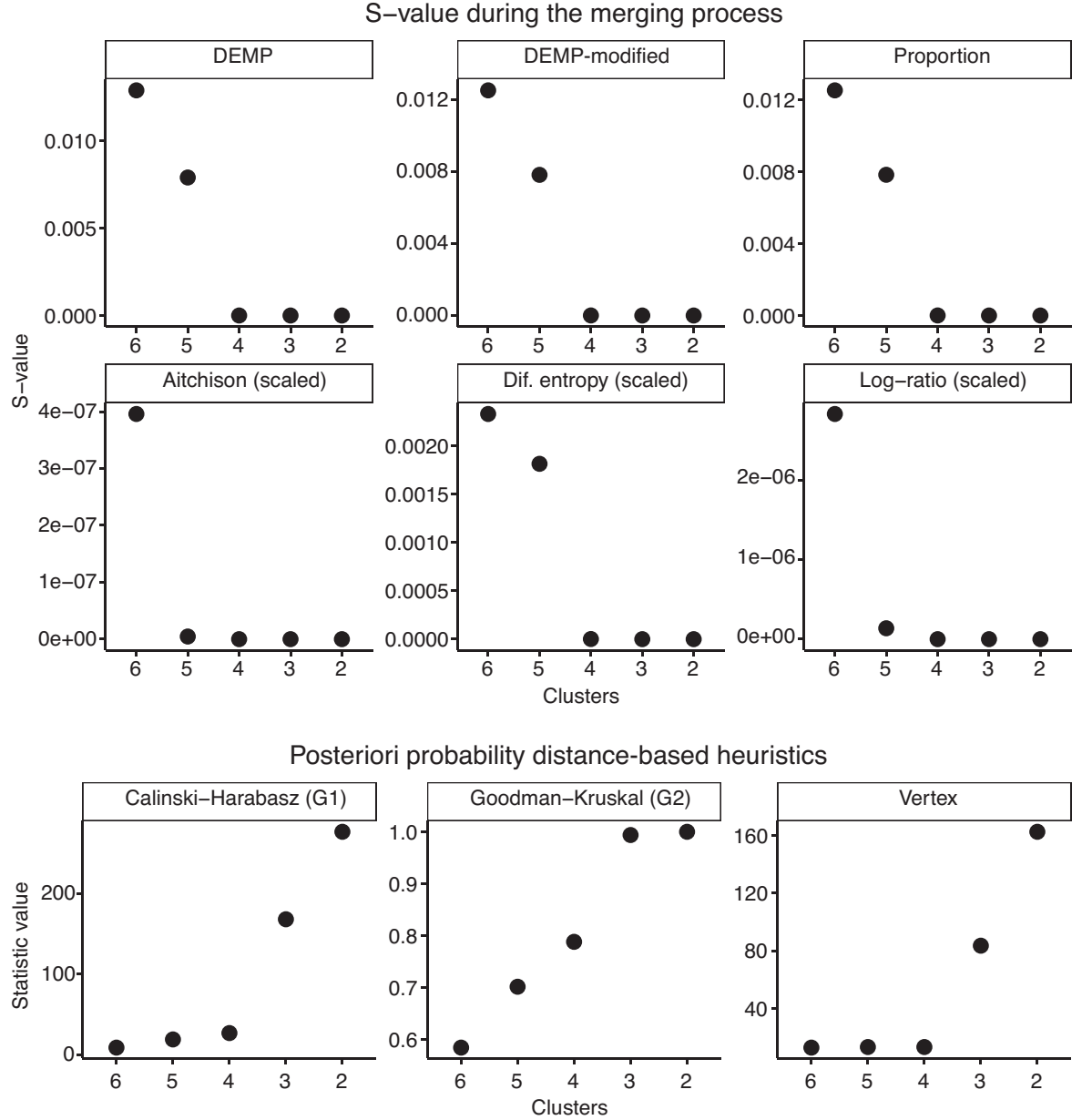
To decide the number of clusters, we plot the  $S$ -values (Figure 9 (top)). The distance-based statistics are plotted in Figure 9 (bottom). In this case, it is difficult to decide the final number of clusters. The  $S$ -values of DEMP, DEMP modified and proportion approaches suggest separating the sample into four clusters. In contrast, the Aitchison distance approach and the log-ratio approach suggest five clusters, and with the difference of entropies, it is difficult to decide. Using the distance-based criterias, we see that the G1 and G2 criteria suggest separating the sample into either three or four clusters. Finally, the closeness to the vertex criteria suggests that the sample is close to the vertex once the sample has been merged into three clusters. With all this information at hand, it seems reasonable to chose either three or four clusters. Figure 10 shows the sample after separating into three clusters. The first cluster contains the mixture components one, two and three; all of them represented by sows with a high amount of feeding time. The second cluster contains the mixture components five and six; this cluster is characterized by sows with high amounts of bed time. Finally, the third cluster is only formed by the component four, which equates to a higher amount of passage time. The reader interested in how to compare groups in compositional data-sets can consult Martín-Fernández et al. (2015).

### 6.3 Simulated example varying the component overlapping in the mixture model

We compared the different  $S_{\omega, \lambda}$  approaches to build a hierarchy in different scenarios as regards the overlapping between mixture components. According Maitra and Melnykov (2010), the maximum overlapping between two components, denoted as  $\varphi$ , is the probability that a given observation generated from one component has a higher posterior probability of being classified to the other component than to itself. In this example, the datasets were generated using the R package 'MixSim' (Melnykov, 2012), and the Gaussian mixture were estimated using the R package 'mclust' (Scrucca et al., 2016). The arguments required to generate the datasets from a mixture of elliptical Gaussian components are: the number of components ( $K_0$ ), the dimension of the observations ( $D$ ) and the maximum overlapping ( $\varphi$ ). The centre and covariance parameters of each mixture component are accordingly calculated by the algorithm (Melnykov, 2012).

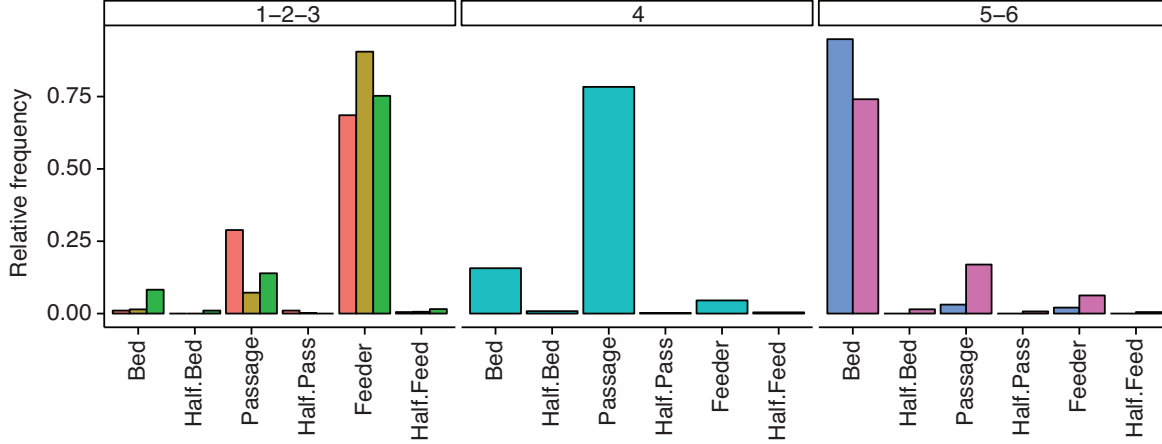
For our illustration purposes, we considered:  $D \in \{2, 3, 4, 5, 6\}$ ,  $K_0 \in \{3, 4, 5\}$  and  $\varphi \in \{0.01, 0.02, \dots, 0.5\}$ . That is, a total of 750 different cases. For each case, 200 datasets were simulated. Indeed, the dataset simulation summarized in Table 3 proceeded as follows:

1. We generated 200 datasets with 500  $D$ -dimensional observations, each coming from a mixture with  $K_0$  elliptical Gaussian components with maximum overlapping  $\varphi$ . We denote by  $\mathcal{C}_0$  the corresponding clustering, where each



**Figure 9** Different criterias to decide the number of clusters for the pigs sample: (Top)  $S$ -values of the different approaches. By rows: DEMP approach (Hennig, 2010) ( $S_{\omega_{prop}, \lambda_{DEMP}}$ ), DEMP-modified approach (Longford and Bartošová, 2014) ( $S_{\omega_{prop}, \lambda_{DEMP_m}}$ ), posterior probability approach ( $S_{\omega_{prop}, \lambda_{prop}}$ ), Aitchison distance with proportional weighing scaled ( $\phi_{dist} \circ S_{\omega_{prop}, \lambda_{dist}}$ ), difference of entropies approach scaled ( $\phi_{\Delta Ent} \circ S_{\omega_{const}, \lambda_{\Delta Ent}}$ ) and log-ratio approach with proportional weighing scaled ( $\phi_{log} \circ S_{\omega_{prop}, \lambda_{log}}$ ). (Bottom) Distance-based criteria for the posterior probability vectors: G1, G2 and closeness to the vertex (Section 5.2)

Merging the components of a finite mixture using posterior probabilities 23



**Figure 10** Pig dataset components after clustering the six mixture components in a 3-FMM. For each cluster, the relative amount of time seen in each location is shown

observation is classified depending on the component from where it was generated.

2. For each of the 200 datasets, we fitted a spherical Gaussian mixture. The number of components  $K$  was estimated using the BIC criteria.
3. We considered all possible combinations of functions  $\omega$  and  $\lambda$ . That is, we took  $\omega \in \{\omega_{\text{cnst}}, \omega_{\text{prop}}, \omega_{\text{dich}}\}$  and  $\lambda \in \{\lambda_{\Delta\text{Ent}}, \lambda_{\text{DEMP}}, \lambda_{\text{DEMP}_m}, \lambda_{\text{prop}}, \lambda_{\text{dist}}, \lambda_{\text{log}}\}$ . For those cases where  $K$  was higher than  $K_0$ , we constructed the hierarchy  $\mathcal{P}_1, \dots, \mathcal{P}_K$ . When  $K$  was lower or equal to  $K_0$ , the dataset was discarded (strikeout numbers in Table 3).
4. To evaluate the performance of a merging algorithm, we applied the Equation (2.3) to the partition  $\mathcal{P}_k$ , when  $k = K_0$ , to form the clustering  $C^*$ . Using the adjusted Rand index, we compared the original clustering  $C_0$  with the corresponding clustering  $C^*$ .

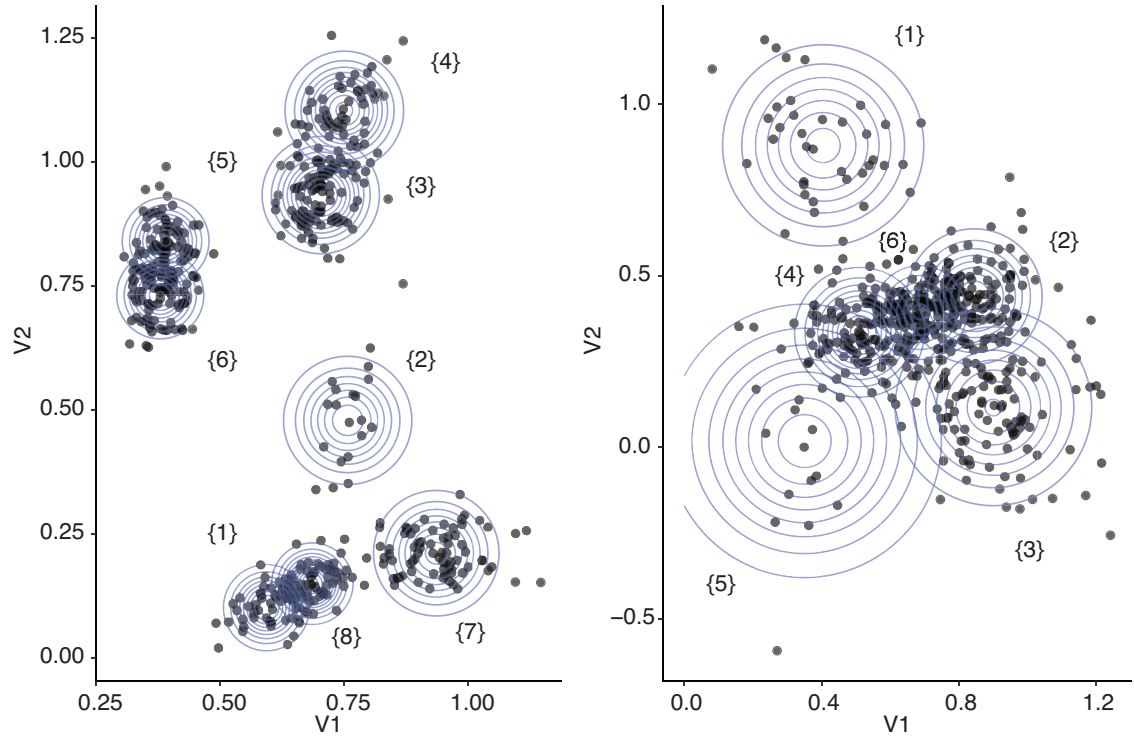
As expected, for most of the datasets the number of fitted spherical components was greater than the original number of elliptical components (Table 3). Figure 11 shows two datasets generated from a five elliptical Gaussian components. The dataset represented in Figure 11 (left) was generated using a low maximum overlapping ( $\varphi = 0.05$ ); whereas for the dataset in Figure 11 (right), we used a high maximum overlapping ( $\varphi = 0.45$ ). Using the BIC criteria, the estimated number of spherical Gaussian components was respectively eight and six.

Figures 12 and 13 show the  $S$ -values obtained using some combinations of functions  $\omega$  and  $\lambda$ : the DEMF-approach (Hennig, 2010) ( $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}}}$ ), the DEMF-modified approach (Longford and Bartošová, 2014) ( $S_{\omega_{\text{prop}}, \lambda_{\text{DEMP}_m}}$ ), the posterior probability approach ( $S_{\omega_{\text{prop}}, \lambda_{\text{prop}}}$ ), the Aitchison distance with proportional weighing scaled ( $\phi_{\text{dist}} \circ S_{\omega_{\text{prop}}, \lambda_{\text{dist}}}$ ), the difference of entropies approach scaled ( $\phi_{\Delta\text{Ent}} \circ S_{\omega_{\text{cnst}}, \lambda_{\Delta\text{Ent}}}$ ) and the log-ratio approach with proportional weighing scaled

24 *Marc Comas-Cufí et al.*

**Table 3** Number of simulated datasets according to its distribution by the original number of elliptical Gaussian components ( $K_0$ ) and the fitted spherical Gaussian components ( $K$ ). Strikeout numbers represent discarded cases ( $K \leq K_0$ )

$K$	$K_0$		
	3	4	5
1	<del>87</del>	<del>66</del>	<del>53</del>
2	<del>840</del>	<del>697</del>	<del>521</del>
3	<del>4891</del>	<del>3140</del>	<del>2003</del>
4	11 517	<del>7032</del>	<del>4546</del>
5	14 173	10 911	<del>7760</del>
6	11 079	10 751	9 936
7	5 270	9 295	10 073
8	1 634	5 211	7 915
9	419	2 187	4 238
10	80	558	2 000
11	10	117	755
12	0	28	181
13	0	7	19



**Figure 11** Dataset generated using a five elliptical Gaussian component mixture: (left) low overlap ( $\varphi = 0.05$ ); (right) high overlap ( $\varphi = 0.45$ ). Isodensity curves for respectively eight and six spherical Gaussian components

## Merging the components of a finite mixture using posterior probabilities 25

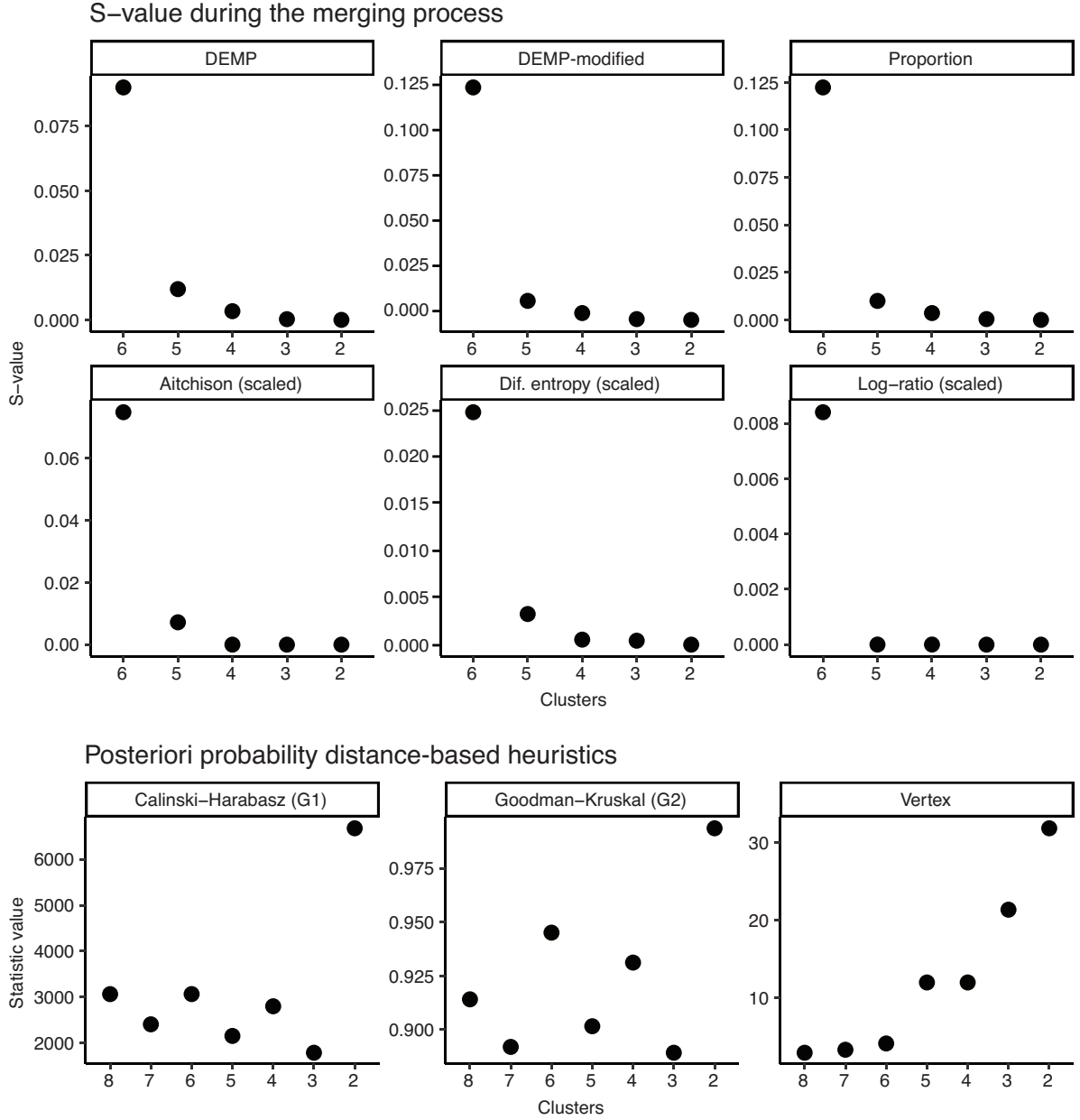
$(\phi_{\log} \circ S_{\omega_{\text{prop}}, \lambda_{\log}})$ . For the dataset with low overlapping (Figure 11 (left)), all the  $S$ -values suggested to form five clusters (Figure 12 (top)). On the other hand, it is more difficult to identify the number of clusters suggested by the distance-based criteria (Figure 12 (bottom)). However, whereas both the G1 and G2 are difficult to interpret, the vertex criteria starts to increase after merging the sample into five components, coherent with the  $S$ -values behaviour. For the high overlapping case (Figure 11 (right)), the  $S$ -values in the Figure 13 (top) suggest to consider only two clusters. Similarly to the low overlapping case, in this scenario the G1 and G2 are difficult to interpret (Figure 13 (bottom)). The vertex criteria slightly increases until merging three components, and jumps into a higher value when two components are considered, a behaviour that is again coherent with the  $S$ -values.

To evaluate the results obtained by the  $S_{\omega, \lambda}$  approaches, we split the datasets of the 750 different cases into three different groups by the level of maximum overlapping: low ( $\varphi \leq 0.15$ ); medium ( $0.15 < \varphi \leq 0.35$ ); and high ( $0.35 < \varphi$ ). With the purpose of improving the interpretation, we also constructed the 'Random' hierarchy, that is the hierarchy obtained by merging two components at random in each step. Table 4 shows the mean and the half width (HW) of its 95% confidence interval of the adjusted Rand index for each method and situation (low, medium and high maximum overlapping). As expected, the Rand index decreases, when the overlap increases. In Table 4, the methods are ordered according to the result obtained in the low maximum overlapping situation. Importantly, this order is mostly preserved in the other overlapping scenarios. No relevant differences were appreciated as regards to the HW of the intervals. According to this classification, the best three methods were the approaches  $S_{\omega_{\text{prop}}, \lambda_{\Delta\text{Ent}}}$ ,  $S_{\omega_{\text{dich}}, \lambda_{\text{prop}}}$  and  $S_{\omega_{\text{dich}}, \lambda_{\text{DEMP}_m}}$ . The corresponding confidence intervals suggest that these methods are not significantly different. The simple method  $S_{\omega_{\text{prop}}, \lambda_{\text{prop}}}$  ranked close to the best methods in all scenarios. Methods presented in Longford and Bartošová (2014), Hennig (2010) and Baudry et al. (2010), respectively, ranked in sixth, seventh and ninth position. The methods proposed in this article ranked in the middle of the classification. Note that three methods,  $S_{\omega_{\text{cst}}, \lambda_{\text{DEMP}}}$ ,  $S_{\omega_{\text{cst}}, \lambda_{\text{prop}}}$  and  $S_{\omega_{\text{cst}}, \lambda_{\log}}$ , were not significantly better than the random merging. In general, the results for the function  $\omega_{\text{cst}}$  were the worst. Except the bad case of  $S_{\omega_{\text{dich}}, \lambda_{\text{DEMP}}}$ , the other two functions  $\omega_{\text{prop}}$  and  $\omega_{\text{dich}}$  provide their best results with the functions  $\lambda_{\Delta\text{Ent}}$ ,  $\lambda_{\text{prop}}$ ,  $\lambda_{\text{DEMP}_m}$  and  $\lambda_{\text{DEMP}}$ .

Following Hamby (1994), to globally evaluate the results obtained by the  $S_{\omega, \lambda}$  approaches for the 750 different cases, we fitted a linear model predicting the adjusted R and index using the following covariates: the method  $S_{\omega, \lambda}$ , the dimension  $D$ , the difference between the number of components estimated and the number of components simulated  $K - K_0$ , and the maximum overlapping  $\varphi$ . All the coefficients of the model were significant where all the associated p-values were ' $< 0.001$ ' and the  $R^2$  obtained was 0.4406. Right-hand column in Table 4 shows the coefficient of each method. The codification of the categorical variable associated to the method indicates that the constant coefficient of the model (intercept) corresponds to the Random method (0.601). The coefficient of the other methods should be interpreted



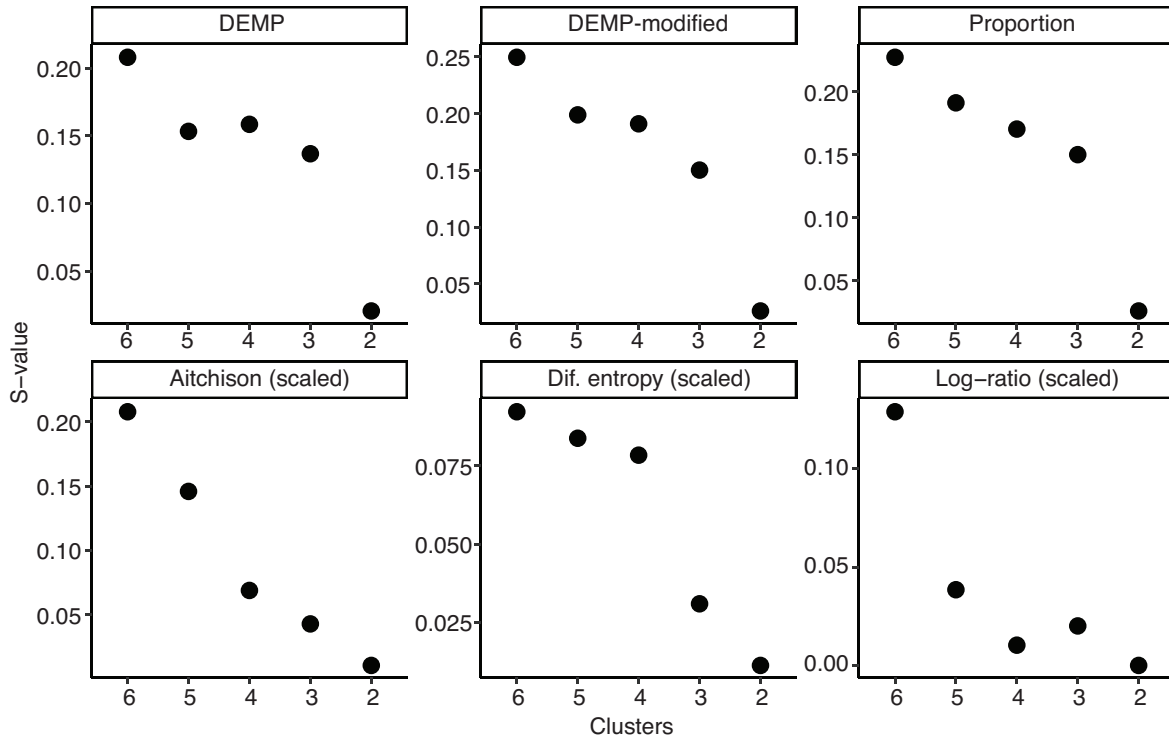
26 Marc Comas-Cufí et al.



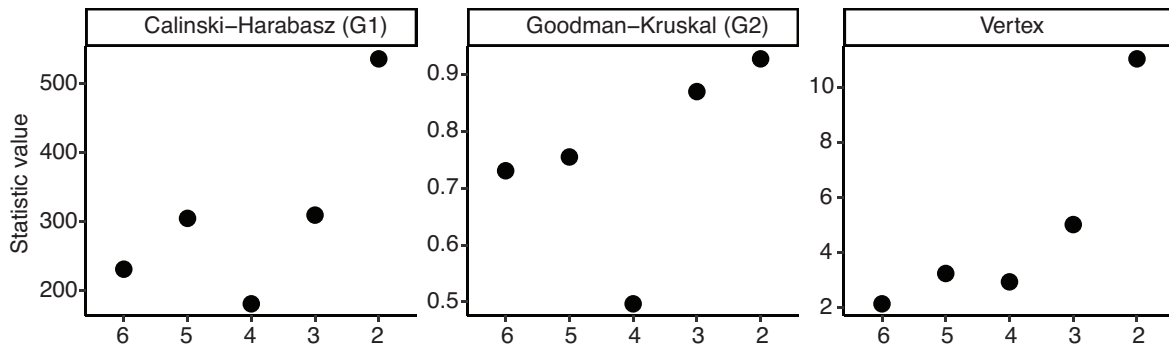
**Figure 12** Low overlapping case: different criteria to decide the number of clusters. (Top) S-values of different approaches. By rows: DEMP approach (Hennig, 2010) ( $S_{\omega_{prop}, \lambda_{DEMP}}$ ), DEMP-modified approach (Longford and Bartošová, 2014) ( $S_{\omega_{prop}, \lambda_{DEMP_m}}$ ), posterior probability approach ( $S_{\omega_{prop}, \lambda_{prop}}$ ), Aitchison distance with proportional weighing scaled ( $\phi_{dist} \circ S_{\omega_{prop}, \lambda_{dist}}$ ), difference of entropies approach scaled ( $\phi_{\Delta Ent} \circ S_{\omega_{const}, \lambda_{\Delta Ent}}$ ) and log-ratio approach with proportional weighing scaled ( $\phi_{log} \circ S_{\omega_{prop}, \lambda_{log}}$ ); (Bottom) Distance-based criteria: G1, G2 and closeness to the vertex (Section 5.2)

*Merging the components of a finite mixture using posterior probabilities* 27

S-value during the merging process



Posteriori probability distance-based heuristics



**Figure 13** High overlapping case: different criteria to decide the number of clusters. (Top) S-values of different approaches. By rows: DEMP approach (Hennig, 2010) ( $S_{\omega_{prop}, \lambda_{DEMP}}$ ), DEMP-modified approach (Longford and Bartošová, 2014) ( $S_{\omega_{prop}, \lambda_{DEMP_m}}$ ), posterior probability approach ( $S_{\omega_{prop}, \lambda_{prop}}$ ), Aitchison distance with proportional weighing scaled ( $\phi_{dist} \circ S_{\omega_{prop}, \lambda_{dist}}$ ), difference of entropies approach scaled ( $\phi_{\Delta Ent} \circ S_{\omega_{cnst}, \lambda_{\Delta Ent}}$ ) and log-ratio approach with proportional weighing scaled ( $\phi_{log} \circ S_{\omega_{prop}, \lambda_{log}}$ ). (Bottom) Distance-based criteria: G1, G2 and closeness to the vertex (Section 5.2)

28 *Marc Comas-Cufí et al.*

**Table 4** Mean of the adjusted Rand index for each approach  $S_{\omega,\lambda}$  by three overlapping levels: low ( $\varphi \leq 0.15$ ), medium ( $0.15 < \varphi \leq 0.35$ ) and high ( $0.35 < \varphi$ ); Random means the random merging method. In parentheses the half width (HW) of the 95% confidence interval; right-hand column: beta ( $\beta$ ) coefficient in the linear model

	Adjusted Rand index (HW)			$\beta$
	$\varphi \leq 0.15$	$0.15 < \varphi \leq 0.35$	$0.35 < \varphi$	
$\omega_{prop} - \lambda_{\Delta Ent}$	0.836 (0.009)	0.539 (0.006)	0.359 (0.006)	0.282
$\omega_{dich} - \lambda_{prop}$	0.835 (0.009)	0.538 (0.006)	0.359 (0.006)	0.281
$\omega_{dich} - \lambda_{DEMP_m}$	0.835 (0.009)	0.538 (0.006)	0.358 (0.006)	0.281
$\omega_{prop} - \lambda_{prop}$	0.834 (0.008)	0.534 (0.006)	0.356 (0.006)	0.278
$\omega_{dich} - \lambda_{\Delta Ent}$	0.833 (0.009)	0.535 (0.006)	0.356 (0.006)	0.278
$\omega_{prop} - \lambda_{DEMP_m}$	0.833 (0.009)	0.530 (0.006)	0.352 (0.006)	0.275
$\omega_{prop} - \lambda_{DEMP}$	0.826 (0.009)	0.528 (0.006)	0.353 (0.006)	0.272
$\omega_{prop} - \lambda_{log}$	0.819 (0.009)	0.521 (0.006)	0.347 (0.006)	0.266
$\omega_{dich} - \lambda_{log}$	0.812 (0.009)	0.513 (0.006)	0.342 (0.006)	0.259
$\omega_{cnst} - \lambda_{\Delta Ent}$	0.795 (0.008)	0.520 (0.006)	0.353 (0.006)	0.259
$\omega_{prop} - \lambda_{dist}$	0.796 (0.009)	0.500 (0.006)	0.332 (0.006)	0.245
$\omega_{dich} - \lambda_{dist}$	0.795 (0.009)	0.499 (0.006)	0.332 (0.006)	0.245
$\omega_{cnst} - \lambda_{dist}$	0.696 (0.008)	0.477 (0.006)	0.329 (0.006)	0.203
$\omega_{cnst} - \lambda_{DEMP_m}$	0.517 (0.007)	0.397 (0.005)	0.292 (0.006)	0.102
$\omega_{dich} - \lambda_{DEMP}$	0.393 (0.006)	0.324 (0.005)	0.249 (0.006)	0.020
Random	0.364 (0.006)	0.305 (0.005)	0.237 (0.005)	0.601
$\omega_{cnst} - \lambda_{DEMP}$	0.335 (0.006)	0.280 (0.005)	0.219 (0.005)	-0.024
$\omega_{cnst} - \lambda_{prop}$	0.334 (0.006)	0.280 (0.005)	0.219 (0.005)	-0.025
$\omega_{cnst} - \lambda_{log}$	0.304 (0.005)	0.273 (0.005)	0.217 (0.005)	-0.038

as the increase or decrease in relation to the random merging. Note that both the sign and order of these coefficients are in full agreement with the value of the mean. The negative coefficients correspond to the approaches that have worse mean than the Random method. The coefficients with the greatest value correspond to the best methods, and they have the same order than the mean values in the left-hand column. The coefficient of the variable associated to the dimension was  $-0.028$ , suggesting that when the dimension increases, the results are worse. The same interpretation has the coefficient of the maximum overlapping  $\varphi$  that took the value  $-0.956$ , with a decrease of the adjusted R and index, when the overlap increases. The coefficient of the difference  $K - K_0$  in the model was  $0.023$ , suggesting that a high number of components estimated captures reasonably well the grouping of the dataset. The merging of the  $K$  components preserves the structure to finally fit the components simulated.

## 7 Final remarks

When FMM is used in clustering, the question ‘is a cluster determined by a unique component?’ emerges. Different authors have proposed scenarios where it seems reasonable to argue that a cluster can be better modelled by more than one single component or, equivalently, modelled by an FMM itself. In these same scenarios, the approaches proposed in this article may be of interest.

*Merging the components of a finite mixture using posterior probabilities* 29

In this article, we propose a generic approach to build a hierarchy of mixture components, which relies only on the posterior probability vectors, and therefore, it is independent from the family of probability distributions. This generic approach allows us to both integrate some criterias that had appeared earlier in the literature and to propose some new techniques. As the posterior probability vectors belong to the simplex sample space, we use the log-ratio methodology developed for this type of vector to develop these new techniques. All the methods described in this article can be applied to any FMM.

To decide the final number of clusters, we have proposed combining the information given by the scaled  $S$ -values, and we have also developed some criterias based on the location of the posterior probabilities. For the latter, we propose the G1 and G2 indices, using Aitchison distances between posterior probabilities vectors and the closeness to the vertex criteria. To the best of our knowledge, it is the first time that these two options have been proposed to study cluster structure. This new approach allows the structure of numerical and categorical datasets based on the considered model to be studied.

Finally, the log-ratio approaches introduced in this article (Section 4) and the indices defined on the posterior probabilities (Section 5.2) use the geometric structure of the simplex space. Working with the posterior probabilities as an element of the simplex space allows the results obtained here to be extended to other methods. For example, in fuzzy clustering, it would be interesting to analyse if the role played by the weights is equivalent to the role played by the posterior probabilities.

## Acknowledgements

This research has been supported by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS (Ref: MTM2015-65016-C2-1(2)-R). The authors gratefully acknowledge the constructive comments of the anonymous referees which have undoubtedly helped to significantly improve the quality of this article.

## References

- Aitchison J (1986). *The statistical analysis of compositional data (Monographs on statistics and applied probability)*. London: Chapman & Hall Ltd. Reprinted with additional material, Caldwell, NJ: The Blackburn Press, 2003.
- (2002) Simplicial inference. *Algebraic Methods in Statistics and Probability*, 287, 1–22.
- Baudry JP, Raftery AE, Celeux G, Lo K and Gottardo R (2010) Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19, 332–53.
- Benaglia T, Chauveau D, Hunter DR and Youn R (2009) mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32, 1–29.
- Comas-Cufí M, Martín-Fernández JA and Mateu-Figueras G (2016) Logratio methods in mixture models for compositional data sets. *SORT*, 40, 349–74.

30 *Marc Comas-Cufí et al.*

- Fraley C and Raftery AE (1998) How many clusters? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578–88.
- (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–31.
- Frey BJ and Dueck D (2007) Clustering by passing messages between data points. *Science*, **315**, 972–76.
- Hamby D (1994) A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, **32**, 135–54.
- Hennig C (2010) Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, **4**, 3–34.
- Keribin C (1998) Consistent estimate of the order of mixture models. *Comptes Rendues de l'Academie des Sciences, Série I-Mathématiques*, **326**, 243–48.
- (2000) Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, **62**, 49–66.
- Lee HJ and Cho S (2004) Combining Gaussian mixture models. In *Intelligent Data Engineering and Automated Learning—IDEAL 2004 SE–98*, edited by Z Yang, H Yin and R Everson. Volume 3177 of *Lecture Notes in Computer Science*, pages 666–71. Exeter, UK: Springer Berlin Heidelberg.
- Longford NT and Bartošová J (2014) A confusion index for measuring separation and clustering. *Statistical Modelling*, **14**, 229–55.
- Maitra R and Melnykov V (2010) Simulating data to study performance of finite mixture modelling and clustering algorithms. *Journal of Computational and Graphical Statistics*, **19**, 354–76.
- Martín-Fernández JA, Daunis-i-Estadella J and Mateu-Figueras G (2015) On the interpretation of differences between groups for compositional data. *SORT*, **39**, 231–52.
- Martín-Fernández JA and Thio-Henestrosa S (editors) (2016) *The compositional data analysis: CoDaWork, L'Escala, Spain, June 2015. Springer Proceedings in Mathematics & Statistics*, 187. New York, NY: Springer International Publishing.
- McLachlan GJ and Rathnayake S (2014) On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4**, 341–55.
- Melnykov V (2013) On the distribution of posterior probabilities in finite mixture models with application in clustering. *Journal of Multivariate Analysis*, **122**, 175–89.
- (2016) Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics*, **25**, 66–90.
- Melnykov V, Chen, WC and Maitra R (2012) MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, **51**, 1–21.
- Milligan GW and Cooper MC (1985) An examination of procedures for determining the number of clusters. *Psychometrika*, **50**, 159–79.
- Palarea-Albaladejo J and Martín-Fernández JA (2015) zCompositions: R packages for multivariate imputation of nondetecteds and zeros in compositional data sets. *Chemo-metrics and Intelligent Laboratory Systems*, **143**, 85–96.
- Palarea-Albaladejo J, Martín-Fernández JA and Soto JA (2012) Dealing with distances and transformations for fuzzy C-means clustering of compositional data. *Journal of Classification*, **29**, 144–69.
- Pastore A and Tonellato SF (2013) *A merging algorithm for Gaussian mixture components* (Series No. 04/WP/2013). [SSRN Electronic Journal]. Department of Economics Research Paper, University Ca' Foscari of Venice. URL [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2233307](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2233307) (last accessed on 23 November 2017).
- Pawlowsky-Glahn V and Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *SERRA*, **15**, 384–98.
- Punzo A (2014) Flexible mixture modelling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, **14**, 257–91.
- R Development Core Team (2015) *R: A language and environment for statistical computing*.

*Merging the components of a finite mixture using posterior probabilities* 31

- Vienna: R Foundation for Statistical Computing. URL <http://www.r-project.org> (last accessed on 23 November 2017).
- Ray S and Lindsay BG (2005) Topography of multivariate normal mixtures. *Annals of Statistics*, **33**, 2042–65.
- Scrucca L, Fop M, Murphy TB and Raftery AE (2016) mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, **8**, 289–317.



### 4.3 Computational Statistics & Data Analysis

El tercer article (enviat) cobreix els objectius Obj. 2 i Obj. C descrits a la secció 2.1. L'article proposa una nova distribució de probabilitat pel modelatge de dades provinents de comptatges. La distribució proposada apareix de la mixtura d'una distribució normal definida en el Símplex amb la distribució multinomial clàssica. A l'article es compara la capacitat de modelització enfront de la distribució equivalent basada en la distribució Dirichlet.

L'article ha estat enviat a la revista Computational Statistics & Data Analysis.

Enviat: Abril 2018

Factor d'impacte: 1.693 (Q1).





## Capítol 5

# Resultats i discussió

En aquesta secció repassem els principals resultats que es deriven d'aquesta tesi, principalment pel que fa als models de mixtures finites de distribucions definides en el Símplex, els mètodes per combinar les components d'una mixtura finita de distribucions a partir de les probabilitats a posterior de pertinença a la component, i finalment, la distribució per dades de comptatge construïda com la mixtura de la distribució multinomial com a funció nucli i la distribució normal definida al Símplex com a funció de barreja.

### 5.1 Mixtures finites de distribucions definides en el Símplex

Per il·lustrar els diferents mètodes existents per la definició de mixtures finites de distribucions en el Símplex, es considerarà un conjunt de dades de diferents tipus de vidre provinents d'envasos, de focus d'automòbil i de les finestres del vehicle. Aquest conjunt de dades va ser extret del repositori online *UCI Machine Learning repository*. Per a poder visualitzar els resultats en un diagrama ternari (veure Secció 3.1) de les mostres de vidre únicament considerarem les característiques que fan referència a la quantitat de Calci (Ca), de Silici (Si) i d'Alumini (Al). La mostra de vidres la podem veure a la Figura 5.1. A la part superior esquerra del gràfic es veu la localització global de les mostres de vidre, a l'estar situades molt a prop del vèrtex del Silici podem concloure que a totes les observacions hi predomina aquest element. En el gràfic principal s'ha ampliat el triangle inferior dret del diagrama ternari, concretament la regió que conté mostres amb una quantitat superior al 80% de Silici.

Una de les principals limitacions en ajustar un model de mixtura de

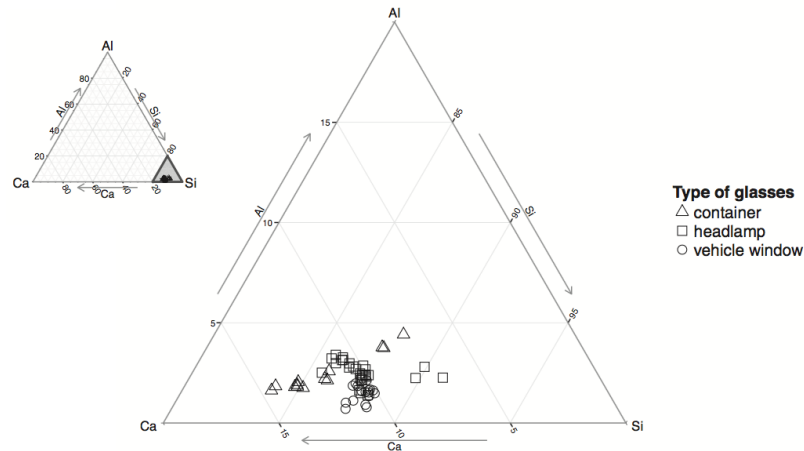


Figura 5.1: Observacions de tres components provinents de vidres de diferents zones.

distribucions de l'espai real a un conjunt de dades definides en el Símplex, és el fet que la suma de les components de cada una de les observacions és una constant. Aquesta característica fa que la matriu de covariància de les dades sigui singular. A la pràctica, això implica que moltes funcions de densitat no estiguin correctament definides. Per resoldre aquest problema, una de les aproximacions proposades per diferents autors consisteix en la no consideració d'una de les components. Com que aquesta component eliminada es pot recuperar fàcilment a partir de la resta, el procés fa que no es perdi informació. L'enfocament fou considerat per Papageorgiou *et al.* (2001) per classificar un conjunt de 45 peces de ceràmica a partir dels elements químics de cada una de les peces, els autors modelaren el conjunt de dades a partir d'una mixtura finita de distribucions normals definida en totes excepte una de les components.

Utilitzant el conjunt de dades introduït al començament d'aquesta secció i seguint el mètode proposat, hem ajustat una mixtura de distribucions normals a les dades considerant únicament l'Alumini i el Silici. A la Figura 5.2 (esquerra) hi han representades les corbes d'igual probabilitat de la funció de distribució resultant. Mirant el gràfic es pot observar com algunes de les corbes estan definides en la zona de valors d'Alumini negatiu (zona inferior a la línia ratllada). Quan aquesta nova distribució la transformem altre cop al Símplex i la visualitzem en el diagrama Ternari (Figura 5.2 (dreta)) veiem

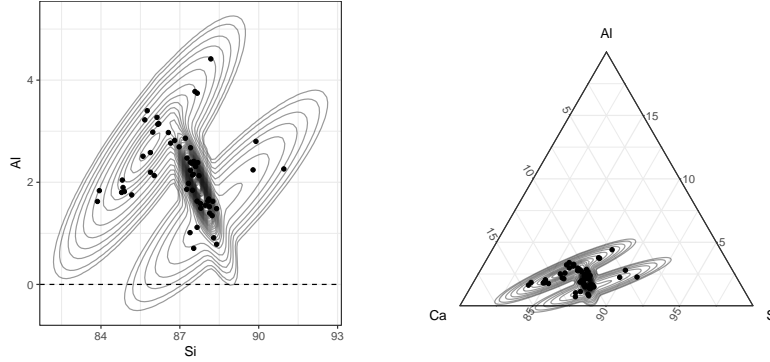


Figura 5.2: Mixtura finita gaussiana definida a  $\mathbb{R}^2$  a través de l'eliminació d'una component d'una mostra definida a  $\mathcal{S}^3$ .

com la distribució passa a estar definida en una regió impossible. Com a conclusió, acabem de veure que al considerar l'enfocament utilitzat a (Papa-georgiou *et al.*, 2001) estem assignant probabilitat positiva a esdeveniments que són impossibles.

Com a solució al problema anterior, es podria treballar directament amb distribucions definides en el Símplex. Per exemple, la distribució Dirichlet és la distribució més coneguda que té el seu domini a l'espai del Símplex, per tant, és natural que a la literatura s'hi hagi considerat la mixtura d'aquesta distribució (Albert i Gupta, 1982; Bouguila *et al.*, 2004; Calif *et al.*, 2011). Tal com s'ha comentat a la Secció 3.1, la distribució Dirichlet té la limitació que totes les components han estat generades com a variables independents univariants seguint una distribució Gamma. Això fa que la capacitat de modelització d'aquesta distribució sigui molt limitada. Això ho podem veure fàcilment amb l'exemple dels vidres on s'hi ha ajustat una mixtura de tres distribucions Dirichlet (Figura 5.3). Com s'hi pot veure, a diferència de l'enfocament anterior, és difícil identificar-hi cap tendència a les tres components, i únicament sembla identificar-s'hi acumulacions de punts al voltant de tres centres. Una altra limitació important d'aquest enfocament, és la manca d'algoritmes eficients per a estimar-ne els paràmetres. Així per exemple, en l'estimació dels paràmetres de la mixtura de la Figura 5.3 per màxima versemblança, s'han hagut d'utilitzar mètodes d'aproximació específics per les dades d'aquest problema.

Finalment, seguint l'aproximació introduïda a Mateu-Figueras (2003) per definir la distribució normal a  $\mathcal{S}^D$  podem treballar directament a l'espai de coordenades isomètriques definides a  $\mathbb{R}^{D-1}$ . D'aquesta manera, podem

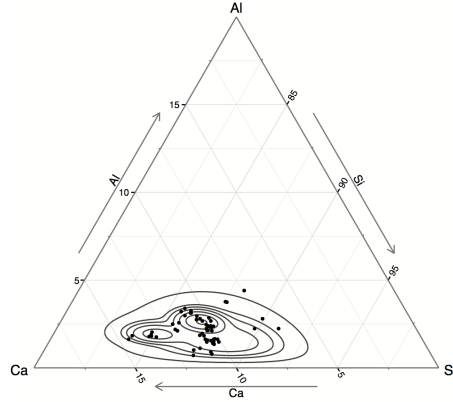


Figura 5.3: Mixtura finita de distribucions Dirichlet ajustada al conjunt reduït de vidres.

dir que una composició aleatòria  $\boldsymbol{x}$  segueix una mixtura de distribucions normal a  $\mathcal{S}^D$  si les coordenades isomètriques  $\boldsymbol{h}$  segueixen una mixtura de distribucions normals a  $\mathbb{R}^{D-1}$ . A la Figura 5.4(esquerra) podem veure-hi l'ajust de la mixtura de tres components normal a l'espai de coordenades isomètriques ( $\mathbb{R}^2$ ). La mixtura identifica tres direccions marcadament diferents dins el conjunt de dades. A la Figura 5.4(dreta) tenim la mateixa distribució representada a  $\mathcal{S}^3$ .

Aquest nou enfocament és el més coherent amb l'estructura geomètrica del Símplex, i resol totes les limitacions de les propostes anteriors: la distribució únicament dona probabilitat a esdeveniments possibles, l'estructura de covariància provinent de la distribució normal permet identificar direccions rellevants a les dades, i finalment, l'estimació dels paràmetres es pot fer directament amb l'algoritme clàssic EM (Secció 3.2) dins l'espai de coordenades.

El ventall de distribucions que sorgeixen al definir mixtures de distribucions directament a l'espai de coordenades, permet estendre les mixtures de distribucions en el Símplex amb totes les mixtures de distribucions actuals que estan definides a l'espai real: mixtures de normal asimètriques, mixtures de distribucions  $T$ , etc.

Les conclusions explicades han servit per construir el fil principal de l'article:

Marc Comas-Cufí, Josep A. Martín-Fernández i Glòria Mateu-Figueras (2016). Log-ratio methods in mixture models for compositional data sets.

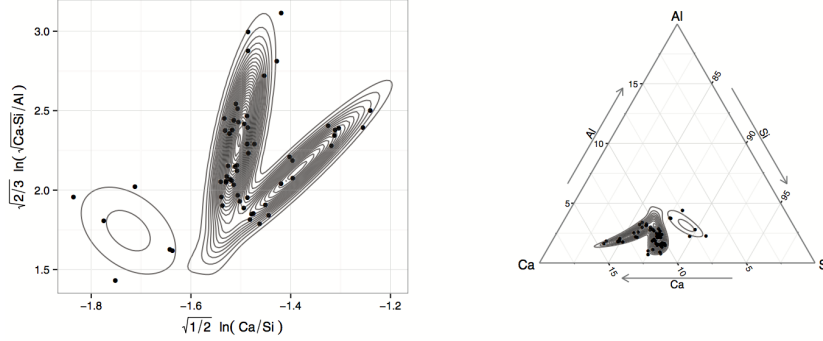


Figura 5.4: Mixtura finita de distribucions normals sobre  $\mathcal{S}^3$  ajustada al conjunt reduït de vidres. A l'esquerra la representació de la mixtura a l'espai de coordenades. A la dreta la representació de la mixtura al diagrama ternari.

*Statistics and Operations Research Transactions*, 40(2): 349–374.

Aquest article serveix per assentar les bases de com definir una mixtura de distribucions en el Símplex mitjançant l'espai de coordenades isomètriques. A més, en un futur, aquest article servirà com a referència principal a l'utilitzar aquest enfocament en treballs aplicats.

## 5.2 Combinació de les components d'una mixtura

A la Secció 3.2 hem comentat que un cop ajustats els paràmetres d'una mixtura de distribucions finita, per cada observació podem calcular el que s'anomenen les probabilitats a posteriori de pertinença a una component (equació 3.9). A partir únicament d'aquestes probabilitats, a la literatura s'han presentat diferents enfocaments per a decidir quines components poden ser considerades components que modelen un únic conjunt de dades (Hennig, 2010; Baudry *et al.*, 2010; Longford i Bartosova, 2014; Melnykov, 2016).

A l'article

Marc Comas-Cufí, Josep A. Martín-Fernández i Glòria Mateu-Figueras (2017). Merging the components of a finite mixture using posterior probabilities . *Statistical Modelling*, In press.

disponible a la pàgina 71 s'introdueix una nova formulació que permet

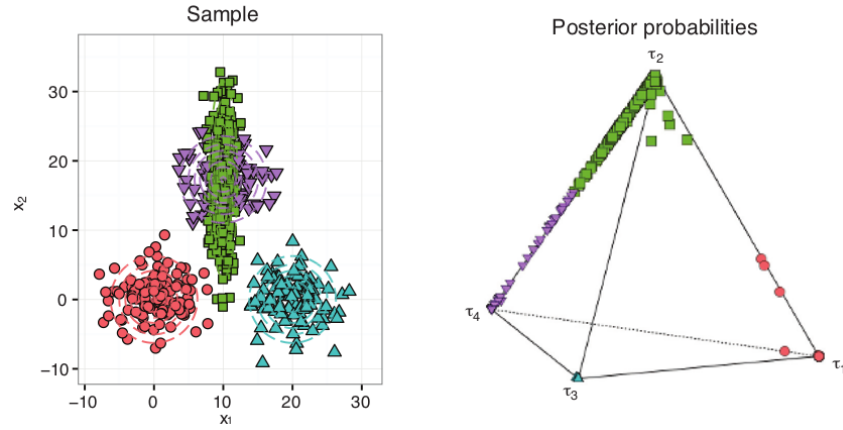


Figura 5.5: Mixtura de quatre distribucions normals i la representació de les probabilitats a posteriori a  $\mathcal{S}^4$ .

unificar els actuals mètodes de combinació de components a partir de les probabilitats a posteriori. Al final, d'una manera informal i reduint-ne la notació, la decisió d'agrupar es fa mitjançant una expressió del tipus

$$S_{a,b} = \frac{\sum_{i=1}^n \omega_i(a, b) \lambda_i(a, b)}{\sum_{i=1}^n \omega_i(a, b)} \quad (5.1)$$

on s'avalua quant bo seria combinar les components  $a$  i  $b$ . Les funcions  $\omega$  i  $\lambda$  són funcions que únicament depenen de les probabilitats a posteriori del fet que una observació pertanyi a les diferents components. El rol que agafa la funció  $\lambda$  és el d'utilitat, mesurant quant bo és per un individu combinar les components  $a$  i  $b$ . La funció  $\omega$  dona més o menys rellevància als resultats finals depenen de la localització de l'observació. A l'article es mostra com les propostes (Hennig, 2010; Baudry *et al.*, 2010; Longford i Bartosova, 2014) són equivalents a l'expressió 5.1 per diferents definicions de les funcions  $\omega$  i  $\lambda$ . A més, s'aprofita la nova formulació per definir nous enfocaments aplicant la metodologia logquocient al vector de probabilitats a posteriori.

Per a decidir el nombre final de grups, o dit d'altra manera, per decidir el criteri de parada en el procés de combinar components. Es consideren dues propostes noves. La primera consisteix a analitzar els valors  $S_{a,b}^*$  òptims obtinguts en el procés de combinar components i observar en quin moment el valor  $S_{a,b}^*$  decau. La segona proposta consisteix a estudiar la localització de les probabilitats a posteriori dins el Símplex. Per exemple, si mirem l'esquerra de la Figura 5.5 podem veure un conjunt de dades acolorides se-

gons la classificació obtinguda per probabilitat a posterior màxim quan es considera una mixtura de 4 components. Les corbes d'isoprobabilitat d'aquesta mixtura de distribucions també han estat representades en el gràfic. Aquestes probabilitats a posteriori són elements de  $\mathcal{S}^4$ , i per tant, poden ser representats en un diagrama quaternari (a la dreta de la Figura 5.5).

Per il·lustrar la metodologia es presenten diferents exemples: un exemple provinent de Baudry *et al.* (2010), un exemple basat amb combinar les components d'una mixtura de distribucions multinomials amb dades obtingudes del paquet *zCompositions* (Palarea-Albaladejo i Martín-Fernández, 2015). Finalment, es realitza una simulació per a veure quin és el comportament dels diferents mètodes en un escenari concret.

### 5.3 La distribució logquocient-normal multinomial

Un cas particular de distribució analitzada en el primer article, és la distribució obtinguda en compondre la distribució normal en el Símplex com a funció nucli amb la distribució categòrica com a funció de barreja. Aquesta distribució és la coneguda mixtura finita de distribucions normal en el Símplex. Com que els paràmetres de la distribució multinomial són elements del Símplex, en la part final de la tesi es va proposar intercanviar el rol de la funció nucli i funció barreja. Concretament, es va proposar estudiar la composició de la distribució multinomial com a funció nucli i la distribució normal en el Símplex com a funció barreja. Aquesta nova distribució anomenada la distribució logquocient-normal multinomial és el que motiva l'article enviat recentment a la revista indexada *Computational Statistics & Data Analysis*:

Marc Comas-Cufí, Josep A. Martín-Fernández, Glòria Mateu-Figueras i Javier Palarea-Albaladejo (2018). Modelling count data using the logratio-normal-multinomial distribution. *Computational Statistics & Data Analysis*, Submitted.

A l'article, a part de derivar algunes de les principals propietats de la nova distribució, es proposa un nou mètode d'estimació basat en la combinació de simulacions de Monte Carlo i iteracions de l'algoritme EM presentat a la Secció 3.2. Per obtenir un millor estimador dels paràmetres de la nova distribució, en lloc de treballar amb generadors de valors pseudoaleatoris, es proposa treballar amb generadors de valors quasi aleatoris: seqüències de Halton o Sobol.



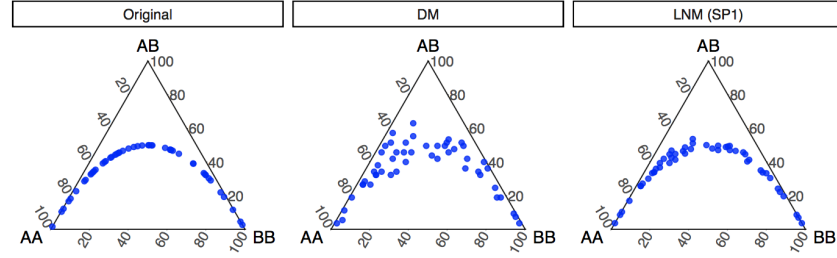


Figura 5.6: Probabilitats reals d'una mostra en equilibri de Hardy-Weinberg i les estimacions fetes amb la distribució Dirichlet-multinomial i la logquocient-normal multinomial.

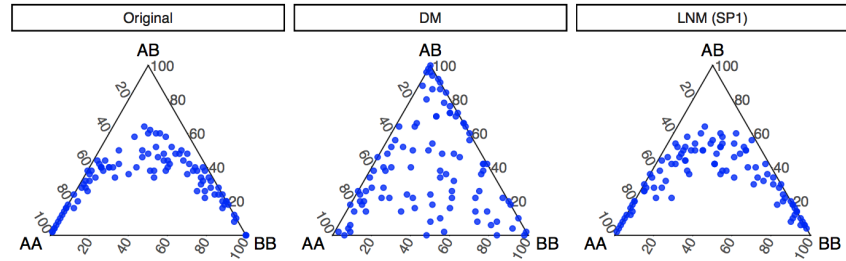


Figura 5.7: Mostra en equilibri de Hardy-Weinberg i generació aleatòria feta amb la distribució Dirichlet-multinomial i la logquocient-normal multinomial ajustades a la mostra original.

Finalment, a l'article es realitza una comparació de la capacitat per modelar diferents situacions entre la distribució logquocient-normal multinomial (LNM) i la més coneguda distribució Dirichlet-multinomial (DM). Per exemple, a la Figura 5.6 es poden veure probabilitats situades en l'equilibri de Hardy-Weinberg definit com la recta composicional  $2 \log(AB) - \log(AA) - \log(BB) = 0$ . Al centre, les estimacions de les probabilitats de la distribució multinomial condicionades a la mostra observada i els paràmetres estimats d'una distribució DM. A la dreta, les estimacions equivalents realitzades amb la distribució LNM. Com es pot observar, les estimacions realitzades amb la distribució LNM són més properes a l'equilibri de Hardy-Weinberg que les estimacions fetes amb la distribució DM.

Aquesta diferència s'accentua més a la Figura 5.7. A l'esquerra es pot veure una mostra (comptatges) d'una població en equilibri de Hardy-Weinberg, concretament és la població generada amb les probabilitats de la Figura 5.6 (esquerra). Al centre, es pot veure una generació aleatòria

realitzada amb una distribució DM utilitzant els paràmetres estimats per màxima versemblança. A la dreta, es veu una generació aleatòria realitzada amb la distribució LNM utilitzant els paràmetres estimats per màxima versemblança. Com es pot apreciar, la sobre dispersió generada per la distribució DM és molt superior a la dispersió observada a la mostra original.

A part de l'exemple de Hardy-Weinberg, a l'article es presenten dos escenaris més: una simulació basada en la distribució multinomial i un exemple que intenta recrear un sondeig realitzat per predir resultats electorals. En tots dos casos, la distribució LNM obté millors resultats.

## 5.4 Conclusions

Tal com s'ha observat en els diferents treballs d'aquesta tesi, ja sigui perquè estem treballant amb mostres definides en el Símplex o bé estem modelant paràmetres d'aquest mateix espai, els models de mixtures de distribucions es poden beneficiar de l'anàlisi de dades composicional.

El primer article ha servit per formalitzar la definició de les mixtures de distribucions definides en el Símplex a partir de l'espai de coordenades. Aquest enfocament permet estendre la flexibilitat a l'hora de definir distribucions en l'espai real directament al Símplex.

A partir del segon article d'aquesta tesi s'ha mostrat que les eines composicionals poden ser utilitzades més enllà de les observacions. En aquests treballs s'ha mostrat que l'anàlisi composicional també pot ser introduït en l'anàlisi o modelització dels paràmetres de diferents models de probabilitat: en el segon article l'anàlisi composicional ha permès analitzar les probabilitats a posteriori des d'un punt de vista composicional i construir nous mètodes per a decidir si dues agrupacions formen o no una única agrupació, en el tercer article l'anàlisi composicional ha servit per a modelar els paràmetres de la funció nucli d'una mixtura, construint d'aquesta manera una nova distribució dins l'espai de comptatges.

## 5.5 Futures línies d'investigació

Durant la realització de la tesi s'han obert noves línies d'investigació relacionades amb els resultats obtinguts en aquest compendi.

- El primer i segon article va obrir la porta a una col·laboració amb la Universitat de Florència per analitzar les característiques geoquímiques d'aigües. El treball aplicat barrejarà els resultats obtinguts en el primer i segon article.

- El tercer article ha fet patent la necessitat de trobar noves maneres d'estimar els paràmetres de la distribució logquocient normal multinomial que siguin més eficients d'un punt de vista computacional.
- El tercer article ha obert una nova línia d'investigació pel reemplaçament de zeros provinents de comptatges.
- El tercer article també obre la porta a una nova metodologia per a la modelització de la sobredispersió en models lineals generalitzats.

# Bibliografía

- AITCHISON, J. I SHEN, S.M. Logistic-Normal Distributions: Some Properties and Uses. *Biometrika* **67**(2):261–272 (1980)
- AITCHISON, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**:139–177 (1982). Reprinted in 2003 with additional material by The Blackburn Press.
- AITCHISON, J. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (1986)
- AITKIN, M., FRANCIS, B., HINDE, J. I DARNELL, R. *Statistical Modelling in R*. Oxford University Press, New York (2009)
- ALBERT, J.H. I GUPTA, A.K. Mixtures of Dirichlet Distributions and Estimation in Contingency Tables. *The Annals of Statistics* **10**(4):1261–1268 (1982)
- ANDREWS, J.L. I MCNICHOLAS, P.D. Model-based Clustering, Classification, and Discriminant Analysis via Mixtures of Multivariate t-distributions: The tEIGEN Family. *Statistics and Computing* **22**:1021–1029 (2012)
- BANFIELD, J. I RAFTERY, A.E. Model-based Gaussian and Non-Gaussian Clustering. *Biometrics* **49**:803–821 (1993)
- BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J.A. I PAWLOWSKY-GLAHN, V. The Mathematics of Compositional Data Analysis. *Austrian Journal of Statistics* **45**(4):57–71 (2016)
- BAUDRY, J.P., RAFTERY, A.E., CELEUX, G., LO, K. I GOTTARDO, R. Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics* **19**(2):332–353 (2010)

- BILLHEIMER, D., GUTTORP, P. I FAGAN, W.F. Statistical Interpretation of Species Composition. *Journal of the American Statistical Association* **96**(456):1205–1214 (2001)
- BÖHNING, D. *Computer-assisted Analysis of Mixtures and Applications : Meta-analysis, Disease Mapping, and Others*. Chapman and Hall, Boca Raton (1999)
- BOUGUILA, N., ZIOU, D. I VAILLANCOURT, J. Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and its Application. *IEEE Transactions on Image Processing* **13**:1533–1543 (2004)
- BROWNE, R.P. I MCNICHOLAS, P.D. A Mixture of Generalized Hyperbolic Distributions. *Canadian Journal of Statistics* **43**(2):176–198 (2015)
- CALIF, R., EMILIOLO, R. I SOUBDHAN, T. Classification of Wind Speed Distributions Using a Mixture of Dirichlet Distributions. *Renewable Energy* **36**:3091–3097 (2011)
- CELEUX, G. I GOVAERT, G. Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28**(5):781–793 (1995)
- DAY, N.E. Estimating the Components of a Mixture of Normal Distributions. *Biometrika* **56**(3):463–474 (1969)
- DEMPSTER, A.P., LAIRD, N.M. I RUBIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**:1–38 (1977)
- EGOZCUE, J.J. I PAWLOWSKY-GLAHN, V. Simplicial Geometry for Compositional Data. *Geological Society, London, Special Publications* **264**:145–159 (2006)
- EGOZCUE, J.J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. I BARCELÓ-VIDAL, C. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology* **35**:279–300 (2003)
- EVERITT, B.S. I HAND, D.J. *Mixtures of Discrete Distributions*. Chapman and Hall, London (1981)
- FERRER-ROSSELL, B., COENDERS GALLART, G. I MARTÍNEZ GARCÍA, E. Segmentation by Tourist Expenditure Composition, an Approach with Compositional Data Analysis and Latent Classes. *Tourism analysis* **21**(6):589–602 (2016)

- FRALEY, C. I RAFTERY, A.E. MCLUST: Software for Model-based Cluster Analysis. *Journal of Classification* **16**:297–306 (1999)
- GOLDBERGER, J. I ROWEIS, S. Hierarchical Clustering of a Mixture Model. A Y. Weiss, P.B. Schölkopf i J.C. Platt (editors), *NIPS 2005. Advances in Neural Information Processing Systems 18*, pàgs. 505–512. MIT Press, Vancouver (2005)
- HASSELBLAD, V. Estimation of Parameters for a Mixture of Normal Distributions. *Technometrics* **8**(3):431–444 (1966)
- HASSELBLAD, V. Estimation of Finite Mixtures of Distributions from the Exponential Family. *Journal of the American Statistical Association* **64**(328):1459–1471 (1969)
- HENNIG, C. Methods for Merging Gaussian Mixture Components. *Advances in Data Analysis and Classification* **4**(1):3–34 (2010)
- JOHNSON, N.L., KOTZ, S. I BALAKRISHNAN, K. *Discrete Multivariate Distributions*. John Wiley & Sons, New York (1997)
- LEE, H.J. I CHO, S. Combining Gaussian Mixture Models. A Z. Yang, H. Yin i R. Everson (editors), *Intelligent Data Engineering and Automated Learning – IDEAL 2004 SE - 98, Lecture Notes in Computer Science*, volum 3177, pàgs. 666–671. Springer, Heidelberg (2004)
- LEE, S.X. I MCLACHLAN, G.J. On the Fitting of Mixtures of Multivariate Skew t-distributions via the EM algorithm. *arXiv* (2011): 1109.4706.
- LEE, S.X. I MCLACHLAN, G.J. Model-based Clustering and Classification with Non-Normal Mixture Distributions. *Statistical Methods and Applications* **22**(4):427–454 (2013)
- LEE, S.X. I MCLACHLAN, G.J. Finite Mixtures of Multivariate Skew t-distributions: Some Recent and New Results. *Statistics and Computing* **24**(2):181–202 (2014)
- LI, J. Clustering Based on a Multilayer Mixture Model. *Journal of Computational and Graphical Statistics* **14**(3):547–568 (2005)
- LIN, T.I. Robust Mixture Modeling using Multivariate Skew t Distributions. *Statistics and Computing* **20**(3):343–356 (2010)
- LINDSAY, B.G. *Mixture Models: Theory, Geometry and Applications*, volum 5. Institute of Mathematical Statistics, Hayward (1995)

- LONGFORD, N.T. I BARTOSOVA, J. A Confusion Index for Measuring Separation and Clustering. *Statistical Modelling* **14**(3):229–255 (2014)
- MARITZ, J.L. I LWIN, L. *Empirical Bayes Method*. Chapman and Hall, New York (1989)
- MARTÍN-FERNÁNDEZ, J.A. *Medidas de Diferencia y Clasificación Automática no Paramétrica de Datos Composicionales*. Tesis Doctoral, Universitat Politècnica de Catalunya (2001)
- MATEU-FIGUERAS, G. *Models de Distribució sobre el Simplex*. Tesis Doctoral, Universitat Politècnica de Catalunya (2003)
- MATEU-FIGUERAS, G., PAWLOWSKY-GLAHN, V. I EGOZCUE, J.J. The Normal Distribution in Some Constrained Sample Spaces. *SORT* **37**(1):29–56 (2013).
- MCLACHLAN, G.J. I BASFORD, K. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York (1988)
- MCLACHLAN, G.J. I PEEL, D. *Finite Mixture Models, Willey Series in Probability and Statistics*. John Wiley & Sons, New York (2000)
- MCNICHOLAS, P.D. *Mixture Model-Based Classification*. Taylor & Francis, Boca Raton (2016)
- MELNYKOV, V. On the Distribution of Posterior Probabilities in Finite Mixture Models with Application in Clustering. *Journal of Multivariate Analysis* **122**:175–189 (2013)
- MELNYKOV, V. Merging Mixture Components for Clustering through Pairwise Overlap. *Journal of Computational and Graphical Statistics* **25**(1):66–90 (2016)
- MELNYKOV, V., CHEN, W.C. I MAITRA, R. MixSim : An R Package for Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software* **51**:1–25 (2012)
- PALAREA-ALBALADEJO, J. I MARTÍN-FERNÁNDEZ, J.A. zCompositions - R Packages for Multivariate Imputation of Left-censored Data under a Compositional Approach. *Chemometrics and Intelligent Laboratory Systems* **143**:85–96 (2015)

- PAPAGEORGIU, I., BAXTER, M.J. I CAU, M.A. Model-based Cluster Analysis of Artefact Compositional Data. *Archaeometry* **43**(4):571–588 (2001)
- PASTORE, A. I TONELLATO, S.F. A Merging Algorithm for Gaussian Mixture Components. *SSRN Electronic Journal* **4** (2013)
- PAWLOWSKY-GLAHN, V. I BUCCIANTI, A. *Compositional Data Analysis: Theory and Applications*. John Wiley, Chichester (2011)
- PAWLOWSKY-GLAHN, V., EGOZCUE, J.J. I TOLOSANA-DELGADO, R. Lecture Notes on Compositional Data Analysis (2010)
- PAWLOWSKY-GLAHN, V., EGOZCUE, J.J. I TOLOSANA-DELGADO, R. *Modelling and analysis of compositional data*. Wiley (2015)
- PEARSON, K. Mathematical Contributions to the Theory of Evoluton. On a Form of Spurious Correlation which may Arise when Indices are Used in the Measurement of Organs. *Proceedings of the Royal Society of London* **60**:489–502 (1897)
- RAY, S. I LINDSAY, B.G. The Topography of Multivariate Normal Mixtures. *The Annals of Statistics* **33**(5):2042–2065 (2005)
- TIEDEMAN, D.V. On the Study of Types. A S.B. Sells (editor), *Symposium on Pattern Analysis*, pàgs. 1–14. Air University, U.S.A.F. School of Aviation Medicine, Randolph Field (1955)
- TITTERINGTON, D.M., SMITH, A.F.M. I MAKOV, U.E. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York (1986)
- VIVES-MESTRES, M. *Gràfic de Control  $T^2$  de Hotelling per a Dades Composicionals*. Tesis Doctoral, Universitat de Girona (2014)
- WOLFE, J.H. *Object Cluster Analysis of Social Areas*. Tesis Doctoral, University of California, Berkeley (1963)
- WOLFE, J.H. A Computer Program for the Maximum Likelihood Analysis of Type. Inform tècnic, Naval Personnel Research Activity, San Diego (1965)
- WOLFE, J.H. NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. Inform tècnic, Naval Personnel Research Activity, San Diego (1967)



WOLFE, J.H. Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research* **5**(3):329–350 (1970)

XIA, F., CHEN, F., FUNG, W.K. I LI, H. A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics* **69**(4):1053–1063 (2013)