SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Statistics and Quantitative Methods (DISMEQ)

Ph.D. in Statistics and Mathematical Finance Cycle: XXXI

Curriculum in Statistics

A FAMILY OF FLEXIBLE MIXTURE DISTRIBUTIONS FOR CONSTRAINED DATA

Candidate: Ascari Roberto

Registration number: 735831

Tutor: Professor Andrea Ongaro

Supervisor: Professor Andrea Ongaro

Coordinator: Professor Giorgio Vittadini

ACADEMIC YEAR: 2017/2018

A family of Flexible mixture distributions for constrained data

---

Roberto Ascari

January, 2019
Version: Final Version

University of Milano-Bicocca

Department of Statistics and Quantitative Methods (DISMEQ)

A family of Flexible mixture distributions for constrained data

Roberto Ascari

Supervisor Andrea Ongaro

January, 2019

Roberto Ascari
A family of Flexible mixture distributions for constrained data
January, 2019
Supervisor: Andrea Ongaro

University of Milano-Bicocca

Department of Statistics and Quantitative Methods (DISMEQ)
Piazza dell'Ateneo Nuovo, 1
20126, Milan

Acknowledgement

To my FARMA Family,
the best group of human beings I could have wished to share this journey with.

—

Abstract

Some kind of data are defined on unusual mathematical spaces instead of classical ones as $\mathbb{R}^D$. For instance, compositional data belong to the D-dimensional simplex, defined as:

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^\intercal \in \mathbb{R}^D : x_i > 0, \sum_{i=1}^{D} x_i = 1 \right\}.$$

This means that data $\mathbf{x}$ are positive vectors subject to unit-sum constraint (i.e. proportions). Note that compositional data are frequent in many disciplines (e.g. geology, medicine, economics, psychology, environmetrics, etc.); therefore, their proper treatment is a relevant issue.
The Dirichlet is one of the most known distribution defined on the simplex. Although it has several mathematical properties, in many real applications it does not fit the data well, due to its extreme forms of simplicial independence or stiffness in modelling cluster structure and the covariance matrix. Moreover, the Dirichlet distribution allows for only one finite mode. The purpose of this thesis is to compare some distributions proposed in the literature to overcome these drawbacks. In particular, the main aspects of Additive Logistic-Normal (ALN, proposed by Aitchison [3]) and Flexibile Dirichlet (FD, proposed by Ongaro and Migliorati [63, 72]) distributions are recalled. The FD has a particular finite mixture structure (with Dirichlet components) that allows for multimodality and a more flexibile structure of the covariance matrix. In particular, the covariance between distinct elements of a vector with FD density is negative; this is coherent with the unit-sum constraint imposed by the simplex, however, in some applications, such a covariance may be positive. For this reason, a new generalization of both the Dirichlet and the FD distributions has been proposed: the Extended Flexible Dirichlet [8, 61, 71]. This distribution can be obtained normalizing a particular basis $\mathbf{Y} = (Y_1, \ldots, Y_D)^\intercal$, where:

$$Y_r = W_r + Z_r \cdot U_r, \qquad r = 1, \ldots, D,$$

$W_r \sim$ Gamma(α_r, β) are independent random variables, $U_r \sim$ Gamma(τ_r, β) are independent of each other and independent of each W_r and $\mathbf{Z} = (Z_1, \ldots, Z_D)^\intercal$ is a further independent random variable distributed according to a Multinomial$(1, \mathbf{p})$. Then $\mathbf{X} = \dfrac{\mathbf{Y}}{\sum_{r=1}^{D} Y_r} \sim EFD(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$.

The EFD preserves a finite mixture structure as the FD does, but it exhibits some relevant features when compared to the FD, such as a more flexible cluster structure and a (even strong) positive dependence for some pairs of variables. This work completes some theoretical and computational aspects related to this model. In particular, it is possible to obtain Maximum Likelihood estimates through the EM algorithm [28], a well known procedure to find maximizers that often suffers from dependence on the starting point. For this reason, a simulation study aimed at selecting the best initialization procedure among three different proposal have been set up.

An important and significant part of this thesis regards the proposal of a new extension of the Flexible Dirichlet. Both the FD and the EFD distributions allow for a number $k < D$ of potential modes. Even this new model, called Double Flexible Dirichlet (DFD), has a finite mixture structure, as we may write:

$$DFD(\mathbf{x}; \boldsymbol{\alpha}, \tau, \mathbf{P}) = \sum_{i=1}^{D} \sum_{j=1}^{D} p_{i,j} Dir(\mathbf{x}; \boldsymbol{\alpha} + \tau (\mathbf{e}_i + \mathbf{e}_j)),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^\intercal \in \mathbb{R}_D^+$, $\tau > 0$ and $\mathbf{P}$ is a $D \times D$ symmetric matrix with generic (i, j)-th element equal to $p_{i,j}$ subject to constraints:

$$\begin{cases} p_{i,j} > 0 \\ \sum_{i=1}^{D} \sum_{j=1}^{D} p_{i,j} = 1 \end{cases} .$$

The DFD allows for $k < \dfrac{D(D+1)}{2}$ modes (one for each cluster in the mixture). This model also allows for positive correlation among two distinct elements of the composition, despite the presence of the unit-sum constraint. Lot of theoretical properties have been proved and computational aspects have been handled through the R software. The main drawback of this model is the high number of parameters to estimate. This penalizes the DFD model when one compares it with other models through criteria such as AIC and BIC. Moreover, the DFD assumes that the $\dfrac{D(D+1)}{2}$ cluster are placed in a very rigid scheme.

These models (Dirichlet, ALN, FD, EFD and DFD) have been compared through simulation studies and analizing two datasets: the olive oil data from the R package "pdfcluster" and a dataset regarding the results of the Italian general election held on 4 March 2018. Both the EFD and the DFD have shown interesting features that produce a good fit to real data. The EFD considers a lower number of clusters than

the DFD model, but these clusters can be placed almost everywhere in the simplex. If more clusters than the number of elements of the composition or are present some clusters are located with a configuration that cannot be recovered by the FD and the EFD models, the DFD model can be a good solution. Future proposal will be designed to weak the structure imposed by the Double Flexible Dirichlet.

Furthermore, a new model for multivariate constrained count data have been implemented. It is based on a compound Multinomial distribution. Compound distributions are probability distributions obtained by a two-steps approach. The first step consists in assuming that the parameter of the distribution of a random vector X is not constant but follows a specific distribution itself. Then, this joint distribution is marginalized, integrating over the parametric space. This approach leads to more flexible distributions. The Dirichlet-Multinomial is one of the most known compound distributions for multivariate count data. Let $\mathbf{X}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ and $\boldsymbol{\Pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, then the marginal distribution of $\mathbf{X}$ is the Dirichlet-Multinomial distribution. Because of the severe covariance structure imposed by the Dirichlet prior, covariance among distinct elements of $\mathbf{X}$ assumes only negative values and this could be unrealistic in some particular scenarios. A new distribution for count data, called Extended Flexible Dirichlet-Multinomial (EFDM), can be obtained by compounding the Multinomial model with an EFD prior over the parameters $\boldsymbol{\Pi}$. Thanks to the covariance structure of the EFD, the EFDM allows for positive dependence for some pairs of counts. Here some theoretical properties of the EFD-Multinomial distribution are shown, and a preliminary simulation study is performed to evaluate the behavior of two estimation procedures under several scenarios, including positively correlated counts.

Purpose of the thesis is to construct distributions that overcome the drawbacks of simpler distrubutions (Dirichlet, ALN and Multinomial). The reason why this aspect has been considered is that real data can violate the structure assumed by simpler models for a variety of aspects, such as:

- Distributions could not be able to model the data dependence structure because of a poor parametrization (i.e. covariances proportional to the product of the expectations, as in the Dirichlet and Multinomial cases).

- Although the sum constraint naturally induces a negative dependence, positive correlations could be found in sample correlation matrices.

- Real data can show multimodality.

Clustering aspects are not aims of the thesis and, therefore, they are not going to be considered.

Contents

1 Introduction

> *Thus the battle of the statistical knights who search for the holy grail of a parametric class [...] is obviously not over.*
>
> **— John Aitchison.**

Vectors of proportions arise in a great variety of fields: geostatistics, chemistry, economics, medicine, biology, psychology, sociology, environmetrics, politics, agricolture and many others. Supposing that a whole can be split into D mutually exclusive and exhaustive categories, vectors describing the percentage of the amount of each category on the total are called **compositional data**. Let $\mathbf{y} = (y_1, \dots, y_D)^\intercal \in \mathbb{R}_+^D$ be a vector of D positive elements collected on a statistical unit. If $y^+ = \sum_{r=1}^{D} y_r$, then each $x_r = \frac{y_r}{y^+}$ $(r = 1, \dots, D)$ represents the percentage of characteristic y_r among the whole y^+. The vector $\mathbf{y}$ is called basis, whereas $\mathbf{x} = (x_1, \dots, x_D)^\intercal$ is defined **composition** and lies on the D-part **simplex** which is defined as follows:

Definition 1. *The set* $\mathcal{S}^D = \left\{ \boldsymbol{x} = (x_1, \dots, x_D)^\intercal : x_r > 0, r = 1, \dots, D; \sum_{r=1}^{D} x_r = 1 \right\}$ *is a* $(D-1)$*-dimensional subset of* $\mathbb{R}_+^D$ *and is called* D*-part simplex.*

An equivalent definition is:

Definition 2. *The set* $\mathcal{S}_a^D = \left\{ \boldsymbol{x} = (x_1, \dots, x_{D-1})^\intercal : x_r > 0, r = 1, \dots, D-1; \sum_{r=1}^{D-1} x_r < 1 \right\}$ *is a* $(D-1)$*-dimensional subset of* $\mathbb{R}_+^{D-1}$ *and is called* D*-part simplex.*

The latter definition highlights the true dimensionality of the simplex and its asymmetric form since it excludes the last element of the vector (the so called "fill-up" value $x_D = 1 - x_1 - \cdots - x_{D-1}$). Figures 1.1 and 1.2 show 2-parts and 3-parts simplices (colored in red) using both the symmetric and asymmetric definitions.

Looking at the symmetric version of S^2 and S^3, it is easy to see that the D-part simplex is an object with $(D-1)$ dimensions laying into a D-dimensional space. The following example is aimed at showinbg the structure of a compositional dataset.

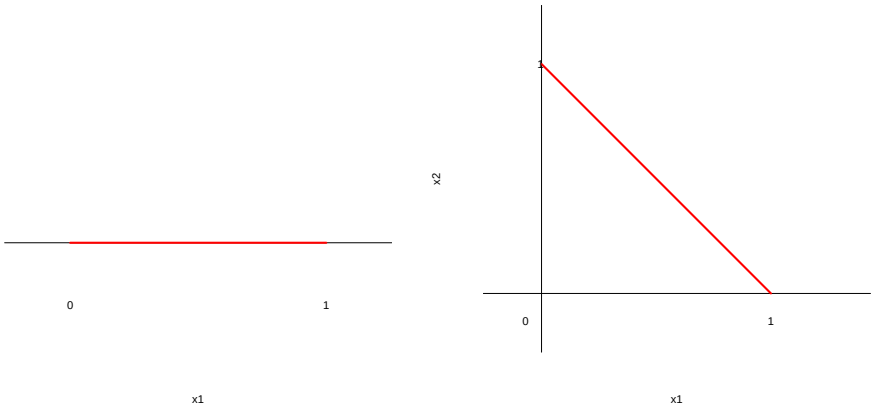

Fig. 1.1: Representation of the 2-parts simplex with the asymmetric definitions (left panel) and the symmetric one (right panel).

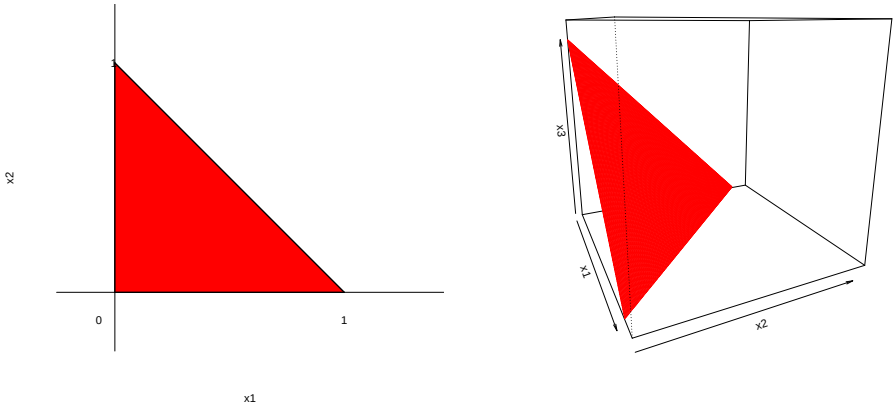

Fig. 1.2: Representation of the 3-parts simplex with the asymmetric definitions (left panel) and the symmetric one (right panel).

Example 1. *Data in Table 1.1 are a subset of a dataset proposed by Aitchison [3] and represent how an academic statistician splits his day into groups of activities over 4 randomly chosen days. Please note that each row satisfies the constraint imposed by the simplex (i.e. each row belongs to* $\mathcal{S}^5$*).*

Day	Teaching	Administration	Research	Other	Sleep	Total
1	0.144	0.179	0.107	0.353	0.217	1
2	0.162	0.107	0.132	0.345	0.254	1
3	0.153	0.131	0.138	0.311	0.267	1
4	0.177	0.140	0.132	0.241	0.310	1

Tab. 1.1: Activity patterns of a statistician during 4 days [3].

Compositional data require a very specific treatment because of their intrinsic dependence structure. Let $\mathbf{Y}$ be a random vector with positive and independent elements

and $Y^+ = \sum_{r=1}^{D} Y_r$; then the covariance among two arbitrary elements of $\mathbf{X} = \mathbf{Y}/Y^+$, say X_r and X_h $(r \neq h)$, is different from 0 in general, although $Y_r \perp\!\!\!\perp Y_h$. In 1897 Karl Pearson [76] pointed out this phenomenon as "spurious correlation" refering to the fact that correlation can possibly be observed where no real association exists. The warning of Pearson has been unheard for many years because of the absence of a coherent statistical methodology. Only in 1986 Aitchison, in the first edition of his monograph [3], proposed a complete and rigorous statistical procedure for analyzing compositional data. The spurious correlation issue can be generalized saying that compositions do not share the dependence structure of their generating basis.

Compositions are not necessarily defined as positive values subject to an unit-sum constraint. This means that a value equal to 0 can appear. Aitchison himself began his monograph with the informal definition:

"*Any vector* $\boldsymbol{x}$ *with non-negative elements* $x_1, \ldots, x_D$ *representing proportions of some whole is subject to the obvious constraint* $x_1 + \cdots + x_D = 1$."[3]

In Chapter 2, he specified that when zeros are present in a dataset, some special adaptations are required. This is the reason why the first formal definition of compositional data and of the simplex do not include any zero values. Standard distributions are defined on this definition of the simplex (i.e. they do not put any mass on the boundary of the simplex).

This work is aimed at presenting some specific distributions defined for compositional data and proposing a new one. Each of these has, of course, advantages and disadvantages so that a "perfect" parametric model for compositional data does not exist. In chapter 2 some useful characteristics of compositional data are provided. Particular emphasis has been put on the definition of subcomposition and amalgamation (i.e. two particular transformations of a composition $\mathbf{x}$), because of their role in future chapters. Chapter 3 introduces some distributions defined for compositional data, listing their advantages and their drawbacks. In particular, the Dirichlet and the Additive Logistic-Normal are likely two of the most known distributions defined for compositional data. The first one is very simple and characterized by several mathematical and statistical properties (for example its log-likelihood function is bounded from above if at least two observations are collected [72] and it is a conjugate prior for the Multinomial model [38, 39]) whereas the second one overtakes the constraint imposed by the support, mapping the simplex into a standard Euclidean space [3–5]. The latter approach allows to use standard statistical methods for normally distributed data, but suffers from the limitation of intrinsic unimodality implied by the multivariate Normal distribution. The Flexible Dirichlet (FD) [63,

72] and the Extended Flexible Dirichlet (EFD) [61, 71], their properties are also presented. This chapter contains two works developed during my Ph.D. period: an alternative Bayesian estimation procedure for the FD (Section 3.3.2, based on [7]) and a simulation study aimed at evaluating the performance of the Maximum Likelihood estimator for the EFD (Section 3.4.4, based on [8]), a distribution introduced by Ongaro and Migliorati [61, 71].

Chapter 4 is the core of this thesis: it introduces a new distribution, the Double Flexible Dirichlet (DFD), whose support is the simplex. The chapter shows the DFD statistical properties and an estimation procedure based on the Expectation-Maximization algorithm. A particular finite mixture structure[38, 58, 60] can be obtained for the DFD model, allowing for a great flexibility in the probability density function (i.e. it can take on several shapes, including multimodality). Thanks to this flexibility, the DFD model gains two peculiarities: it can consider a large number of mixture components and it allows for positive covariances among distinct elements in the composition. These covariances are strictly connected to the covariances of the corresponding elements of the basis. These aspects make the DFD an interesting model to consider when a great number of clusters is expected and/or when they are located in particular ways on the simplex. To evaluate the performance of the Maximum Likelihood estimator for the parameters of a Double Flexible Dirichlet model, a simulation study has been conducted.

Chapter 5 presents two applications to real datasets: the italian general election data (Section 5.1) and the olive oil data (Section 5.2), a compositional dataset presented by Forina in 1983 [36] and used by several authors for model-based clustering [10, 40, 87]. The aim of these applications is to fit models presented in the previous chapters and assess their fit through graphical tools and quantitative criterions.

Finally, in Chapter 6 a way to use distributions for compositional data to extend the well-known Multinomial distribution is discussed. The Multinomial distribution models discrete vectors of integers subject to a sum-constraint. Even if one could treat them as compositions (dividing each element by their sum), their support is not the simplex. Support of the Multinomial distribution is a set of points inside the simplex; therefore this kind of data (referred as "count data" even if they are constrained counts) should be modelled with discrete distributions instead of continuous ones. Particular focus has been put on a new model, the Extended Flexible Dirichlet-Multinomial (EFDM, Section 6.3), obtained compounding the Multinomial distribution with the EFD. Thanks to the dependence structure imposed by the EFD, the EFDM allows for positive covariances among Multinomial counts.

The title refers to a **family** of flexible distributions, but the term "family" is here not used as in the "exponential family" context. Here it defines a set of distributions with

a common "goal" rather than a common general form for the probability density function. Indeed, these Flexible distributions are constucted in order to obtain a more flexible modelization for both the covariance matrix and the (possible) cluster structure of a random vector subject to a sum constraint.

2 Characteristics of Compositional Data

Compositional data are positive vectors subject to an unit-sum constraint. These data can be constructed from a **basis**, that is an unconstrained positive vector $\mathbf{y} = (y_1, \ldots, y_D)^\intercal \in \mathbb{R}_+^D$. A composition is uniquely identified by a basis through the constraining/closure operator $\mathcal{C}(\cdot) : \mathbb{R}_+^D \to \mathcal{S}^D$. The closure operator normalizes its argument $\mathbf{y}$, as follows:

$$\mathbf{x} = \mathcal{C}(\mathbf{y}) \equiv \mathbf{y}/y^+,$$

where $y^+ = \sum_{r=1}^{D} y_r$ is the size of a the basis $\mathbf{y}$. A particular composition $\mathbf{x}$ can be generated by several bases. Indeed, it is possible to define the set $\mathcal{B}(\mathbf{x})$ of all the possible bases leading to the same composition $\mathbf{x}$:

$$\mathcal{B}(\mathbf{x}) = \{\mathbf{y} : \mathbf{y} = \gamma \cdot \mathbf{x}, \quad \gamma \in (0, +\infty)\}.$$

The red line in Figure 2.1 represents $\mathcal{B}(\mathbf{x})$ where $\mathbf{x} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^\intercal$.

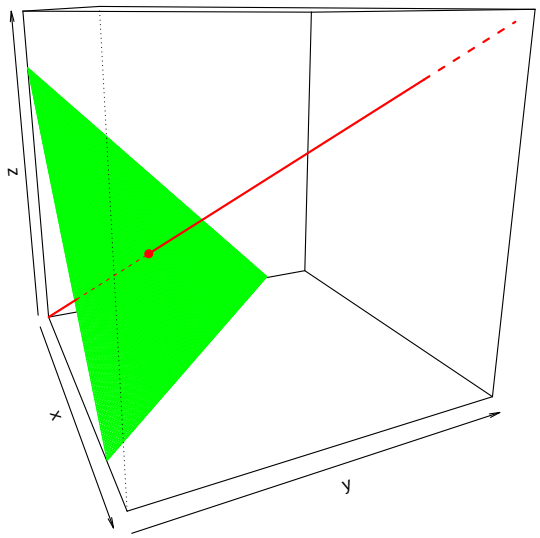

Fig. 2.1: Relationship between the composition $\mathbf{x} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^\intercal$ (dot) and the set $\mathcal{B}(\mathbf{x})$ (line).

A basis is uniquely determined by its size and its composition, since $\mathbf{y} = y^+ \cdot \mathbf{x}$. In what follows, some useful definitions are provided.

Definition 3. *Let $\boldsymbol{y} = (y_1, \ldots, y_D)^\intercal$ be a D-dimensional vector and $\boldsymbol{a} = (a_0, a_1, \ldots, a_{C-1}, a_C)^\intercal$ a vector of non negative integers such that $a_0 = 0 < a_1 < \cdots < a_C = D$. Then the **partition of order C-1** induced by $\boldsymbol{a}$ of the vector $\boldsymbol{y}$ is:*

$$y_1, \ldots y_{a_1} \left| y_{a_1+1}, \ldots, y_{a_2} \right|, \ldots, \left| y_{a_{C-1}+1}, \ldots, y_{a_C} \right.$$

Definition 4. *Let $\boldsymbol{y} = (y_1, \ldots, y_D)^\intercal$ be a D-dimensional vector and $\boldsymbol{a} = (a_0, a_1, \ldots, a_{C-1}, a_C)^\intercal$ a vector of non negative integers such that $a_0 = 0 < a_1 < \cdots < a_C = D$. Then an **amalgamation** based on the partition of order $C-1$ induced by $\boldsymbol{a}$ is the vector of totals of the C subsets:*

$$\boldsymbol{y}^+ = \left(y_1^+, \ldots, y_C^+\right)^\intercal, \qquad \text{where } y_i^+ = \sum_{j=a_{i-1}+1}^{a_i} y_j.$$

Example 2. *Recall data in Example 1. Table 2.1 contains compositions obtained amalgamating the components "Teaching & Administration" and "Sleep & Other":*

Day	Teach. + Adm.	Research	Sleep + Other
1	0.323	0.107	0.570
2	0.269	0.132	0.599
3	0.284	0.138	0.578
4	0.317	0.132	0.551

Tab. 2.1: Activity patterns of a statistician during 4 days (amalgamation)[3].

Definition 5. *Let C and D be two integers such that $0 < C \leq D$. Then, an **amalgamation matrix** $\boldsymbol{A} \in \mathcal{M}(C, D)$ is any matrix with D entries equal to 1 and the remaining equal to 0. The 1s must be allocated one in each column and at least one in each row.*

Example 3. *Each row of Table 2.1 can be obtained by the product of the amalgamation matrix $\boldsymbol{A}$ and the corresponding (transposed) row of Table 1.1, where:*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

Definition 6. *Let $\boldsymbol{b} = (b_1, \ldots, b_{C-1}, b_C)^\intercal$ $(C \leq D)$ be a vector of non negative integers such that $0 < b_1 < \cdots < b_C$. The vector $\boldsymbol{s} = \mathcal{C}((x_{b_1}, \ldots, x_{b_C})^\intercal)$ is called **subcomposition**.*

The subcomposition represents the composition of a subset of components. As a result, it represents a possible way to define marginals in the compositional framework.

Proposition 1. *A subcomposition can be obtained in two ways:*

- *closing a subset of components of a composition, as in Definition 6*
- *closing the equivalent elements of the basis:* $\boldsymbol{s} = \mathcal{C}((y_{b_1}, \ldots, y_{b_C})^\intercal)$.

Proof. Let $\mathbf{y} = (y_1, \ldots, y_D)^\intercal$ be a generic basis and $\mathbf{x} = \mathcal{C}(\mathbf{y}) = (x_1, \ldots, x_D)^\intercal$ its composition. Let $\mathbf{b} = (b_1, \ldots, b_{C-1}, b_C)^\intercal$ a vector of non negative integers and $\mathbf{s}$ a subcomposition as in Definition (6). Then:

$$\begin{aligned} \mathbf{s} &= \mathcal{C}((x_{b_1}, \ldots, x_{b_C})^\intercal) = \left(\frac{x_{b_1}}{\sum_{i=1}^{C} x_{b_i}}, \ldots, \frac{x_{b_C}}{\sum_{i=1}^{C} x_{b_i}} \right)^\intercal \\ &= \left(\frac{y_{b_1}/y^+}{\sum_{i=1}^{C} y_{b_i}/y^+}, \ldots, \frac{y_{b_C}/y^+}{\sum_{i=1}^{C} y_{b_i}/y^+} \right)^\intercal = \left(\frac{y_{b_1}}{\sum_{i=1}^{C} y_{b_i}}, \ldots, \frac{y_{b_C}}{\sum_{i=1}^{C} y_{b_i}} \right)^\intercal = \mathcal{C}((y_{b_1}, \ldots, y_{b_C})^\intercal). \end{aligned}$$

□

The selection of the dimensions to keep in the composition can be made through a selecting matrix, as defined by Aitchison ([3]):

Definition 7. *Let C and D be two integers such that $0 < C < D$. Then, a* ***selecting matrix*** $\boldsymbol{S} \in \mathcal{M}(C, D)$ *is any matrix with C entries equal to 1 and the remaining ones equal to 0. There must be exactly a single 1 in each row and at most one in each column.*

Example 4. *Let be $D = 6$ and suppose that the interest is in the subcomposition $\boldsymbol{s} = \mathcal{C}((x_1, x_3, x_5)^\intercal)$. Then, $\boldsymbol{s} = \mathcal{C}(\boldsymbol{S} \cdot \boldsymbol{x})$, where:*

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Example 5. *From Table 2.2, it is possible to find subcompositions referred to the components "Teaching", "Administration" and "Research" of Table 1.1.*

Day	Teaching	Administration	Research
1	0.335	0.416	0.249
2	0.404	0.267	0.329
3	0.363	0.310	0.327
4	0.394	0.312	0.294

Tab. 2.2: Activity patterns of a statistician during 4 days (subcomposition)[3].

Definition 8. *Let $\boldsymbol{z} = (z_1, \ldots, z_D)^\intercal$ be a vector with positive elements and $\boldsymbol{x}$ a composition. It is possible to define the* ***perturbation*** *operator as $\oplus : \mathbb{R}_+^D \times \mathcal{S}^D \to \mathcal{S}^D$ such that:*

$$\boldsymbol{p} = \boldsymbol{z} \oplus \boldsymbol{x} = \mathcal{C}((z_1 x_1, \ldots, z_D x_D)^\intercal).$$

Definition 9. *Let $\boldsymbol{x} = (x_1, \ldots, x_D)^\intercal$ be a composition and c be a real number. Then, the* ***power transformation*** *operator is defined as $\otimes : \mathcal{S}^D \times \mathbb{R} \to \mathcal{S}^D$ such that:*

$$\boldsymbol{h} = c \otimes \boldsymbol{x} = \mathcal{C}\left((x_1^c, \ldots, x_D^c)^\intercal\right).$$

It is worth noting that the simplex allows for a vector space structure, where perturbation and power play role of addition and scalar multiplication, respectively. Indeed, for $\mathbf{x}, \mathbf{z}, \mathbf{p} \in \mathcal{S}^D$ and $c, d \in \mathbb{R}$ the following axioms hold (some of them have been proved by Billheimer et al. [16]):

- Commutativity of perturbation: $\mathbf{x} \oplus \mathbf{z} = \mathbf{z} \oplus \mathbf{x}$.

$$\mathbf{x} \oplus \mathbf{z} = \mathcal{C}((x_1 z_1, \ldots, x_D z_D)^\intercal) = \mathcal{C}((z_1 x_1, \ldots, z_D x_D)^\intercal) = \mathbf{z} \oplus \mathbf{x}.$$

- Associativity of perturbation: $(\mathbf{x} \oplus \mathbf{z}) \oplus \mathbf{p} = \mathbf{x} \oplus (\mathbf{z} \oplus \mathbf{p})$

$$\begin{aligned}
(\mathbf{x} \oplus \mathbf{z}) \oplus \mathbf{p} &= \mathbf{x}' \oplus \mathbf{p} = \mathcal{C}((x_1' p_1, \ldots, x_D' p_D)^\intercal) \\
&= \left(\frac{x_1' p_1}{\sum_{r=1}^D x_r' p_r}, \ldots, \frac{x_D' p_D}{\sum_{r=1}^D x_r' p_r} \right)^\intercal \\
&= \left(\frac{x_1 z_1 p_1}{\sum_{r=1}^D x_r z_r p_r}, \ldots, \frac{x_D z_D p_D}{\sum_{r=1}^D x_r z_r p_r} \right)^\intercal \\
&= \left(\frac{x_1 z_1 p_1}{\sum_{r=1}^D x_r z_r p_r} \cdot \frac{\sum_{h=1}^D z_h p_h}{\sum_{h=1}^D z_h p_h}, \ldots, \frac{x_D z_D p_D}{\sum_{r=1}^D x_r z_r p_r} \cdot \frac{\sum_{h=1}^D z_h p_h}{\sum_{h=1}^D z_h p_h} \right)^\intercal \\
&= \left(\frac{x_1 p_1'}{\sum_{r=1}^D x_r p_r'}, \ldots, \frac{x_D p_D'}{\sum_{r=1}^D x_r p_r'} \right)^\intercal \\
&= \mathcal{C}((x_1 p_1', \ldots, x_D p_D')^\intercal) = \mathbf{x} \oplus \mathbf{p}' = \mathbf{x} \oplus (\mathbf{z} \oplus \mathbf{p}),
\end{aligned}$$

  where $x_r' = \dfrac{x_r z_r}{\sum_{h=1}^D x_h z_h}$ and $p_r' = \dfrac{z_r p_r}{\sum_{h=1}^D z_h p_h}$.

- Neutral element of perturbation: the D-dimensional vector $\mathbf{u} = \left(\frac{1}{D}, \ldots, \frac{1}{D}\right)^\intercal$ allows for $\mathbf{x} \oplus \mathbf{u} = \mathbf{x}$.

$$\mathbf{x} \oplus \mathbf{u} = \mathcal{C}\left(\left(x_1 \frac{1}{D}, \ldots, x_D \frac{1}{D}\right)^\intercal\right)$$

$$
\begin{aligned}
&= \left(\frac{x_1 \frac{1}{D}}{\sum_{h=1}^{D} x_h \frac{1}{D}}, \ldots, \frac{x_D \frac{1}{D}}{\sum_{h=1}^{D} x_h \frac{1}{D}} \right)^{\intercal} \\
&= \left(\frac{x_1 \frac{1}{D}}{\frac{1}{D} \sum_{h=1}^{D} x_h}, \ldots, \frac{x_D \frac{1}{D}}{\frac{1}{D} \sum_{h=1}^{D} x_h} \right)^{\intercal} = \mathcal{C}((x_1, \ldots, x_D)^{\intercal}) = \mathbf{x}.
\end{aligned}
$$

- Inverse element of perturbation: for each $\mathbf{x} \in \mathcal{S}^D$ it is possible to define the element $\mathbf{x}^{-1} \equiv \mathcal{C}\left(\left(\frac{1}{x_1}, \ldots, \frac{1}{x_D}\right)^{\intercal}\right)$ such that $\mathbf{x}^{-1} \oplus \mathbf{x} = \mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{u}$.

$$
\mathbf{x} \oplus \mathbf{x}^{-1} = \mathcal{C}\left(\left(x_1 \frac{1}{x_1}, \ldots x_D \frac{1}{x_D}\right)^{\intercal}\right) = \mathcal{C}((1, \ldots, 1)^{\intercal}) = \left(\frac{1}{D}, \ldots \frac{1}{D}\right)^{\intercal} = \mathbf{u}.
$$

- Associativity of power transformation: $c \otimes (d \otimes \mathbf{x}) = (c \cdot d) \otimes \mathbf{x}$.

$$
\begin{aligned}
c \otimes (d \otimes \mathbf{x}) &= c \otimes \mathcal{C}\left(\left(x_1^d \ldots x_D^d\right)^{\intercal}\right) \\
&= c \otimes \left(\frac{x_1^d}{\sum_{h=1}^{D} x_h^d}, \ldots, \frac{x_1^d}{\sum_{h=1}^{D} x_h^d} \right)^{\intercal} = c \otimes \left(\frac{x_1^d}{L}, \ldots, \frac{x_1^d}{L} \right)^{\intercal} \\
&= \mathcal{C}\left(\left(\left(\frac{x_1^d}{L} \right)^c, \ldots, \left(\frac{x_D^d}{L} \right)^c \right)^{\intercal} \right) \\
&= \left(\frac{x_1^{c \cdot d}}{L^c} \frac{L^c}{\sum_{r=1}^{D} x_r^{c \cdot d}}, \ldots, \frac{x_D^{c \cdot d}}{L^c} \frac{L^c}{\sum_{r=1}^{D} x_r^{c \cdot d}} \right)^{\intercal} \\
&= \left(\frac{x_1^{c \cdot d}}{\sum_{r=1}^{D} x_r^{c \cdot d}}, \ldots, \frac{x_D^{c \cdot d}}{\sum_{r=1}^{D} x_r^{c \cdot d}} \right)^{\intercal} \\
&= \mathcal{C}\left(\left(x_1^{c \cdot d}, \ldots, x_D^{c \cdot d} \right)^{\intercal} \right) = (c \cdot d) \otimes \mathbf{x}.
\end{aligned}
$$

- Distributivity of power transformation w.r.t. perturbation: $c \otimes (\mathbf{x} \oplus \mathbf{z}) = (c \otimes \mathbf{x}) \oplus (c \otimes \mathbf{z})$

$$
\begin{aligned}
c \otimes (\mathbf{x} \oplus \mathbf{z}) &= c \otimes \left(\frac{x_1 z_1}{\sum_{h=1}^{D} x_h z_h}, \ldots, \frac{x_D z_D}{\sum_{h=1}^{D} x_h z_h} \right)^{\intercal} \\
&= \mathcal{C}\left(\left(\left(\frac{x_1 z_1}{\sum_{h=1}^{D} x_h z_h} \right)^c, \ldots, \left(\frac{x_D z_D}{\sum_{h=1}^{D} x_h z_h} \right)^c \right)^{\intercal} \right) \\
&= \left(\frac{x_1^c z_1^c}{\cancel{\left(\sum_{h=1}^{D} x_h z_h \right)^c}} \cdot \frac{1}{\sum_{r=1}^{D} \left(\frac{x_r z_r}{\cancel{\sum_{h=1}^{D} x_h z_h}} \right)^c}, \ldots, \frac{x_D^c z_D^c}{\cancel{\left(\sum_{h=1}^{D} x_h z_h \right)^c}} \cdot \frac{1}{\sum_{r=1}^{D} \left(\frac{x_r z_r}{\cancel{\sum_{h=1}^{D} x_h z_h}} \right)^c} \right)^{\intercal} \\
&= \left(\frac{x_1^c z_1^c}{\sum_{r=1}^{D} (x_r z_r)^c} \cdot \frac{\left(\sum_{h=1}^{D} x_h^c \right) \left(\sum_{h=1}^{D} z_h^c \right)}{\left(\sum_{h=1}^{D} x_h^c \right) \left(\sum_{h=1}^{D} z_h^c \right)}, \ldots, \frac{x_D^c z_D^c}{\sum_{r=1}^{D} (x_r z_r)^c} \cdot \frac{\left(\sum_{h=1}^{D} x_h^c \right) \left(\sum_{h=1}^{D} z_h^c \right)}{\left(\sum_{h=1}^{D} x_h^c \right) \left(\sum_{h=1}^{D} z_h^c \right)} \right)^{\intercal}
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{x_1^c z_1^c}{\left(\sum_{h=1}^{D} x_h^c\right)\left(\sum_{h=1}^{D} z_h^c\right)} \cdot \frac{\left(\sum_{h=1}^{D} x_h^c\right)\left(\sum_{h=1}^{D} z_h^c\right)}{\sum_{r=1}^{D} (x_r z_r)^c}, \ldots, \frac{x_D^c z_D^c}{\left(\sum_{h=1}^{D} x_h^c\right)\left(\sum_{h=1}^{D} z_h^c\right)} \cdot \frac{\left(\sum_{h=1}^{D} x_h^c\right)\left(\sum_{h=1}^{D} z_h^c\right)}{\sum_{r=1}^{D} (x_r z_r)^c} \right)^\intercal \\
&= \mathcal{C}\left(\left(\frac{x_1^c}{\sum_{h=1}^{D} x_h^c} \cdot \frac{z_1^c}{\sum_{h=1}^{D} z_h^c}, \ldots, \frac{x_D^c}{\sum_{h=1}^{D} x_h^c} \cdot \frac{z_D^c}{\sum_{h=1}^{D} z_h^c} \right)^\intercal \right) \\
&= \left(\frac{x_1^c}{\sum_{h=1}^{D} x_h^c}, \ldots, \frac{x_D^c}{\sum_{h=1}^{D} x_h^c} \right)^\intercal \oplus \left(\frac{z_1^c}{\sum_{h=1}^{D} z_h^c}, \ldots, \frac{z_D^c}{\sum_{h=1}^{D} z_h^c} \right)^\intercal = (c \otimes \mathbf{x}) \oplus (c \otimes \mathbf{z})
\end{aligned}
$$

- Distributivity of power transformation with respect to "classical" addition: $(c+d) \otimes \mathbf{x} = (c \otimes \mathbf{x}) \oplus (d \otimes \mathbf{x})$.

$$
\begin{aligned}
(c+d) \otimes \mathbf{x} &= \mathcal{C}\left(\left(x_1^{c+d}, \ldots, x_D^{c+d} \right)^\intercal \right) \\
&= \left(\frac{x_1^{c+d}}{\sum_{h=1}^{D} x_h^{c+d}} \cdot \frac{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)}{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)}, \ldots, \frac{x_D^{c+d}}{\sum_{h=1}^{D} x_h^{c+d}} \cdot \frac{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)}{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)} \right)^\intercal \\
&= \left(\frac{x_1^{c+d}}{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)} \cdot \frac{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)}{\sum_{h=1}^{D} x_h^{c+d}}, \ldots, \frac{x_D^{c+d}}{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)} \cdot \frac{\left(\sum_{r=1}^{D} x_r^c\right)\left(\sum_{r=1}^{D} x_r^d\right)}{\sum_{h=1}^{D} x_h^{c+d}}, \right)^\intercal \\
&= \mathcal{C}\left(\frac{x_1^c}{\sum_{r=1}^{D} x_r^c} \cdot \frac{x_1^d}{\sum_{r=1}^{D} x_r^d}, \ldots, \frac{x_D^c}{\sum_{r=1}^{D} x_r^c} \cdot \frac{x_D^d}{\sum_{r=1}^{D} x_r^d} \right)^\intercal \\
&= \left(\frac{x_1^c}{\sum_{r=1}^{D} x_r^c}, \ldots, \frac{x_D^c}{\sum_{r=1}^{D} x_r^c} \right)^\intercal \oplus \left(\frac{x_1^d}{\sum_{r=1}^{D} x_r^d}, \ldots, \frac{x_D^d}{\sum_{r=1}^{D} x_r^d} \right)^\intercal \\
&= (c \otimes \mathbf{x}) \oplus (d \otimes \mathbf{x})
\end{aligned}
$$

- Neutral element of power transformation: $1 \otimes \mathbf{x} = \mathbf{x}$.

$$
1 \otimes \mathbf{x} = \mathcal{C}\left(\left(x_1^1, \ldots, x_D^1 \right)^\intercal \right) = \mathcal{C}\left((x_1, \ldots, x_D)^\intercal \right) = \mathbf{x}
$$

The above proofs show that the simplex defines a vector space with basic operations defined by perturbation and power transformation. This vector space can be complemented by a distance measure [2]:

$$d_A(\mathbf{x}, \mathbf{z}) = \left\{ \sum_{k=1}^{D} \left[\log \frac{x_k}{\mu_0(\mathbf{x})} - \log \frac{z_k}{\mu_0(\mathbf{z})} \right]^2 \right\}^{\frac{1}{2}}, \tag{2.1}$$

where $\mathbf{x}, \mathbf{z}$ are vectors on the simplex. Please note that $\mu_0(\mathbf{x}) = \left(\prod_{k=1}^{D} x_k \right)^{\frac{1}{D}}$ is the geometric mean of the components of $\mathbf{x}$. Function (2.1) satisfies the usual conditions for a metric and has three more properties (proofs can be found in [74]):

- $d_A(\cdot, \cdot)$ does not change if the components of the compositions are permuted.
- $d_A(\cdot, \cdot)$ is perturbation invariant: $d_A(\mathbf{x}, \mathbf{z}) = d_A(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{z})$.
- $d_A(\cdot, \cdot)$ is scale invariant: $d_A(c \otimes \mathbf{x}, c \otimes \mathbf{z}) = |c| \cdot d_A(\mathbf{x}, \mathbf{z})$.

Since it is possible to define a metric $d_A(\cdot, \cdot)$ that is perturbation invariant, the simplex is also a metric vector space (usually referred to as the "Aitchison geometry on the simplex" [74]).
By including also the inner product:

$$\langle \mathbf{x}, \mathbf{z} \rangle = \frac{1}{2D} \sum_{i=1}^{S} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{z_i}{z_j}, \tag{2.2}$$

it is possible to show that the D-part simplex is a $(D-1)$-dimensional Hilbert space [74, 75].

The analysis of compositional data raises some issues, the more obvious being "negative bias" problem. Let $\mathbf{X}$ be a random vector whose support is the simplex. Then,

$$0 = \mathrm{Cov}(X_r, 1) = \mathrm{Cov}\left(X_r, \sum_{h=1}^{D} X_h \right) = \sum_{h=1}^{D} \mathrm{Cov}(X_r, X_h), \qquad r = 1, \ldots, D \tag{2.3}$$

and then $\sum_{h \neq r} \mathrm{Cov}(X_r, X_h) = -\mathrm{Var}(X_r)$. Although no assumptions on the distribution of $\mathbf{X}$ have been made, as a consequence of the unit-sum constraint at least one covariance among two distinct components has to be negative. Since equation 2.3 holds for every $r = 1, \ldots, D$, at least one covariance in each row (and column) of the covariance matrix must be negative and the sum of each row (column) must equal 0. Since covariances are not truly free to vary neither are correlations. In

other words, a generic correlation coefficient among components of $\mathbf{X}$ can not vary in the usual range $(-1, 1)$. Then a question raises: is still the zero value a reasonable threshold for no linear association?

Given a covariance matrix (suffering from the negative bias problem) of compositional data is obtained, it could be of interest to investigate the connection with the covariance matrix of the corresponding basis. In general, it is known that the closure operator alters the covariance structure and induces negative correlation.

Example 6. *Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\intercal$ be a random vector defining a basis (i.e. $\boldsymbol{Y} \in \mathbb{R}^3_+$) with indipendent components. Let $Y_r \sim Gamma(\alpha_r, 1)$, $Y^+ = \sum_{r=1}^3 Y_r$ and $\boldsymbol{X} = \boldsymbol{Y}/Y^+$. Then it can be shown that the covariance matrices of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are:*

$$\Sigma_{\boldsymbol{Y}} = \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{bmatrix}, \qquad \Sigma_{\boldsymbol{X}} = \begin{bmatrix} \frac{\alpha_1(\alpha^+ - \alpha_1)}{(\alpha^+)^2(\alpha^+ + 1)} & -\frac{\alpha_1\alpha_2}{(\alpha^+)^2(\alpha^+ + 1)} & -\frac{\alpha_1\alpha_3}{(\alpha^+)^2(\alpha^+ + 1)} \\ -\frac{\alpha_1\alpha_2}{(\alpha^+)^2(\alpha^+ + 1)} & \frac{\alpha_2(\alpha^+ - \alpha_2)}{(\alpha^+)^2(\alpha^+ + 1)} & -\frac{\alpha_2\alpha_3}{(\alpha^+)^2(\alpha^+ + 1)} \\ -\frac{\alpha_1\alpha_3}{(\alpha^+)^2(\alpha^+ + 1)} & -\frac{\alpha_2\alpha_3}{(\alpha^+)^2(\alpha^+ + 1)} & \frac{\alpha_3(\alpha^+ - \alpha_3)}{(\alpha^+)^2(\alpha^+ + 1)} \end{bmatrix}.$$

Details on $\Sigma_{\boldsymbol{X}}$ will be given in Section 3.1.

In the previous example, closing a random vector with independent components generates negative dependence among elements. This is coherent with the observations made by Pearson [76], in 1897. He noted that ratios with independent numerators and the same denominator have a non null correlation. In general, closing a random vector modifies the correlation structure of the vector.

2.1 How to plot compositional data

A problem of compositional data regards visualization. The two main tools to compositional data visualization with $D = 3$ are the Harker diagrams and the ternary diagram. An Harker diagram is simply a scatterplot of two components. Supposing that $D = 3$, then 3 scatterplots of the same compositional dataset (with components called X_1, X_2 and X_3) can be obtained: X_1 Vs. X_2, X_1 Vs. X_3 and X_2 Vs. X_3. Since each of these represents the relationship between different components, one should report all the possible scatterplots of component pairs.

Ternary diagrams [50] are intrinsecally connected to the symmetric version of S^3. They are formed by a triangle representing the boundary of the simplex and points

represent closed compositions. In particular, vertices of this triangle are the 3 compositions $\mathbf{e}_1 = (1,0,0)^\intercal$, $\mathbf{e}_2 = (0,1,0)^\intercal$ and $\mathbf{e}_3 = (0,0,1)^\intercal$ and the edges are the sets of compositions with one null component. Gerald van den Boogaart and Tolosana-Delgado [18] represent in a very clear how a ternary diagram should be read (Figure 2.2 is Figure 2.2 in their book).

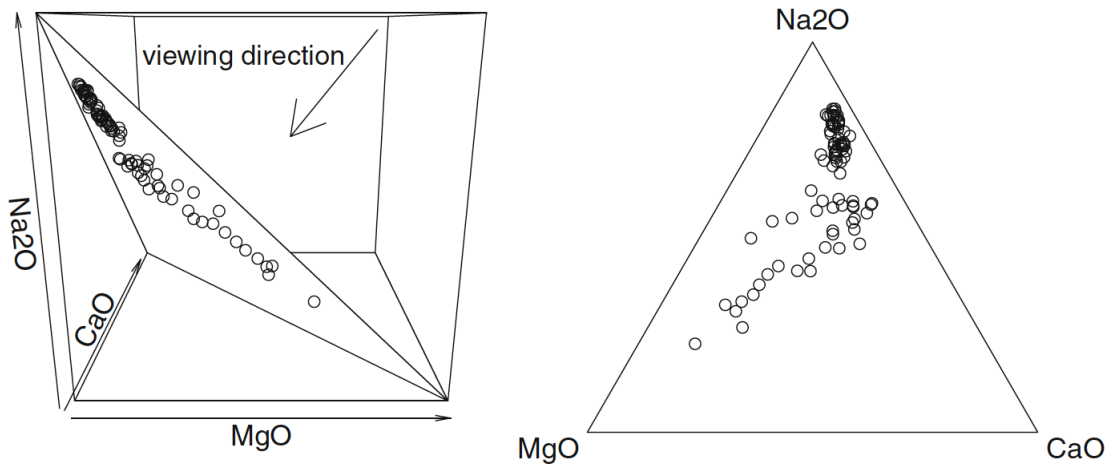

Fig. 2.2: Interpretation of a ternary diagram. Figure taken by Gerald van den Boogaart and Tolosana-Delgado [18].

Ternary diagrams are a very powerful tool, since they help to represent three variables into a two dimensional graphic. Of course, this is possible thanks to the unit-sum constraint imposed by the simplex. In next sections both Harker and ternary diagrams will be used to describe compositional data.

2.2 Simplicial Independencies

Because of the unit-sum constraint and the resulting "negative bias" issue, the usual definition of independence can not be used when dealing with compostional data (i.e. components of a composition always depends on each other). For this reason, Aitchison and other authors [4, 23–27, 52] proposed several forms of independence on the simplex.
Let $\mathbf{X}$ be a random vector obtained closing a basis $\mathbf{Y}$. Several independence properties can be described in terms of subcompositions and amalgamation.

Let $Y_1, \dots Y_{a_1} | Y_{a_1+1}, \dots, Y_{a_2} |, \dots, | Y_{a_{C-1}+1}, \dots, Y_{a_C}$ be a partition of order $C-1$ induced by the vector $\mathbf{a} = (a_0, a_1, \dots, a_C)^\intercal$ as in Definition 3. Then it is possible to define the subcomposition $\mathbf{S}_l$:

$$\mathbf{S}_l = \frac{(X_{1+a_{l-1}}, \dots, X_{a_l})^\intercal}{X_l^+}, \qquad l = 1, \dots, C, \tag{2.4}$$

where $X_l^+ = (X_{1+a_{l-1}} + \cdots + X_{a_l})$ is an element of the corresponding amalgamation vector $\mathbf{X}^+ = (X_1^+, \ldots, X_C^+)^\intercal$.

Definition 10 (Compositional Invariance)**.** *If a D-part composition $\boldsymbol{X} = \mathcal{C}(\boldsymbol{Y})$ is independent of the size of its basis Y^+, then the basis $\boldsymbol{Y}$ is said to be compositionally invariant.*

While compositional invariance is essentially a property of a basis in relation to the corresponding composition, next forms of independence are specifically defined for partitions of order 1. Assuming that $a_1 = k$, it follows that $C = 2$ and therefore $\mathbf{a} = (0, k, D)^\intercal$. If a generic form of independence holds for every $k = 1, \ldots, D-1$, then that independence property is said to be **complete**.

Definition 11 (Subcompositional independence)**.** *The composition $\boldsymbol{X}$ is said to have a subcompositional independence property if the two subcompositions $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are independent: $\boldsymbol{S}_1 \perp\!\!\!\perp \boldsymbol{S}_2$.*

Definition 12 (Subcompositional invariance)**.** *A D-part composition $\boldsymbol{X}$ is subcompositional invariant if $(\boldsymbol{S}_1, \boldsymbol{S}_2)^\intercal \perp\!\!\!\perp \boldsymbol{X}^+$.*

Definition 13 (Neutrality on the left)**.** *A neutrality on the left property means that $\boldsymbol{S}_1 \perp\!\!\!\perp (\boldsymbol{S}_2, \boldsymbol{X}^+)^\intercal$. As a result, the subcomposition $\boldsymbol{S}_1$ is not influenced by $\boldsymbol{X}_2 = \boldsymbol{S}_2 \cdot X_2^+$.*

Definition 14 (Neutrality on the right)**.** *Anagously to Definition 13, a neutrality on the right property means that $\boldsymbol{S}_2 \perp\!\!\!\perp (\boldsymbol{S}_1, \boldsymbol{X}^+)^\intercal$.*

If a composition has both neutrality properties is said to be **neutral**.

Definition 15 (Partition independence)**.** *A D-part composition $\boldsymbol{X}$ has the partition independence property if $\boldsymbol{S}_1 \perp\!\!\!\perp \boldsymbol{S}_2 \perp\!\!\!\perp \boldsymbol{X}^+$.*

3 Models defined on the simplex

In order to apply standard statistical methods (i.e. Maximum Likelihood estimation, computation of confidence intervals, etc.) an assumption on the distribution generating an observed sample is usually required. In this Section some parametric distributions defined on the simplex are proposed. It is important to remark that the D-part simplex is a $(D-1)$-dimensional object laying into a D-dimensional space. This implies that each distribution defined on the simplex must be characterized by a probability density function that is a density with respect to $(D-1)$-dimensional Lebesgue measure [37]. Then, most of the density functions proposed in this section are functions of $D-1$ variables, being the D-th equal to 1 minus the other. Without loss of generality, it is possible to define $x_D = 1 - \sum_{r=1}^{D-1} x_r$ and write these densities as function of D elements.

It is possible to distinguish two approaches in modelling compositional data: the so-called "staying in the simplex" and the transformation approach. The first one defines distribution whose support is the simplex, whereas the second one looks for a suitable transformation in order to map each composition into a different (simplier and unconstrained). The first complete methodology built up for compositional data is based on the latter approach. It has been developed by Aitchison [3] and it has given rise to the "leave the simplex" branch of compositional data analysis.

3.1 The Dirichlet Distribution

The most popular distribution defined on the simplex is the Dirichlet distribution. This is one of the multivariate generalizations of the Beta, that describes data defined in the interval $(0,1)$. Let $\mathbf{X}$ be a D-dimensional random vector distributed according to a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^\intercal$, $\alpha_r > 0$, $r = 1, \ldots, D$. Then, $\mathbf{X} \sim \mathcal{D}(\boldsymbol{\alpha})$ and its probability density function is:

$$f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha^+)}{\prod_{r=1}^{D} \Gamma(\alpha_r)} \prod_{r=1}^{D} x_r^{\alpha_r - 1}, \tag{3.1}$$

where $\mathbf{x} \in \mathcal{S}^D$ and $\alpha^+ = \sum_r \alpha_r$. This distribution is very popular in Bayesian inference, since it can easily represent prior information on probabilities and because the

Dirichlet distribution is a conjugate prior for the Multinomial distribution, making information on the posterior distribution easy to obtain.
Evaluating the form of (3.1) for several configurations of $\boldsymbol{\alpha}$, some interesting considerations can be made. If $\alpha_1 = \cdots = \alpha_D$, then the density is symmetric. This setup can be very useful in Bayesian inference in order to define a non-informative prior on a probability vector. Another configuration of interest is $\alpha_1 = \cdots = \alpha_D = 1$, that corresponds to the uniform distribution on the simplex. If each α_r is greater than 1, then the Dirichlet has an unique (finite) mode. On the contrary, if one or more α_r are less than 1, the density has a peak in each vertex associated to those $\alpha < 1$. Figure 3.1 illustrates different contour plots with different parameter configurations. Comparing top-left and top-right panels of Figure 3.1, it is possible to note that as α^+ increases, it occurs a shift of the probability mass towards a barycenter. Therefore, the parameter α^+ can be thought as a concentration parameter meaning that the greater the α^+ is, the larger probability mass is concentrated around a point in the simplex.
Differently, bottom-left panel shows a situation with three peaks, one in each vertex, due to the fact that all parameters are less than 1. In such a situation, a finite mode does not exist.

The Dirichlet distribution is connected to the Gamma one, since it can be obtained by normalizing a vector of independent Gamma elements. The Gammas elements have a common rate parameter equal to β and can have (possibly) different shape parameter. More formally, let $\mathbf{Y} = (Y_1, \ldots, Y_D)^\intercal$ be the vector of Gamma elements, where $Y_r \perp\!\!\!\perp Y_h$ for every $r \neq h$ and $Y_r \sim \text{Gamma}(\alpha_r, \beta)$. Then, $\mathbf{Y}/Y^+ \sim \mathcal{D}(\boldsymbol{\alpha})$. In the light of this relationship, observations from a Dirichlet distribution can be drawn as follows:

1. Draw y_r from $Y_r \sim \text{Gamma}(\alpha_r, \beta)$, $r = 1, \ldots, D$.

2. Compute $y^+ = \sum_r \alpha_r$.

3. Normalize $\mathbf{y} = (y_1, \ldots, y_D)^\intercal$ to get the vector $\mathbf{x} = \mathcal{C}(\mathbf{y}) = \mathbf{y}/y^+$, which is Diriclet-distributed with parameter $\boldsymbol{\alpha}$.

Example 6 in Section 2 pratically illustrates the relationship among Gamma and Dirichlet distributions. In particular, it enables the comparison between the covariance matrix of the Dirichlet (provided in the next subsection) and the one connected to its Gamma-basis. Since the elements of the Gamma basis are independent, the observed Dirichlet covariances are completely due to the unit-sum constraint and, consequently, to the closure operator.

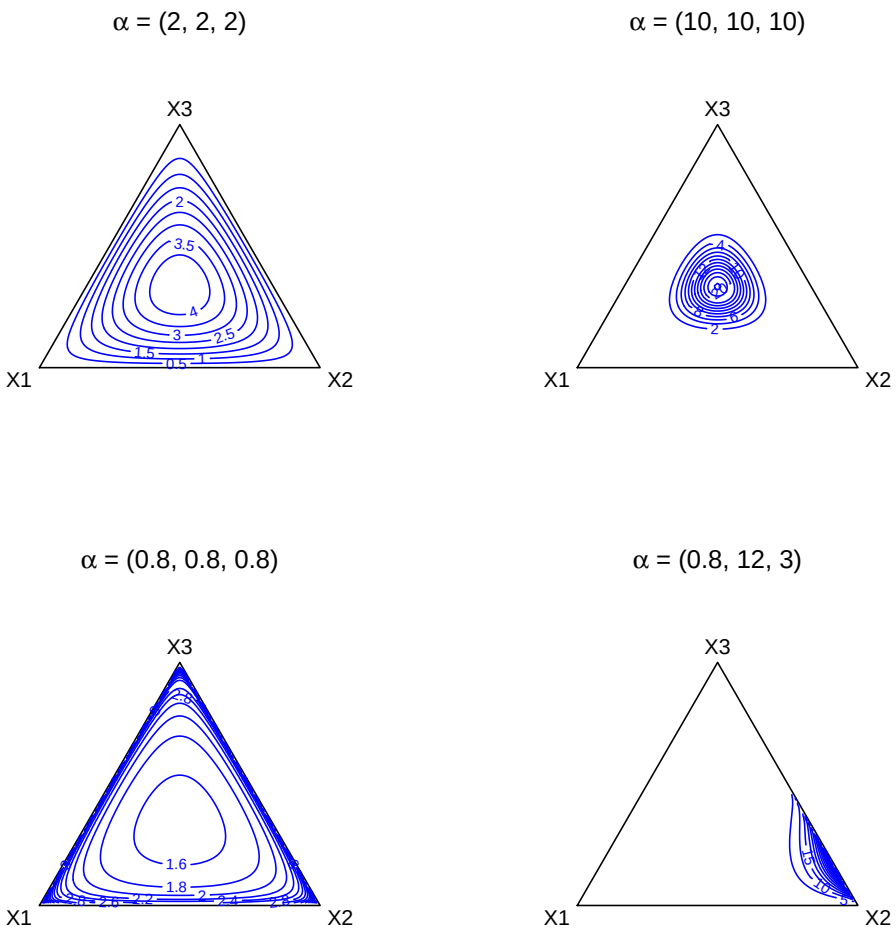

Fig. 3.1: Contour Plots of Dirichlet with different parameters.

Moments and properties

Let $\mathbf{X} = (X_1, \ldots, X_D)^\intercal \sim \mathcal{D}(\boldsymbol{\alpha})$, then the first two order moments can be written in a simple form as follows. Let r and h be two integers such that $r, h = 1, \ldots, D$ and $r \neq h$, then:

$$\mathbb{E}[X_r] = \frac{\alpha_r}{\alpha^+} \tag{3.2}$$

$$\text{Var}(X_r) = \frac{\mathbb{E}[X_r](1 - \mathbb{E}[X_r])}{\alpha^+ + 1} \tag{3.3}$$

$$\text{Cov}(X_r, X_h) = -\frac{\mathbb{E}[X_r]\mathbb{E}[X_h]}{\alpha^+ + 1} \tag{3.4}$$

From the above moments, some remarks can be made:

- The mean vector $\boldsymbol{\mu} = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_D])^\intercal$ is proportional to the parameter $\boldsymbol{\alpha}$.

- If the mean vector is kept fixed, variances and covariances are entirely defined as functions of α^+ only.

- If X_r and X_h have the same expected value, then they are forced to have the same variance too.

- Covariances among distinct elements of $\mathbf{X}$ are always negative and proportional to the product of their expectations.

From these remarks it is immediate to note that the covariance structure of the Dirichlet distribution is quite rigid. There are many real applications where such a rigid structure is not reasonable. For example, considering the unit-sum constraint imposed by the simplex, the negative linear dependence in (3.4) usually makes sense, but it would be just as plausisible to allow two components to be positively associated. As a simple example, suppose a set of data concerning the distribution of a family's income divided into four categories: "Food", "Clothes", "Savings" and "Other". If "Food" and "Clothes" expenditures depend on the number of family members, a positive (spurious) correlation can be observed among those components.
Moreover, there can be situations where components with the same expected value do not have equal variances, which is not coherent with the Dirichlet distribution.

Despite its rigid structure, the Dirichlet distribution has several properties that make it appealing and that motivate its widespread use in compositional data analysis:

Proposition 2 (Closure under Amalgamation)**.** *Let* $\boldsymbol{X} = (X_1, \ldots, X_D)^\intercal \sim \mathcal{D}(\boldsymbol{\alpha})$*; if* $\boldsymbol{X}^+ = \left(X_1^+, \ldots, X_C^+\right)^\intercal$ *and* $\boldsymbol{\alpha}^+ = \left(\alpha_1^+, \ldots, \alpha_C^+\right)^\intercal$ *are the amalgamations of the vectors* $\boldsymbol{X}$ *and* $\boldsymbol{\alpha}$*, induced by the same partition, then* $\boldsymbol{X}^+ \sim \mathcal{D}(\boldsymbol{\alpha}^+)$.

From Proposition 2 it is possible to derive the distribution of marginals. For examples, the univariate marginals are:

$$(X_r, 1-X_r)^\intercal = \left(X_r, \sum_{i \neq r} X_i\right)^\intercal \sim \mathcal{D}(\alpha_r, \alpha^+ - \alpha_r) \equiv \text{Beta}(\alpha_r, \alpha^+ - \alpha_r), \qquad r = 1, \ldots, D \tag{3.5}$$

Proposition 3 (Closure under Conditioning)**.** *Suppose that* $\boldsymbol{X} = (X_1, \ldots, X_D)^\intercal \sim \mathcal{D}(\boldsymbol{\alpha})$ *and let the vector* $\boldsymbol{X}$ *be split into two subvectors,* $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^\intercal$*, where* $\boldsymbol{X}_1 = (X_1, \ldots, X_k)^\intercal$ *and* $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_D)^\intercal$ *for some* $k \in \{1, \ldots, D\}$*. Then,*

$$\left(\frac{\boldsymbol{X}_1}{1 - x_2^+} \middle| \boldsymbol{X}_2 = \boldsymbol{x}_2 \right) \sim \mathcal{D}(\boldsymbol{\alpha}_1), \tag{3.6}$$

where $\boldsymbol{\alpha}_1 = (\alpha_1, \ldots, \alpha_k)^\intercal$ *and* x_2^+ *is the sum of the elements in* $\boldsymbol{x}_2$*.*

Please note that 3.6 can be written equivalently as:

$$\left(\frac{\mathbf{X}_1}{1 - x_2^+} \middle| \mathbf{X}_2 = \mathbf{x}_2 \right) \equiv \mathbf{S}_1 | \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{D}(\boldsymbol{\alpha}_1). \tag{3.7}$$

Equation (3.7) does not depend on $\mathbf{x}_2$, implying that the Dirichlet distribution has the neutrality on the left property. Since neutrality on the right can be easily obtained with the same considerations, it is possible to conclude that the Dirichlet has the neutrality property.

Proposition 4 (Independences)**.** *If* $\boldsymbol{X}$ *is Dirichlet-distributed, then it has all the independence forms described in subsection 2.2.*

Proposition 4 guarantees that the Dirichlet is tractable from a theoretical (and computational) point of view but, on the other hand, it can be unrealistic in real data applications. Indeed, it follows that the Dirichlet can only model extreme independence among compositions, but a key role of statistics in real life applications is to investigate the relationship among variables. Real cases where compositions have every simplicial form of independence are uncommon.

In conclusion, the Dirichlet is a straightforward distribution on the simplex (its simplicity is one of its strong features) that allows for a clear parameter interpretation; however, it comes with a rigid covariance structure and a strong set of independencies. Furthermore, it allows only for one finite mode, which is a strong limitation where multimodality occurs. Several authors proposed distibutions aimed at generalizing the Dirichlet [12, 23, 34, 67, 68, 71, 72]. In particular, the Liouville family plays an important role. Marshall and Olkin [59] introduced the Lioville distribution, Gupta and Richards [42–46] studied the multivariate Liouville distribution, and several authors studied variants of this distributions and their properties [41, 78, 83, 85].

3.2 The Additive Logistic-Normal

Aitchison (1986) developed a complete methodology for compositional data analysis, which takes advantage of standard results for gaussian data through proper transformation. The main idea here is to mimic the Lognormal approach: apply some transformation to non-normally distributed data and then analyze the new variable with a methodology built-up for gaussian data. In particular, Aitchison developed the Additive-Logistic Normal distribution, ALN, also called simply Logistic Normal [3, 5], based on the additive log-ratio transformation to compositional vectors.

Definition 16. *Given a vector $\boldsymbol{x} \in \mathcal{S}^D$, the **additive log-ratio transformation** (alr) is the application:*

$$y_r = \ln\left(\frac{x_r}{x_h}\right), \qquad r = 1, \ldots, h-1, h+1, \ldots, D, \tag{3.8}$$

where the component h is the baseline category:

Since it is a 1-to-1 tranformation from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$, it is possible to define its inverse transformation, namely the **additive logistic transformation**:

$$\begin{cases} x_r = \frac{\exp(y_r)}{1+\sum_{v\neq h}\exp(y_v)}, & r = 1, \ldots, h-1, h+1, \ldots, D \\ x_h = \frac{1}{1+\sum_{v\neq h}\exp(y_v)} \end{cases} \tag{3.9}$$

It is immediate to note that (3.8) and (3.9) depend on the baseline category. A similar transformation that does not depend on h is the centered log-ration transformation.

Definition 17. *Given a vector $\boldsymbol{x} \in \mathcal{S}^D$, the **centered log-ratio transformation** (clr) is defined by:*

$$w_r = \ln\left(\frac{x_r}{\mu_0(\boldsymbol{x})}\right), \qquad r = 1, \ldots, D. \tag{3.10}$$

where $\mu_0(\boldsymbol{x})$ denotes the geometric mean of the elements of $\boldsymbol{x}$.

Let $\mathbf{X} \in \mathcal{S}^D$ be a random composition, then by applying the alr transformation, it holds:

$$Y_r = \ln\left(\frac{X_r}{X_D}\right), \qquad r = 1, \ldots, D-1. \tag{3.11}$$

The random vector $\mathbf{X}$ is said to follow an Additive-Logistic Normal distribution $\Big($denoted as: $\mathbf{X} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\Big)$ if the vector $\mathbf{Y} = (Y_1, \ldots, Y_{D-1})^\intercal$ follows a $(D -$

$1)-$dimensional Normal distribution $\left(\text{i.e. } \mathbf{Y} \sim \mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$. $\boldsymbol{\mu}$ is a $(D-1)$-dimensional mean vector while $\boldsymbol{\Sigma}$ is a $(D-1) \times (D-1)$ covariance matrix also known as the **log-ratio covariance matrix** [3], since its generic element $\sigma_{i,j}$ is the covariance of two log-ratio elements with common denominator:

$$\sigma_{i,j} = \text{Cov}\left(\ln \frac{X_i}{X_D}, \ln \frac{X_j}{X_D}\right). \tag{3.12}$$

As the Normal distribution can be thought of as the limit distribution of a phenomenon characterized by additive random errors, the ALN results as the limit distribution of a compositional event subject to random perturbations [18]. From the density function of $\mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ it is easy to compute the density function of an ALN distribution:

$$f_{\mathcal{L}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\left(\prod_{r=1}^{D} x_r\right)\sqrt{(2\pi)^{D-1}|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\left(\ln \frac{\mathbf{x}_{(-D)}}{x_D} - \boldsymbol{\mu}\right)^{\intercal} \boldsymbol{\Sigma}^{-1}\left(\ln \frac{\mathbf{x}_{(-D)}}{x_D} - \boldsymbol{\mu}\right)\right\}, \tag{3.13}$$

where $\mathbf{x} \in \mathcal{S}^D$, $|\boldsymbol{\Sigma}|$ is the determinant of the matrix $\boldsymbol{\Sigma}$ and $\mathbf{x}_{(-D)}$ is the vector $\mathbf{x}$ without the element in the D-th position. This density function depends on $\frac{(D+2)(D-1)}{2}$ parameters. In Figure 3.2 we report some contour plots referred to the ALN distribution.

An important feature of the Additive-Logistic Normal distribution is its connection to the Lognormal distribution [3]:

Proposition 5. *Let $\boldsymbol{W}$ be a D-dimensional random vector distributed according to a multivariate Lognormal [48] with parameters $\boldsymbol{\xi}$ and Ω, $\boldsymbol{W} \sim \text{LogNorm}^D(\boldsymbol{\xi}, \Omega)$. Then, $\boldsymbol{X} = \mathcal{C}(\boldsymbol{W}) \sim \mathcal{L}^{D-1}(\boldsymbol{F}\boldsymbol{\xi}, \boldsymbol{F}\Omega\boldsymbol{F}^{\intercal})$, where $\boldsymbol{F} = \left[\boldsymbol{I}_{(D-1)}| - \mathbb{1}_{(D)}\right]$ is a $(D-1) \times D$ matrix, $\boldsymbol{I}_{(D-1)}$ is the identity matrix of order $(D-1)$ and $\mathbb{1}_{(D)}$ is a vector with D elements equal to 1.*

The most attractive aspect of this transformation approach is the possibility of making use of every statistical tool based on multivariate normality to analyze compositional data. For example, a statistician can test if a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\intercal}$ arises from an ALN distribution in a two-step approach:

1) transform $\mathbf{x}$ into $\mathbf{y}$ through the additive log-ratio transformation

2) test the assumption of multivariate normality of $\mathbf{y}$ [33]

Another example is the possibility to apply hypothesis testing to transformed data and mapping the inferential conclusions back to the simplex.

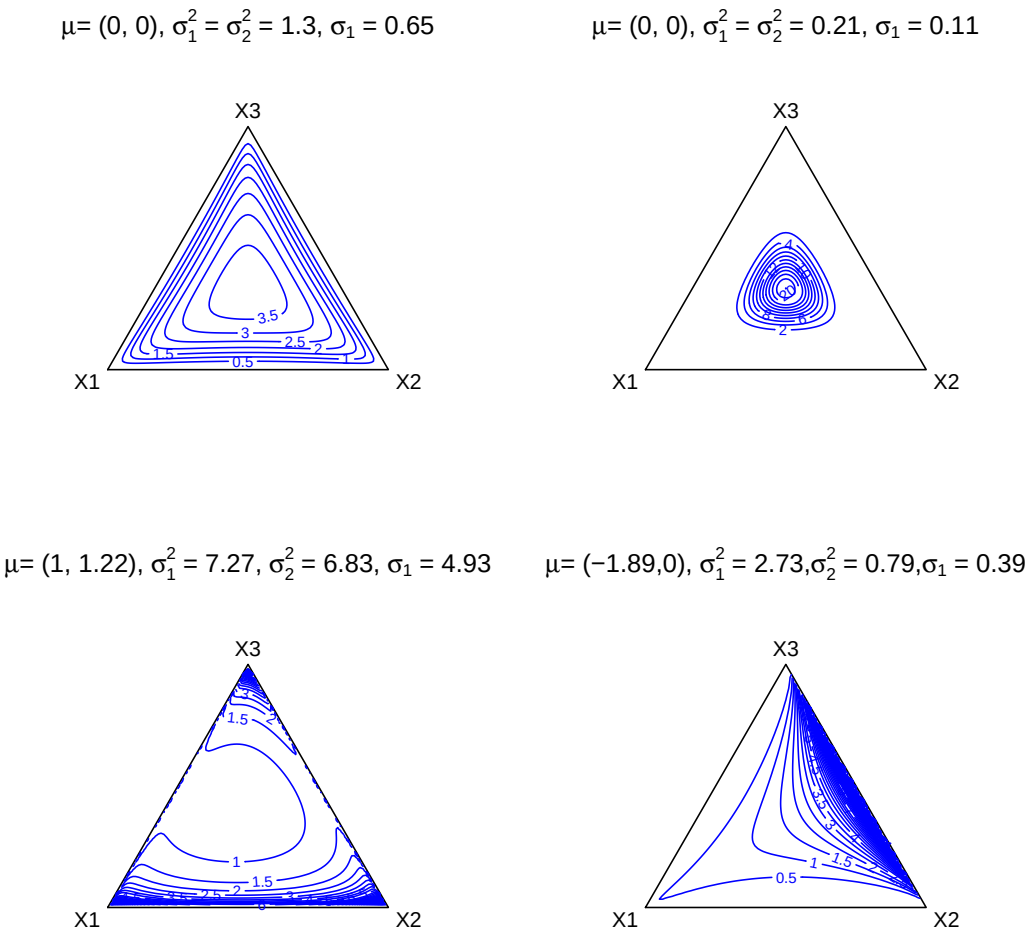

Fig. 3.2: Contour Plots of ALNs with different parameters.

An important limitation of this approach is that the alr transformation depends on a baseline dimension. One might wonder if the order of the components of the composition influence the analysis. In this regard, Aitchison showed that the ALN is closed under permutation of the elements of the compositions [3]:

Proposition 6. *Let* $\boldsymbol{X} \sim \mathcal{L}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{X_P} = \boldsymbol{P} \cdot \boldsymbol{X}$ *the composition ordered according to a permutation matrix* $\boldsymbol{P}$*. Then,* $\boldsymbol{X_P} \sim \mathcal{L}^{D-1}(\mu_{\boldsymbol{P}}, \boldsymbol{\Sigma_P})$*, where:*

$$\boldsymbol{\mu_P} = \boldsymbol{Q_P} \cdot \boldsymbol{\mu}, \qquad \boldsymbol{\Sigma_P} = \boldsymbol{Q_P} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{Q_P^\intercal}, \qquad \boldsymbol{Q_P} = \boldsymbol{F} \cdot \boldsymbol{P} \cdot \boldsymbol{F}^\intercal \cdot \boldsymbol{H}^{-1}.$$

The matrix $\boldsymbol{F}$ *is defined as in Proposition 5, whereas the matrix* $\boldsymbol{H} \in \mathcal{M}(D-1, D-1) = \boldsymbol{I}_{(D-1)} + \boldsymbol{J}_{(D-1)}$ *and* $\boldsymbol{J}_{(D-1)}$ *is a unit matrix.*

Furthermore, other log-ratio models can be considered replacing the alr transformation with a different one. For example, the isometric log-ratio (ilr) transformation [32] allows to define a new distribution. However, since it is possible to show that the ilr transformations are linearly associated with the alr and the Normal distribution is invariant under linear transformations, the ilr model can be expressed as a re-parametrization of the Additive Logistic-Normal distribution.

This "leave the simplex" approach did not satisfy the whole statistical community. Indeed, the transformation approach helps use several tools but it makes the interpretation of the results difficult, especially with respect to the original simplex space. For example, the generic element $\sigma_{i,j}$ in (3.12) is the covariance among the logarithm of two ratios with common denominator: how should it be interpreted? In the discussion of an important paper of Aitchison [4], Fisher made the following statement:

Clearly, the speaker has been very successful in fitting simple models to normal-transformed data; the counterpart to the simplicity of these models is the complexity of corresponding relationships amongst the untransfomed components. [...] I still hold out some hope that simple models of dependence can be found, peculiar to the simplex. [...] Meanwhile, I shall analyse data with the normal-transform method.

In conclusion, the ALN (and all the log-ratio based approach for compositional data analysis) takes advantage of a very simple idea (mapping the simplex into $\mathbb{R}^{D-1}$) and, in such a way, it makes possible to use the standard and powerful statistical methods for inference and modelling based on normality assumption. Nonetheless, this approach introduces some ambiguity. This ambiguity of the ALN distribution does not rely on the particular data transformation, rather than on the interpretation of the results, meaning how expectations and/or covariances of log-ratios should be interpreted with respect to the composition.
Let us consider an example: let $\mathbf{X} = (X_1, X_2, X_3)^\intercal$ be distributed according to an ALN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$), where:

- $\boldsymbol{\mu} = \left(\mathbb{E}\left[\ln \frac{X_1}{X_3}\right], \mathbb{E}\left[\ln \frac{X_2}{X_3}\right]\right)^\intercal$.

- $\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}\left(\ln \frac{X_1}{X_3}\right) & \text{Cov}\left(\ln \frac{X_1}{X_3}, \ln \frac{X_2}{X_3}\right) \\ \text{Cov}\left(\ln \frac{X_2}{X_3}, \ln \frac{X_1}{X_3}\right) & \text{Var}\left(\ln \frac{X_2}{X_3}\right) \end{bmatrix}$.

  $\boldsymbol{\Sigma}$ is the symmetric matrix with generic element $\sigma_{i,j} = \text{Cov}\left(\ln \frac{X_i}{X_3}, \ln \frac{X_j}{X_3}\right)$, $i, j = \{1, 2\}$.

Do the expectations of the log-ratios provide some information on the mean vector of the composition **X**? How is the covariance among two log-ratios connected to the covariance matrix of **X**? A "staying in the simplex" approach allows to avoid this issues.

Furthermore, the ALN distribution does not include the Dirichlet one, althought every Dirichlet distribution can be approximated by the ALN minimizing the Kullback-Leibler divergence measure ([54], see section 3.4.1 for a formal definition) [1]. The alternative approach which will be illustrated in further subsections is built on the definition of a proper distribution on the simplex (that possibly includes the Dirichlet as a particular case), thus avoiding any need for transformation.

3.3 The Flexible Dirichlet

The Dirichlet and the ALN distribution do not allow for multimodality. This characteristic can lower the fit of these models to real data, since they are often clustered, thus showing multimodality. A recent alternative to overcome this issue has been proposed by Ongaro and Migliorati in 2013 [72]. They have developed a new distribution on the simplex with a particular finite mixture structure that allows for a large flexibility both in the density function and in modelling the corresponding covariance matrix. The distribution they developed, referred to as the Flexible Dirichlet distribution, is obtained by normalizing a particular basis **Y**. In order to define the elements of **Y**, the following random variables are introduced:

- $W_r \sim \text{Gamma}(\alpha_r, 1)$, where $W_r \perp\!\!\!\perp W_h$ for $r \neq h$ $(r, h = 1, \dots, D)$
- $U \sim \text{Gamma}(\tau, 1)$, independent of each W_r
- $\mathbf{Z} = (Z_1, \dots, Z_D)^\intercal \sim \text{Multinomial}(1, \mathbf{p})$ independent of the W_r's and the U.

Therefore, $\mathbf{W} = (W_1, \dots, W_D)^\intercal$, U and $\mathbf{Z}$ are jointly independent.

Definition 18. *Let $W_1, \dots, W_D$, U and $\boldsymbol{Z}$ be random variables defined as above. Then, the vector $\boldsymbol{Y} = (Y_1, \dots, Y_D)^\intercal$, where:*

$$Y_r = W_r + U Z_r, \qquad r = 1, \dots, D, \tag{3.14}$$

*follows a **Flexible Gamma** distribution with parameters $\boldsymbol{\alpha}$, τ and $\boldsymbol{p}$, denoted as $FG(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$.*

Due to Definition 18, it is easy to show that $\mathbf{Y}$ is a random vector with positive and dependent elements, whose support is $\mathbb{R}_+^D$. In [72] it is possible to find several properties of this distribution, some are reported below:

- The Flexible Gamma distribution has a finite mixture structure, whose components are vector with independent Gamma elements.

- **Closure under Amalgamation**. Amalgamations of a Flexible Gamma-distributed vector follow a FG distribution themselves. That is, let $\mathbf{Y} \sim \text{FG}(\boldsymbol{\alpha}, \tau, \mathbf{p})$; then $\mathbf{Y}^+ = (Y_1^+, \ldots, Y_C^+)^\intercal \sim \text{FG}(\boldsymbol{\alpha}^+, \tau, \mathbf{p}^+)$, where $\boldsymbol{\alpha}^+ = (\alpha_1^+, \ldots, \alpha_C^+)^\intercal$ and $\mathbf{p}^+ = (p_1^+, \ldots, p_C^+)^\intercal$ are the equivalent amalgamation of the vectors $\boldsymbol{\alpha}$ and $\mathbf{p}$.

- A FG basis is compositionally invariant: $\mathcal{C}(\mathbf{Y}) \perp\!\!\!\perp Y^+$.

It is worth noting that the presence of the Multinomial vector in (3.14) allows for a flexible dependence among elements of the basis. This is a small but significant gain compared with the basis characterizing the Dirichlet distribution, whose elements are independent. In particular, under the Flexible Gamma distribution,

$$\text{Cov}\,(Y_r, Y_h) = -p_r p_h \tau^2, \qquad r \neq h. \tag{3.15}$$

By construction, these covariances are all negative, so that this model does not allow for positive linear dependence among elements of the basis.
Closing the Flexible Gamma basis leads to the distribution defined on the simplex:

Definition 19. *Let $\boldsymbol{Y} \sim FG(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$; then its closed version $\boldsymbol{X} = \mathcal{C}(\boldsymbol{Y})$ is said to be distributed according to a **Flexible Dirichlet** distribution, denoted by $FD(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$.*

Due to the nature of the previously defined random variables, the parametric space of this new simplex distribution is:

$$\Theta_{\text{FD}} = \left\{ (\boldsymbol{\alpha}, \tau, \mathbf{p}) : \boldsymbol{\alpha} \in \mathbb{R}_+^D, \tau \in \mathbb{R}^+, \mathbf{p} \in \mathcal{S}^D \right\}.$$

The crucial characteristic of the Flexible Dirichlet distribution is that, conditioning on $\mathbf{Z}$, it can be represented as a finite mixture model with Dirichlet components. It follows that:

$$f_{FD}(\mathbf{x}; \boldsymbol{\alpha}, \tau, \mathbf{p}) \quad = \quad \sum_{i=1}^{D} p_i f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau\, \mathbf{e}_i)$$

$$= \frac{\Gamma(\alpha^+ + \tau)}{\prod_{r=1}^{D} \Gamma(\alpha_r)} \left(\prod_{r=1}^{D} x_r^{\alpha_r - 1} \right) \sum_{i=1}^{D} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)} x_i^{\tau} \qquad (3.16)$$

for $\mathbf{x} \in \mathcal{S}^D$. The vector $\mathbf{e}_i$ is the i-th element of the usual canonical basis with elements equal to 0 except for the i-th that is equal to 1. It is easy to see that this density coincides with the Dirichlet if and only if $\tau = 1$ and $p_i = \frac{\alpha_i}{\alpha^+}$, $i = 1, \ldots, D$. Thanks to this special case, the FD distribution allows for a severe scheme of independences.

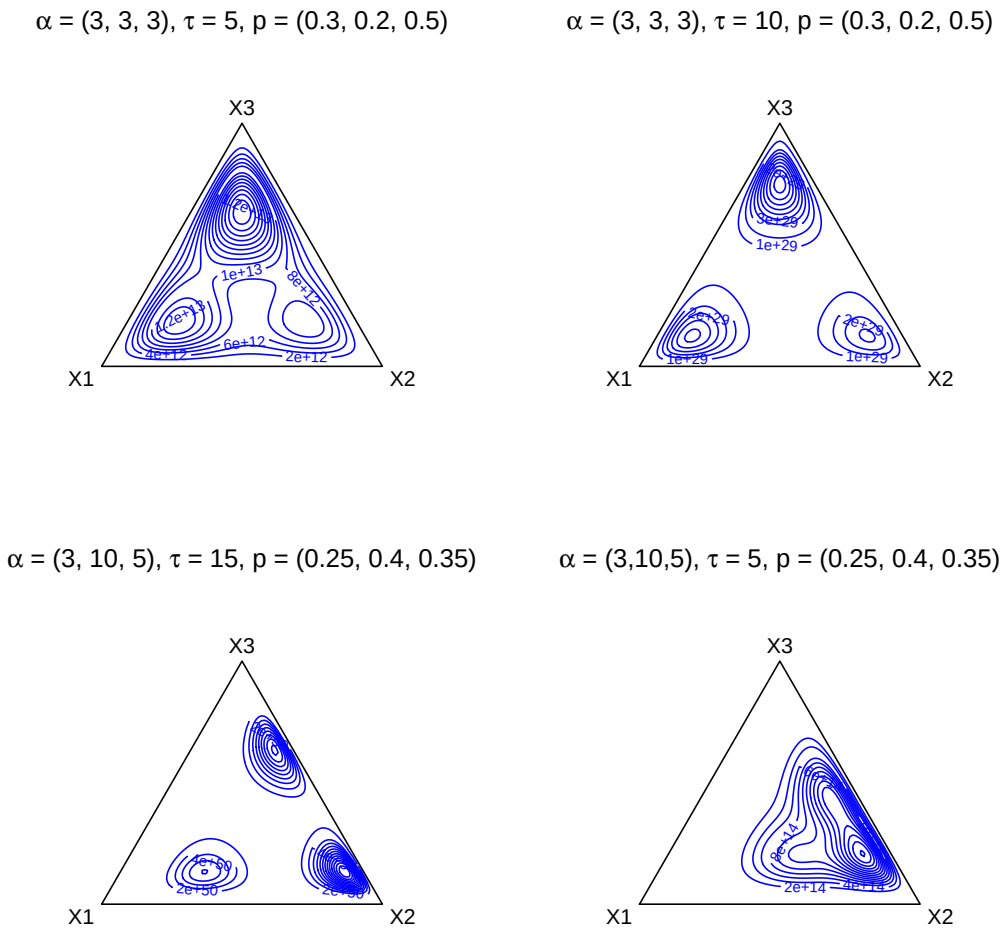

Fig. 3.3: Contour Plots of FD with different parameters.

The density function (3.16), thanks to parameters introduced, can assume several shapes, but, most importantly, it allows for multimodality, as depicted in contour plots in Figure 3.3. This feature is due to the underlying finite mixture structure, that enables to model data collected among several unknown subpopulations [38]. The FD can represent a good model for clustering, since it is a "structured" mixture with

links among the parameters of the components, as each component is parametrized by $\boldsymbol{\alpha} + \tau \mathbf{e}_i$. From the finite mixture theory, it is well known that each mixture component may be interpreted as a cluster; in the FD context, each cluster has the following vector mean:

$$\boldsymbol{\mu}_i^{\text{FD}} = \frac{\boldsymbol{\alpha} + \tau \mathbf{e}_i}{\alpha^+ + \tau} = \left(\frac{\alpha^+}{\alpha^+ + \tau}\right) \frac{\boldsymbol{\alpha}}{\alpha^+} + \left(\frac{\tau}{\alpha^+ + \tau}\right) \mathbf{e}_i, \qquad i = 1, \ldots, D. \tag{3.17}$$

These cluster means deserve a very clear and simple geometric interpretation, as they are linear convex combinations of a common "barycenter" $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$ and the i-th simplex vertex $\mathbf{e}_i$. Thus, the i-th element of $\boldsymbol{\mu}_i^{\text{FD}}$ is higher than the i-th element of $\boldsymbol{\mu}_j^{\text{FD}}$, for every $j \neq i$. The parameter $\frac{\tau}{\alpha^+ + \tau}$ measures the distance between each cluster mean $\boldsymbol{\mu}_i^{\text{FD}}$ and the common barycenter $\bar{\boldsymbol{\alpha}}$ in the direction of $\mathbf{e}_i$.
To better illustrate the cluster structure imposed by this distribution, one can consider the case with $D = 3$. Looking at Figure 3.4 one can see that the i-th cluster mean (blue triangle) is situated on the line connecting $\bar{\boldsymbol{\alpha}}$ (green triangle) to the i-th vertex. Thus, connecting the cluster means one can obtain an equilateral triangle that can be thought of as a re-scaled simplex with the i-th vertex equal to $\boldsymbol{\mu}_i^{FD}$. This "mini-simplex" has edges proportional to the ones of $\mathcal{S}^3$.

Unlike general mixture models, the FD distribution does not have identifiability issues thanks to the particular parameter of each mixture component [63]. Furthermore, thanks to the definition of the Flexible Gamma elements and to its closure under amalgamation, the following propositions hold:

Proposition 7 (Closure under Permutation)**.** *Let $\boldsymbol{X} \sim FD(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$ and let $\tilde{\boldsymbol{X}}$ be any permutation of the elements of $\boldsymbol{X}$. Then, $\tilde{\boldsymbol{X}} \sim FD(\tilde{\alpha}, \tau, \tilde{p})$, where $\tilde{\alpha}$ and $\tilde{p}$ are the corresponding permutation of the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{p}$.*

Proposition 8 (Closure under Amalgamation)**.** *Let $\boldsymbol{X} \sim FD(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$. Then the amalgamation $\boldsymbol{X}^+ = (X_1, \ldots, X_C)^\intercal \sim FD(\boldsymbol{\alpha}^+, \tau, \boldsymbol{p}^+)$.*

Let $\mathbf{X}$ be distributed according to a FD$(\boldsymbol{\alpha}, \tau, \mathbf{p})$ distribution and let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_D)^\intercal$ be a vector of non-negative integers. Then the joint moments of any order are:

$$\mathbb{E}\left[\prod_{r=1}^{D} X_r^{\gamma_r}\right] = \frac{1}{(\alpha^+ + \tau)^{[\gamma^+]}} \prod_{r=1}^{D} \alpha_r^{[\gamma_r]} \sum_{i=1}^{D} \frac{(\alpha_i + \tau)^{[\gamma_i]}}{\alpha_i^{[\gamma_i]}} p_i, \tag{3.18}$$

where $\gamma^+ = \sum_{r=1}^{D} \gamma_r$ and $z^{[\gamma]} = z(z+1)\ldots(z+\gamma-1)$ is the rising factorial (it is important to recall that $z^{[0]} = 1$).
In particular, the first two order moments are:

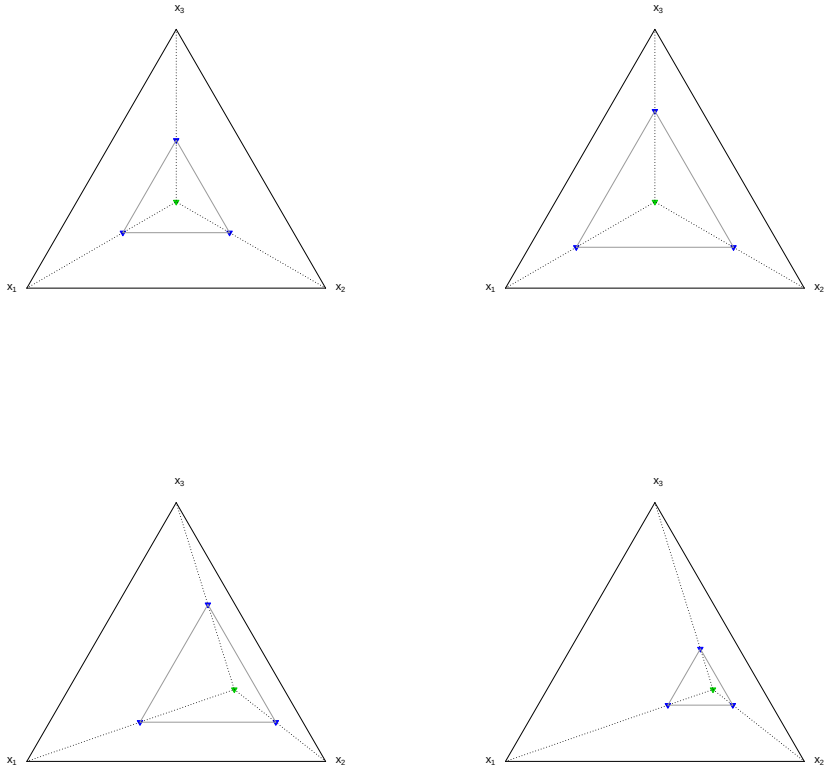

Fig. 3.4: FD cluster means structure. *Top-Left:* $\boldsymbol{\alpha} = (3,3,3)^\intercal$, $\tau = 5$. *Top-Right:* $\boldsymbol{\alpha} = (3,3,3)^\intercal$, $\tau = 10$. *Bottom-Left:* $\boldsymbol{\alpha} = (3,10,5)^\intercal$, $\tau = 15$. *Bottom-Right:* $\boldsymbol{\alpha} = (3,10,5)^\intercal$, $\tau = 5$.

$$\mathbb{E}[X_r] = \frac{\alpha_r + p_r\tau}{\alpha^+ + \tau} = \frac{\alpha_r}{\alpha^+}\left(\frac{\alpha^+}{\alpha^+ + \tau}\right) + p_r\left(\frac{\tau}{\alpha^+ + \tau}\right) \tag{3.19}$$

$$\text{Var}(X_r) = \frac{\mathbb{E}[X_r](1 - \mathbb{E}[X_r])}{\alpha^+ + \tau + 1} + \frac{\tau^2 p_r(1-p_r)}{(\alpha^+ + \tau)(\alpha^+ + \tau + 1)} \tag{3.20}$$

$$\text{Cov}(X_r, X_h) = -\frac{\mathbb{E}[X_r]\,\mathbb{E}[X_h]}{\alpha^+ + \tau + 1} - \frac{\tau^2 p_r p_h}{(\alpha^+ + \tau)(\alpha^+ + \tau + 1)}, \quad r \neq h \tag{3.21}$$

Comparing (3.19 - 3.21) to the corresponding moments of the Dirichlet distribution, it is easy to note that the new parameters entail a more flexible model for the covariance matrix. In particular, the parameters τ and $\mathbf{p}$ allow components with the same mean to have different variances and covariances which are not strictly proportional to the product of expected values. The properties of the Flexible Dirichlet make it a suitable solution for compositional data.

3.3.1 Marginals, Subcompositions and Conditional distributions

Let k be an integer number such that $1 \leq k < D$. Then it is possible to split the vector $\mathbf{X}$ into two subvectors with distinct elements: $\mathbf{X}_1 = (X_1, \ldots, X_k)^\intercal$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_D)^\intercal$. Let X_1^+ and X_2^+ denote the totals of these two subvectors. The quantities $\boldsymbol{\alpha}_l, \mathbf{p}_l, \alpha_l^+$ and p_l^+ $(l = 1, 2)$ are defined in the same way. Finally, $\mathbf{S}_1$ and $\mathbf{S}_2$ are two particular subcompositions: $\mathbf{S}_1 = \mathcal{C}(\mathbf{X}_1)$ and $\mathbf{S}_2 = \mathcal{C}(\mathbf{X}_2)$.

Proposition 9 (Marginal distributions)**.** *Let $\boldsymbol{X}$ be a random vector distributed according to a FD distribution with parameters $\boldsymbol{\alpha}, \tau$ and $\boldsymbol{p}$. Then,*

$$(\boldsymbol{X}_1, 1 - X_1^+)^\intercal \sim FD\left(\left(\boldsymbol{\alpha}_1, \alpha^+ - \alpha_1^+\right)^\intercal, \tau, \left(\boldsymbol{p}_1, 1 - p_1^+\right)^\intercal\right) \tag{3.22}$$

The property of closeness under marginalization is due to the same property of the Flexible Gamma distribution. From (3.24) it is easy to derive the one-dimensional marginals:

$$(X_r, 1 - X_r)^\intercal \sim p_r \text{Beta}\left(\alpha_r + \tau, \alpha^+ - \alpha_r\right) + (1 - p_r)\text{Beta}\left(\alpha_r, \alpha^+ - \alpha_r + \tau\right). \tag{3.23}$$

Proposition 10 (Distribution of subcompositions)**.** *Let $\boldsymbol{X} \sim FD(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$. Then:*

$$\boldsymbol{S}_1 \sim p_1^+ FD\left(\boldsymbol{\alpha}_1, \tau, \frac{\boldsymbol{p}_1}{p_1^+}\right) + (1 - p_1^+)\mathcal{D}(\boldsymbol{\alpha}_1) \tag{3.24}$$

Proposition 11 (Conditional distributions)**.** *Let $\boldsymbol{X} \sim FD(\boldsymbol{\alpha}, \tau, \boldsymbol{p})$. Then:*

$$\boldsymbol{S}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim w(\boldsymbol{x}_2) FD\left(\boldsymbol{\alpha}_1, \tau, \frac{\boldsymbol{p}_1}{p_1^+}\right) + (1 - w(\boldsymbol{x}_2))\mathcal{D}(\boldsymbol{\alpha}_1), \tag{3.25}$$

where $w(\boldsymbol{x}_2) = \dfrac{p_1^+}{p_1^+ + q(\boldsymbol{x}_2)}$ *and* $q(\boldsymbol{x}_2) = \dfrac{\Gamma(\alpha_1^+ + \tau)}{\Gamma(\alpha_1^+)(1 - x_2^+)^\tau} \displaystyle\sum_{i=k+1}^{D} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)} x_i^\tau.$

The above proposition means that the conditional distribution of $\mathbf{S}_1$ given the remaining elements of the composition coincide with a finite mixture of a Dirichlet and a FD distributions. If one of the following holds, then the weights of this finite mixture do not depend on $\mathbf{x}_2$:

- $p_1 = \cdots = p_k = 0$.
- $p_{k+1} = \cdots = p_D = 0$.
- $\tau = 1$ and $\dfrac{p_1}{\alpha_1} = \cdots = \dfrac{p_k}{\alpha_k}$.

Thanks to this special mixture structure, it is easy to derive the conditional moments. For example, the conditional expectation is:

$$\mathbb{E}\left[\mathbf{S}_1|\mathbf{X}_2=\mathbf{x}_2\right]=\frac{\boldsymbol{\alpha}_1}{\alpha_1^+}+\left(\frac{\tau}{\tau+\alpha_1^+}\right)\left(\frac{\mathbf{p}_1}{p_1^+}-\frac{\boldsymbol{\alpha}_1}{\alpha_1^+}\right)w(\mathbf{x}_2). \tag{3.26}$$

It is possible to note that the conditional expectation (3.26) does not depend on $\mathbf{x}_2$ if and only if $p_1^+=0$ or $\boldsymbol{\alpha}_1/\alpha_1^+=\mathbf{p}_1/p_1^+$. Otherwise, it can capture several forms of dependence. For example, let $\mathbf{S}_1=(S_{1,1},S_{1,2},S_{1,3})^\intercal=\dfrac{(X_1,X_2,X_3)^\intercal}{X_1+X_2+X_3}$, then Figure (3.5) shows how the expectation $\mathbb{E}\left[\mathbf{S}_1|X_4=x_4\right]$ varies as a function of x_4, given $\boldsymbol{\alpha}=(3,17,10,5)^\intercal$, $\tau=10$ and $\mathbf{p}=(0.4,0.1,0.25,0.25)^\intercal$. In top panels, it is possible to observe an (increasing or decreasing) S-shape behavior of the conditional expectation, whereas in the bottom panel the regression line is constant. The reason why $\mathbb{E}\left[S_{1,3}|X_4=x_4\right]$ does not change as x_4 varies is that $\dfrac{\alpha_3}{\alpha_1^+}=\dfrac{10}{30}=\dfrac{0.25}{0.75}=\dfrac{p_3}{p_1^+}$, that is one of the condition listed above.

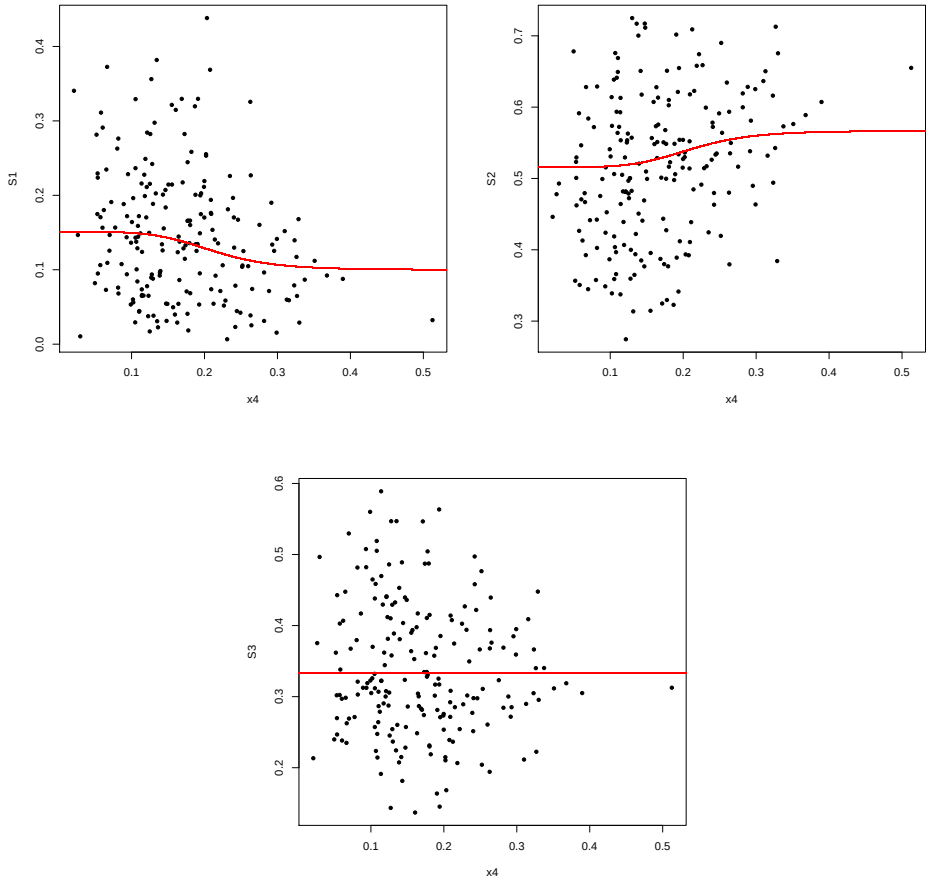

Fig. 3.5: FD's Conditional Expectation with $\boldsymbol{\alpha}=(3,17,10,5)^\intercal$, $\tau=10$ and $\mathbf{p}=(0.4,0.1,0.25,0.25)^\intercal$ - Simulated data. Each color represents a subpopulation.

It is important to note that, thanks to the property of closure under permutation, all the properties previously reported hold not only for subcompositions based on the first k elements of $\mathbf{X}$ but they hold for every subcomposition based on the first k components of any permutation of the elements of $\mathbf{X}$.

Finally, an estimation procedure based on the EM algorithm [28] has been developed to provide Maximum Likelihood Estimates (MLes) for the parameters of a FD distribution [63]. As it is well known that EM algorithm depends on inital values [15, 29, 70], the proposed procedure combines several variants of the EM algorithm and different initialization strategies. The finite mixture structure of the FD justifies the use of the EM algorithm, since it makes possible to treat the estimation procedure as an incomplete data problem. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ be an i.i.d. sample generated from a FD$(\boldsymbol{\alpha}, \tau, \mathbf{p})$. Since the FD$(\cdot)$ can be described by a finite mixture, each observation $\mathbf{x}_s$ $(s = 1, \ldots, n)$ can be thought of as generated from a particular component of that mixture; therefore we may define a latent vector $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\intercal$, where $z_{s,i}$ is equal to 1 if the s-th observation has arisen from the i-th cluster of the mixture (i.e. $S_s = i$) and 0 otherwise. The vector $\mathbf{z}$ can be considered as the missing component vector. Then, it is possible to define the **complete-data log-likelihood** function:

$$L_C(\mathbf{x}, \mathbf{S}; \boldsymbol{\alpha}, \tau, \mathbf{p}) = \prod_{s=1}^{n} \prod_{i=1}^{D} \left\{ p_i \frac{\Gamma(\alpha^+ + \tau)\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)} x_{s,i}^{\tau} \prod_{h=1}^{D} \frac{x_{s,h}^{\alpha_h - 1}}{\Gamma(\alpha_h)} \right\}^{z_{s,i}}. \tag{3.27}$$

This function is the one that will be maximized by the EM algorithm [38]. More details about this estimation procedure can be found in [63].

3.3.2 An alternative estimation procedure: a Bayesian approach

Although there already exists an EM based estimation procedure, a Bayesian approach has been considered and compared to the classical one. The reason why such approach has been considered is that bringing prior information in the analysis (e.g. information on the clusters size $\mathbf{p}$) could be crucial. Furthermore, the interest is also in comparing the results of the two procedures in some challenging scenarios. This work has been presented at the meeting of the CLAssification and Data Analysis Group (CLADAG 2017) [7] and accepted for publication in the CLADAG2017 Springer Book. Bayesian estimation of mixture models often suffers of the label switching problem. Nevertheless, strong identifiability of the FD ensures that the model does not show invariance under permutation of the mixture components. Therefore, no label switching problems arise in the estimation process. Thanks to the already mentioned missing data structure, the Bayesian procedure can rely on a Gibbs sampling algorithm to draw from the posterior distribution [38, 39].

To implement a Bayesian estimation procedure, one needs to define the likelihood function and the priors. The likelihood function suitable for a Gibb sampling algorithm is the complete-data likelihood already defined in (3.27). To simplify prior elicitation, it is useful to assume that $\mathbf{p}$ and $(\boldsymbol{\alpha}, \tau)^\intercal$ have independent prior distributions. In this scenario, a reasonable choice for the prior distribution of $\mathbf{p}$ is $\mathcal{D}(e_0, \dots, e_0)$, $e_0 \in \mathbb{R}^+$. The latter is a common choice in the statistical literaure [30, 38, 39], since the Dirichlet with equal hyperparameters (as the one in the top-left panel in Figure 3.1) is the standard prior for weigths of a finite mixture model and it treats all the components alike ([38]). Another simple choice is to impose independence among τ and each α_i (i.e. $\pi(\boldsymbol{\alpha}, \tau) = \pi(\tau) \prod_{i=1}^{D} \pi(\alpha_i)$) and select a reparametrized exponential prior distribution for each element of the random vector $(\alpha_1, \dots, \alpha_D, \tau)^\intercal$, which greatly simplifies computation of the full conditionals. Thus:

$$\begin{cases} \pi(\alpha_i) \propto a_i^{\alpha_i}, i = 1, \dots, D \\ \pi(\tau) \propto b^\tau \end{cases} \implies \pi(\boldsymbol{\alpha}, \tau) \propto b^\tau \prod_{i=1}^{D} a_i^{\alpha_i}, \tag{3.28}$$

where $(a_1, \dots, a_D, b)^\intercal$ is a vector of hyperparameters and a_i, $b \in (0, 1)$.

Then, the Gibbs sampling algorithm can be described as follows. Let $\mathbf{S}$ denote the vector of missing group labels (i.e. $S_j = i$ means that the j-th observation has arisen from group i). Then, the algorithm is composed by the following steps:

1. Obtain an initial classification $\mathbf{S}^{(0)}$ of data into D groups. Repeat steps 2 and 3 for $m = 1, \dots, B, \dots, B + N$.

2. Given $\mathbf{S}^{(m-1)}$, sample parameters from their full conditionals:

   - Sample $\mathbf{p}^{(m)}$ from $\pi\left(\mathbf{p}|\mathbf{S}^{(m-1)}, \mathbf{x}\right)$
   - Sample $(\boldsymbol{\alpha}^{(m)}, \tau^{(m)})^\intercal$ from $\pi\left(\boldsymbol{\alpha}, \tau|\mathbf{S}^{(m-1)}, \mathbf{x}\right)$

3. Given the new parameters $\left(\boldsymbol{\alpha}^{(m)}, \tau^{(m)}, \mathbf{p}^{(m)}\right)^\intercal$, sample a new partition $\mathbf{S}^{(m)}$ from $\pi\left(\mathbf{S}|\boldsymbol{\alpha}^{(m)}, \tau^{(m)}, \mathbf{p}^{(m)}\right)$

Choosing a Dirichlet prior for $\mathbf{p}$ implies that the full conditional $\pi\left(\mathbf{p}|\mathbf{S}^{(m-1)}, \mathbf{x}\right)$ is a Dirichlet distribution with updated hyperparameters $(e_1, \dots, e_D)^\intercal$, where $e_i = e_0 + N_i\left(\mathbf{S}^{(m-1)}\right)$ and $N_i\left(\mathbf{S}^{(m-1)}\right)$ is the number of data points assigned to group i in partition $\mathbf{S}^{(m-1)}$. Step 3. needs new data partitions $\mathbf{S}^{(m)}$: these can be obtained

by drawing a vector from a Multinomial$(1, \mathbf{p}_s^*)$ and assigning to $S_s^{(m)}$ $(s = 1, \ldots, n)$ the position in which the 1 occurs, where $\mathbf{p}_s^* = (p_{s,1}^*, \ldots, p_{s,D}^*)$ and:

$$p_{s,i}^* = Pr\left(S_s = i | \boldsymbol{\alpha}^{(m)}, \tau^{(m)}, \mathbf{p}^{(m)}\right) = \frac{p_i^{(m)} f_{\mathcal{D}}\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)} + \tau^{(m)} \mathbf{e}_i\right)}{\sum_{k=1}^{D} p_k^{(m)} f_{\mathcal{D}}\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)} + \tau^{(m)} \mathbf{e}_k\right)}, \quad (3.29)$$

$(i = 1, \ldots, D)$.

The main issue with this Gibbs sampling algorithm is the generation of values from the full conditional $\pi(\boldsymbol{\alpha}, \tau | \mathbf{S}^{(m-1)}, \mathbf{x})$. It is possible to show that the latter represents a distribution which results difficult to generate from whatever prior is chosen for $(\boldsymbol{\alpha}, \tau)^\intercal$. Given the joint prior (3.28), the full conditionals are the following:

$$\begin{cases} \pi\left(\alpha_l | \boldsymbol{\alpha}_{(-l)}, \tau, \mathbf{S}, \mathbf{x}\right) \propto \left[\frac{\Gamma(\alpha^+ + \tau)}{\Gamma(\alpha_l)}\right]^n \left[\frac{\Gamma(\alpha_l)}{\Gamma(\alpha_l + \tau)}\right]^{N_l(\mathbf{S})} a_l^{\alpha_l} \prod_{i=1}^{D} \prod_{s:S_s=i} x_{s,l}^{\alpha_l}, \quad l = 1, \ldots, D \\ \pi\left(\tau | \boldsymbol{\alpha}, \mathbf{S}, \mathbf{x}\right) \propto \left[\Gamma(\alpha^+ + \tau)\right]^n \prod_{i=1}^{D} \left[\Gamma(\alpha_i + \tau)\right]^{-N_i(\mathbf{S})} b^\tau \prod_{i=1}^{D} \prod_{s:S_s=i} x_{s,i}^{\tau}. \end{cases}$$

where $\boldsymbol{\alpha}_{(-l)} = \left(\alpha_1, \ldots, \alpha_{(l-1)}, \alpha_{(l+1)}, \ldots, \alpha_D\right)^\intercal$. Unfortunately, these density functions do not characterize any known distribution, so an Inverse Transform Method (ITM, [79]) has been implemented to obtain exact values from these distributions. This method requires the numerical evaluation of $D + 1$ integrals in order to compute the normalization constants for the full conditionals and one more numerical integration to obtain the distribution function of each one of the full conditionals. Finally, the procedure have to numerically find the percentile of order q, where q is drawn from an Uniform distribution on $(0, 1)$. This involves a time-consuming algorithm (i.e. slow convergence of the Gibbs sampler) though, as it has emerged from an exploratory simulation study implemented in R ([77]). This issue can be overcome by considering a new parametrization similar to the one proposed by Migliorati, Di Brisco and Ongaro [62]:

$$\begin{cases} \boldsymbol{\mu} = \frac{\boldsymbol{\alpha}}{\phi} + \tilde{w}\mathbf{p} \\ \phi = \alpha^+ + \tau \end{cases} \quad \begin{cases} \tilde{w} = \frac{\tau}{\phi} \\ \mathbf{p} = \mathbf{p} \end{cases} \quad (3.30)$$

This parametrization allows for an interesting and straightforward interpretation of parameters: the vector $\mathbf{p}$ contains the usual weights of the mixture model, $\boldsymbol{\mu}$ represents the overall mean vector (i.e. $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$), ϕ is a precision parameter and $\tilde{w}$ measures the distance of each cluster mean from the common barycenter $\boldsymbol{\mu}$. It is easy to check that $\boldsymbol{\mu}, \mathbf{p} \in \mathcal{S}^D$ and $\phi \in \mathbb{R}^+$.

Proposition 12. *The new parameter $\tilde{w}$ belongs to the interval* $\left(0, \min\left\{1, \min_j\left\{\frac{\mu_j}{p_j}\right\}\right\}\right)$.

Proof. It is possible to rewrite α_j and α^+ with respect to the new parametrization:

$$\alpha_j = \phi(\mu_j - \tilde{w}p_j), \quad j = 1, \dots, D \qquad\qquad \alpha^+ = \phi(1 - \tilde{w}). \tag{3.31}$$

Since $0 < \frac{\alpha_j}{\alpha^+} < 1$, it follows that $0 < \frac{\phi(\mu_j - \tilde{w}p_j)}{\phi(1-\tilde{w})} < 1$. Then, for every $j = 1, \dots, D$:

$$\begin{aligned} 0 &< \frac{\phi(\mu_j - \tilde{w}p_j)}{\phi(1-\tilde{w})} \\ \Rightarrow 0 &< \mu_j - \tilde{w}p_j \\ \Rightarrow \tilde{w} &< \frac{\mu_j}{p_j} \end{aligned} \tag{3.32}$$

$$\begin{aligned} & \frac{\phi(\mu_j - \tilde{w}p_j)}{\phi(1-\tilde{w})} < 1 \\ \Rightarrow \quad & \mu_j - \tilde{w}p_j < 1 - \tilde{w} \\ \Rightarrow \quad & \tilde{w} < \frac{1-\mu_j}{1-p_j} \end{aligned} \tag{3.33}$$

Since (3.32) holds for every $j = 1, \dots, D$, the following hold:

$$\begin{aligned} & \tilde{w}p_j < \mu_j \\ \Rightarrow \quad & \sum_{h \neq j} p_h \tilde{w} < \sum_{h \neq j} \mu_h \\ \Rightarrow \quad & \tilde{w}(1 - p_j) < (1 - \mu_j) \\ \Rightarrow \quad & \tilde{w} < \frac{1-\mu_j}{1-p_j} \end{aligned}$$

Thus, (3.32) implies (3.33). Since that $\tilde{w} = \frac{\tau}{\phi} < 1$, it follows that $\tilde{w} < \min\left\{1, \min_j\left\{\frac{\mu_j}{p_j}\right\}\right\}$. □

Thanks to Proposition 12, it is possible to define a normalized version of $\tilde{w}$: $w = \frac{\tilde{w}}{\min\left\{1, \min_j\left\{\frac{\mu_j}{p_j}\right\}\right\}}$. In this way the parameter space is variation independent, so that the prior elicitation can rely on independent priors:

$$\begin{cases} \boldsymbol{\mu} \sim \mathcal{D}(e_0, \dots, e_0) \\ w \sim \text{Unif}(0, 1) \end{cases} \qquad \begin{cases} \phi \sim \text{Gamma}(g_1, g_2) \\ \mathbf{p} \sim \mathcal{D}(d_0, \dots, d_0) \end{cases} \tag{3.34}$$

where e_0, d_0, g_1 and g_2 are positive hyperparameters. This set of priors ensures noninformativity or vagueness, in the estimation procedure. Indeed, the Dirichlet distribution with equal hyperparameters treats all the components alike. Moreover, the Gamma distribution is a common choice for the prior of a precision parameter, and, by choosing small values for the rate hyperparameter g_2, vague priors are obtained. By setting $g_1 = g_2$, the prior expectation is equal to 1 and the prior variance is g_2^{-1}. Then a large prior probability is given to observed values all close to zero or one (and this implies that the α_i's in the original parametrization would be less than 1). If some prior information is available and one expects that the precision parameter should be greater than one, then he/she might choose prior distributions for ϕ with higher mean, though still keeping a large variance (i.e. $g_1 = k \cdot g_2$).

A Gibbs sampling algorithm was implemented in the BUGS environment ([57, 69]) to sample from the joint posterior distribution. The likelihood function used in this model is the complete-data likelihood function given by (3.27) according to new parametrization in (3.30).
In order to evaluate the performance of this Gibbs sampling algorithm, samples from a Flexible Dirichlet with $D = 3$ have been simultated considering several parametric configurations. Priors as in (3.34) have been chosen with hyperparameters $e_0 = d_0 = 1$ and $g_1 = g_2 = 0.0001$. The results of two representative parameter configurations are reported below: one characterized by well separated clusters and one with overlapping clusters. The latter is a challenging scenario for every cluster-based approach, due to the difficulties in identifying groups of homogeneous observations. Figure 3.6 shows a simulated dataset for each of these scenarios; data points are colored according to their cluster membership. The left panel is characterized by the parameters $\boldsymbol{\mu} = (0.333, 0.333, 0.333)^\intercal$, $\mathbf{p} = (0.333, 0.333, 0.333)^\intercal$, $\phi = 47$ and $w = 0.362$ whereas the right panel is characterized by $\boldsymbol{\mu} = (0.271, 0.339, 0.390)^\intercal$, $\mathbf{p} = (0.333, 0.333, 0.333)^\intercal$, $\phi = 58.5$ and $w = 0.116$.

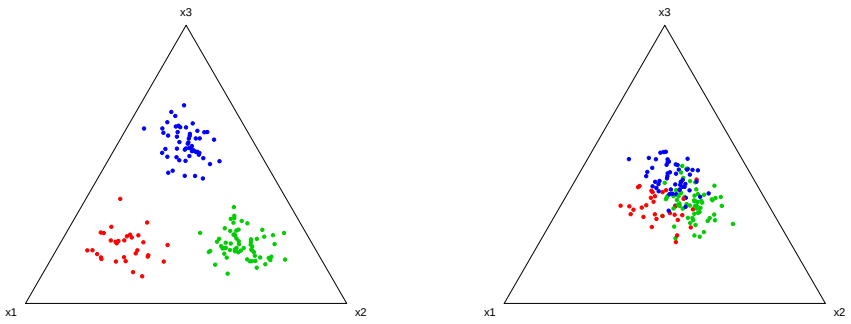

Fig. 3.6: Two datasets simulated from FD with: Each color defines a cluster.

In Appendix 8.1 the interested reader may find the results of the simulation study and all the five configurations of parameters considered.

The simulation consists in generating 200 samples of size 150 for each parameter configuration and, for each of them, initializing an MCMC chain of length 25000 with $B = 10000$ burn-in iteration. To properly treat autocorrelation derived by the use of a MCMC method, a thinning value equal to 10 has been set. Graphical tools (i.e. trace plots and mean plots) have been used to verify the convergence of the chain to the stationary distribution. Tables 3.1 and 3.2 show the mean of the 200 posterior means, the mean of the 200 posterior Standard Deviations (SD), the mean of the Maximum Likelihood estimates (MLe) and the corresponding Standard Errors (SE).

Parameter	True	Post. Mean	Post. SD	MLe	MLe SE
μ_1	0.333	0.334	0.015	0.334	0.015
μ_2	0.333	0.335	0.015	0.335	0.015
μ_3	0.333	0.331	0.015	0.331	0.015
p_1	0.333	0.334	0.038	0.334	0.039
p_2	0.333	0.337	0.039	0.337	0.039
p_3	0.333	0.329	0.038	0.329	0.039
ϕ	47	47.237	3.872	47.824	3.827
w	0.3617	0.361	0.009	0.390	0.018

Tab. 3.1: Simulation results for a well separated clusters scenario.

Parameter	True	Post. Mean	Post. SD	MLe	MLe SE
μ_1	0.271	0.271	0.006	0.271	0.006
μ_2	0.339	0.340	0.006	0.340	0.006
μ_3	0.390	0.389	0.006	0.389	0.006
p_1	0.333	0.366	0.206	0.337	0.155
p_2	0.333	0.335	0.203	0.344	0.162
p_3	0.333	0.299	0.198	0.319	0.171
ϕ	58.5	48.684	8.720	59.332	9.950
w	0.1158	0.066	0.031	0.152	0.050

Tab. 3.2: Simulation results for overlapped clusters.

From Table 3.1 it emerges that, when clusters are well separated, the Bayesian procedure produces more accurate and less variable estimates than the E-M based ones. Nonetheless, if clusters are too closed (Table 3.2), both approaches do not provide unbiased estimation of the parameters, as expected due to the unclear data structure. Though, in this scenario the classical procedure is preferable: the precision parameter ϕ and w are heavily underestimated with the Bayesian approach, while the ML procedure overestimates them only slightly.

One last consideration about the new Bayesian procedure is that it is robust with respect to the choice of the hyperparameters. Even with different values of e_0, d_0, g_1

and g_2, results are similar to the ones reported in Tables 3.1 and 3.2 (in appendix 8.1.2 it is possible to find the results associated to different values of g_1 and g_2). Furthermore, it is also robust with respect to the choice of the loss function: due to the approximate symmetry of each marginal posterior distribution, the posterior means are very close to the posterior medians and posterior modes (see appendix 8.1.1).

In conclusion, the Bayesian approach is very precise when data show well separated clusters, but it does not work as well as the EM algorithm when clusters are overlapping.

3.4 The Extended Flexible Dirichlet

The Flexible Dirichlet can fit real data better than the Dirichlet and the Additive Logistic-Normal distributions in a variety of scenarios. However, it assumes a symmetric structure of the cluster means, as it can be noted by Figure 3.4. This means that the great advantage of using this distribution, with respect the Dirichlet and the ALN ones, depends on a less flexible structure of subpopulation distributions. In order to overcome this aspect, the generating basis can be generalized. Let the generic r-th element of the basis $\mathbf{Y}$ be:

$$Y_r = W_r + U_r Z_r, \qquad r = 1, \dots, D. \tag{3.35}$$

The vectors $\mathbf{W} = (W_1, \dots, W_D)^\intercal$, $\mathbf{U} = (U_1, \dots, U_D)^\intercal$ and $\mathbf{Z} = (Z_1, \dots, Z_D)^\intercal$ are jointly independent. Furthermore, $\mathbf{W}$ and $\mathbf{U}$ have independent elements and $W_r \sim \text{Gamma}(\alpha_r, \beta)$, $U_r \sim \text{Gamma}(\tau_r, \beta)$ and $\mathbf{Z} \sim \text{Multinomial}(1, \mathbf{p})$. This basis is parametrized by the vectors $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\intercal, \boldsymbol{\tau} = (\tau_1, \dots, \tau_D)^\intercal$ and $\mathbf{p} = (p_1, \dots, p_D)^\intercal$ and it can be viewed as a finite mixture of random vectors with independent Gamma components. This allows an easy expression for its density function:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}, \beta) &= \sum_{i=1}^{D} p_i \prod_{r=1}^{D} f_{\mathcal{G}}(y_r; \alpha_r + \tau_r e_{i_r}, \beta) \\ &= \frac{\beta^{\alpha^+}}{\prod_{r=1}^{D} \Gamma(\alpha_r)} e^{-\beta y^+} \left(\prod_{r=1}^{D} y_r^{\alpha_r - 1} \right) \sum_{i=1}^{D} (\beta \cdot y_r)^{\tau_i} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)} p_i. \end{aligned} \tag{3.36}$$

where $f_{\mathcal{G}}(\cdot;\cdot)$ denotes the probability density function of a Gamma random variable. Furthermore,

$$\mathbb{E}\left[\prod_{i=1}^{D} Y_i^{\gamma_i}\right] = \beta^{-\gamma^+}\left(\prod_{r=1}^{D}\alpha_r^{[\gamma_r]}\right)\sum_{i=1}^{D}\frac{(\alpha_i+\tau_i)^{[\gamma_i]}}{\alpha_i^{[\gamma_i]}}p_i, \tag{3.37}$$

where $\gamma^+ = \sum_{r=1}^{D}\gamma_r$, γ_r $(r = 1,\ldots,D)$ are non-negative integers and $z^{[\gamma]} = z(z+1)\ldots(z+\gamma-1)$ is the rising factorial (please note that $z^{[0]} = 1$).

Definition 20. *Let $\boldsymbol{Y} = (Y_1,\ldots,Y_D)^\intercal$ be a basis obtained according to 3.35 and Y^+ its size. Then, the distribution of the random vector $\boldsymbol{X} = \boldsymbol{Y}/Y^+$ is called* ***Extended Flexible Dirichlet*** *and it is denoted by EFD$(\boldsymbol{\alpha},\boldsymbol{\tau},\boldsymbol{p})$.*

Because of the particular mixture structure (3.36), it is possible to show that $\mathbf{X}/Y^+|\mathbf{Z} = \mathbf{e}_i \sim \mathcal{D}(\boldsymbol{\alpha}+\tau_i\mathbf{e}_i)$. Because of the compositional invariance of the Dirichlet distribution, $\mathbf{X}|\mathbf{Z} = \mathbf{e}_i$ is independent of $Y^+|\mathbf{Z} = \mathbf{e}_i$ for every $i = 1,\ldots,D$. Thus it is possible to derive the density function of $(\mathbf{X}, Y^+)^\intercal$ as:

$$f_{(\mathbf{X},Y^+)}(\mathbf{x},y^+;\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p},\beta) = \sum_{i=1}^{D} p_i f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau_i\mathbf{e}_i) f_{\mathcal{G}}(y^+;\alpha^+ + \tau_i,\beta) \tag{3.38}$$

From (3.38) it is easy to compute the marginals of $\mathbf{X}$ and the size Y^+. Specifically:

$$f_{Y^+}(y^+) = \sum_{i=1}^{D} p_i f_{\mathcal{G}}(y^+;\alpha^+ + \tau_i,\beta) \tag{3.39}$$

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}) &= f_{EFD}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}) = \sum_{i=1}^{D} p_i f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau_i\mathbf{e}_i) \\ &= \frac{1}{\prod_{r=1}^{D}\Gamma(\alpha_r)}\left(\prod_{r=1}^{D} x_r^{\alpha_r-1}\right)\sum_{i=1}^{D} p_i \frac{\Gamma(\alpha_i)\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_i+\tau_i)} x_i^{\tau_i} \end{aligned} \tag{3.40}$$

The parametric space of the EFD distribution derives from the definition of the basis in Equation (3.35):

$$\Theta_{\text{EFD}} = \left\{(\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}) : \boldsymbol{\alpha}\in\mathbb{R}_+^D, \boldsymbol{\tau}\in\mathbb{R}_+^D, \mathbf{p}\in\mathcal{S}^D\right\}.$$

From (3.40) it is easy to see that the EFD coincides with the FD distribution if and only if $\tau_1 = \cdots = \tau_D = \tau$. In Figure 3.7 is possible to find some examples of how its density function varies according to the parameter vectors.

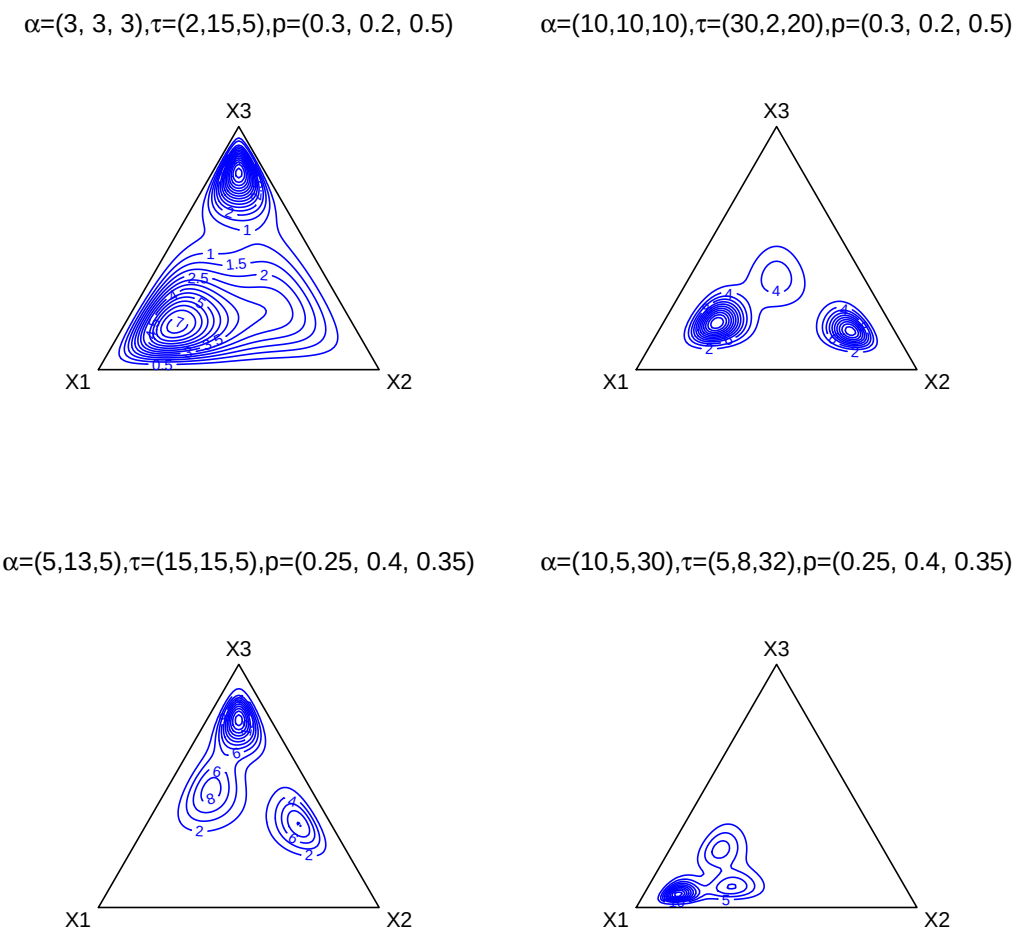

Fig. 3.7: Contour Plots of EFD with different parameters.

The Extendend Flexible Dirichlet has $D-1$ more parameters than the Flexible Dirichlet: these additional parameters allow to introduce more flexibility in the cluster position on the ternary diagram. In particular, each cluster mean can have a different distance from the common barycenter $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$:

$$\boldsymbol{\mu}_i^{EFD} = \frac{\boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i}{\alpha^+ + \tau_i} = \left(\frac{\alpha^+}{\alpha^+ + \tau_i}\right) \frac{\boldsymbol{\alpha}}{\alpha^+} + \left(\frac{\tau_i}{\alpha^+ + \tau_i}\right) \mathbf{e}_i \tag{3.41}$$

From Equation (3.41) it is easy to see that $\boldsymbol{\mu}_i^{EFD}$ is a weighted average of two quantities: $\bar{\boldsymbol{\alpha}}$ and $\mathbf{e}_i$. The higher τ_i is, the further away the i-th cluster is from $\bar{\boldsymbol{\alpha}}$, without depending on $\tau_1, \ldots, \tau_{i-1}, \tau_{i+1}, \ldots, \tau_D$. Thanks to this structure, element i of $\boldsymbol{\mu}_i^{EFD}$ is greater than element i of $\boldsymbol{\mu}_r^{EFD}$ for every $r \neq i$.

Considering the case with $D = 3$, in the FD model the cluster means form an equilateral triangle with edges parallel to the simplex ones. Introducing several τ_i's, such constraint does not hold anymore: connecting the $\boldsymbol{\mu}_i^{EFD}$'s one can obtain any triangle with vertices located on the lines connecting the barycenter $\bar{\boldsymbol{\alpha}}$ to each vertex $\mathbf{e}_i$. Figure 3.8 shows some examples of this cluster pattern.

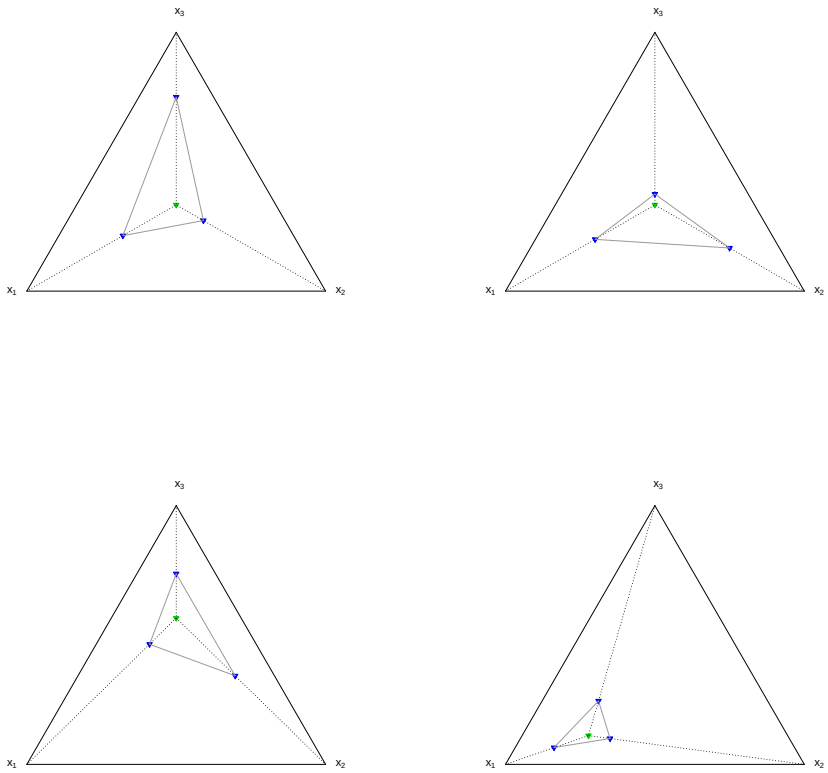

Fig. 3.8: EFD cluster means structure. *Top-Left:* $\boldsymbol{\alpha} = (3, 3, 3)^\intercal$, $\boldsymbol{\tau} = (2, 15, 5)^\intercal$. *Top-Right:* $\boldsymbol{\alpha} = (3, 3, 3)^\intercal$, $\boldsymbol{\tau} = (30, 2, 30)^\intercal$. *Bottom-Left:* $\boldsymbol{\alpha} = (5, 13, 5)^\intercal$, $\tau = (15, 15, 5)^\intercal$. *Bottom-Right:* $\boldsymbol{\alpha} = (10, 5, 30)^\intercal$, $\tau = (5, 8, 32)^\intercal$.

An important feature of the EFD distribution is that it is not compositionally invariant, in general:

Proposition 13. *Let* $\boldsymbol{X} \sim EFD(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$. *Then, it is compositionally invariant if and only if* $\tau_r = \tau \; \forall \; r \in \{1, \ldots, D\}$ *(i.e. if it coincides with the FD).*

Proof. First of all, the conditional distribution of $\mathbf{X}|Y^+ = y^+$ is required. Thanks to equations (3.38) and (3.39) it is possible to compute $f_{\mathbf{X}|Y^+}(\mathbf{x})$:

$$
\begin{aligned}
f_{\mathbf{X}|Y^+}(\mathbf{x}) &= \frac{f_{\mathbf{X},Y^+}(\mathbf{x},y^+)}{f_{Y^+}(y^+)} = \frac{\sum_{i=1}^{D} p_i f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau_i\mathbf{e}_i) f_{\mathcal{G}}(y^+;\alpha^+ + \tau_i,\beta)}{\sum_{r=1}^{D} p_r f_{\mathcal{G}}(y^+;\alpha^+ + \tau_r,\beta)} \\
&= \sum_{i=1}^{D} f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau_i\mathbf{e}_i) \cdot \frac{p_i \frac{\beta^{(\alpha^+ + \tau_i)}}{\Gamma(\alpha^+ + \tau_i)} \left(y^+\right)^{(\alpha^+ + \tau_i - 1)} e^{-(\beta(y^+))}}{\sum_{r=1}^{D} p_r \frac{\beta^{(\alpha^+ + \tau_r)}}{\Gamma(\alpha^+ + \tau_r)} \left(y^+\right)^{(\alpha^+ + \tau_r - 1)} e^{-(\beta(y^+))}} \\
&= \sum_{i=1}^{D} f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau_i\mathbf{e}_i) \cdot \frac{p_i \frac{(y^+)^{\tau_i}}{\Gamma(\alpha^+ + \tau_i)}}{\left(\frac{1}{\beta}\right)^{\tau_i} \sum_{r=1}^{D} p_r \frac{(\beta y^+)^{\tau_r}}{\Gamma(\alpha^+ + \tau_r)}}
\end{aligned}
$$

Defining:

$$
\begin{aligned}
p_i'(y^+) &= \frac{p_i f_{\mathcal{G}}(y^+;\alpha^+ + \tau_i,\beta)}{\sum_{r=1}^{D} p_r f_{\mathcal{G}}(y^+;\alpha^+ + \tau_r,\beta)} \\
&= \frac{p_i \frac{(y^+)^{\tau_i}}{\Gamma(\alpha^+ + \tau_i)}}{\left(\frac{1}{\beta}\right)^{\tau_i} \sum_{r=1}^{D} p_r \frac{(\beta y^+)^{\tau_r}}{\Gamma(\alpha^+ + \tau_r)}},
\end{aligned}
\tag{3.42}
$$

the density function $f_{\mathbf{X}|Y^+}(\mathbf{x})$ defines an EFD$(\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}'(y^+))$ distribution. It is worth to noting that the only quantity influenced by y^+ is the vector $\mathbf{p}'(y^+)$. It is easy to see that if $\tau_i = \tau \ \forall\ i$, then $p_i'(y^+)$ coincides with p_i, $i = 1,\ldots,D$ and then the distribution is compositionally invariant. On the other hand, if the compositional invariance property holds, then $p_i'(y^+)$ does not depend on the size and so does not the ratio $p_i'(y^+)/p_r'(y^+) \propto (y^+)^{\tau_i - \tau_r}$. Then $\tau_i = \tau_r$ for every $i \neq r$. □

Inspecting (3.42) it is possible to study the dependence among $\mathbf{X}|Y^+ = y^+$ and y^+: increasing the size, the weigth associated to the larger τ_i increases. This means that the size does not affect the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ (and, consequently, the cluster means), but it modifies the structure of the weights. In order to illustrate this influence on the weights, the following example is proposed:

Example 7. *Let $\beta = 1$, $\boldsymbol{\alpha} = (5,5,5)^\intercal$ and $\boldsymbol{\tau} = (10,12,8)^\intercal$. Figure 3.9 shows how the vector $\mathbf{p}'(y^+)$ changes as a function of the size y^+ in two different situations: the first with $\boldsymbol{p} = \left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)^\intercal$ and the second with $\boldsymbol{p} = (0.2,0.3,0.5)^\intercal$.*

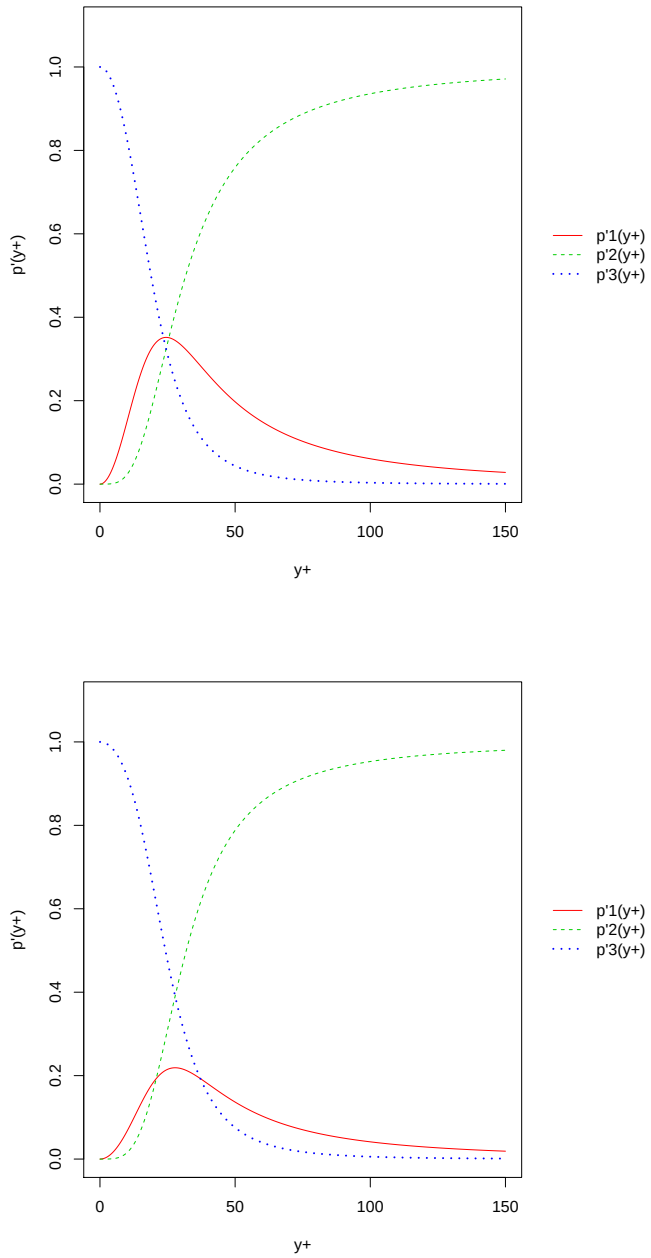

Fig. 3.9: Elements of $\mathbf{p}'(y^+)$ for increasing values of y^+ with $\boldsymbol{\alpha} = (5, 5, 5)^\intercal$, $\boldsymbol{\tau} = (10, 12, 8)^\intercal$ and $\mathbf{p} = (1/3, 1/3, 1/3)^\intercal$ (left) or $\mathbf{p} = (0.2, 0.3, 0.5)^\intercal$ (right).

In both scenarios, low values of y^+ lead the weight associated to the lower τ_i to be close to 1. Increasing the size, this weight goes to zero, whereas the one connected to the highest τ_i approaches 1.

3.4.1 Moments

From (3.40) it is possible to note that the EFD distribution is a finite mixture where the i-th component is distributed according to a $\mathcal{D}(\boldsymbol{\alpha} + \tau_i \mathbf{e}_i)$. This fact makes the joint moments easy to compute:

$$\mathbb{E}\left[\prod_{i=1}^{D} X_i^{\gamma_i}\right] = \prod_{i=1}^{D} \alpha_i^{[\gamma_i]} \sum_{r=1}^{D} \frac{(\alpha_r + \tau_r)^{[\gamma_r]}}{\alpha_r^{[\gamma_r]} (\alpha^+ + \tau_r)^{[\gamma^+]}} p_r. \tag{3.43}$$

In particular, the first two moments of the EFD take the form:

$$\mathbb{E}[X_r] = \alpha_r \, k_1 + \tau_r \frac{p_r}{\alpha^+ + \tau_r}, \tag{3.44}$$

$$\begin{aligned} \text{Var}(X_r) = \alpha_r^2 (k_2 - k_1^2) + \frac{p_r \tau_r (2\alpha_r + \tau_r + 1)}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)} + \\ + \alpha_r k_2 - \frac{p_r^2 \tau_r^2}{(\alpha^+ + \tau_r)^2} - k_1 \frac{2\alpha_r p_r \tau_r}{\alpha^+ + \tau_r}, \end{aligned} \tag{3.45}$$

$$\begin{aligned} \text{Cov}(X_h, X_r) = \frac{\alpha_r p_h \tau_h}{\alpha^+ + \tau_h} \left(\frac{1}{\alpha^+ + \tau_h + 1} - k_1 \right) + \\ + \frac{\alpha_h p_r \tau_r}{\alpha^+ + \tau_r} \left(\frac{1}{\alpha^+ + \tau_r + 1} - k_1 \right) + \\ - \frac{p_h p_r \tau_h \tau_r}{(\alpha^+ + \tau_h)(\alpha^+ + \tau_r)} + \alpha_h \alpha_r (k_2 - k_1^2), \end{aligned} \tag{3.46}$$

$(h, r = 1, \ldots, D;\ h \neq r)$ where:

$$k_1 = \sum_{r=1}^{D} \frac{p_r}{\alpha^+ + \tau_r}, \qquad \text{and} \qquad k_2 = \sum_{r=1}^{D} \frac{p_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)}.$$

The EFD distribution allows for positive covariances. For example, considering the parameters $\boldsymbol{\alpha} = (4, 6, 19)^\intercal$, $\boldsymbol{\tau} = (5, 1, 42)^\intercal$ and $\mathbf{p} = (0.23, 0.12, 0.64)^\intercal$ the covariance and correlation matrices (denoted as $\boldsymbol{\Sigma}$ and $\mathbf{R}$) are:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.00972 & 0.00320 & -0.01180 \\ 0.00320 & 0.00555 & -0.00753 \\ -0.01180 & -0.00753 & 0.02687 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & 0.43547 & -0.72975 \\ 0.43547 & 1 & -0.61621 \\ -0.72975 & -0.61621 & 1 \end{bmatrix}$$

Equation (3.46) is very complicated and its analytical study is particulary heavy. Since the term $\alpha_h \alpha_r (k_2 - k_1^2)$ can assume both positive and negative values, it plays a key role in determining the sign of the covariance. In particular, let T be a discrete

random variable that can assume values $\tau_1, \ldots, \tau_D$ with probabilities $p_1, \ldots, p_D$. Then:

$$\begin{aligned} k_2 - k_1^2 &= \sum_{r=1}^{D} \frac{p_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)} - \left(\sum_{r=1}^{D} \frac{p_r}{\alpha^+ + \tau_r} \right)^2 \pm \sum_{r=1}^{D} \left(\frac{1}{(\alpha^+ + \tau_r)^2} \right) p_r \\ &= \text{Var}\left(\frac{1}{\alpha^+ + T} \right) + \sum_{r=1}^{D} \frac{p_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)} - \sum_{r=1}^{D} \left(\frac{1}{(\alpha^+ + \tau_r)^2} \right) p_r \\ &= \text{Var}\left(\frac{1}{\alpha^+ + T} \right) + \sum_{r=1}^{D} \frac{p_r(\alpha^+ + \tau_r) - p_r(\alpha^+ + \tau_r + 1)}{(\alpha^+ + \tau_r)^2(\alpha^+ + \tau_r + 1)} \\ &= \text{Var}\left(\frac{1}{\alpha^+ + T} \right) - \mathbb{E}\left[\frac{1}{(\alpha^+ + T)^2(\alpha^+ + T + 1)} \right] \end{aligned}$$

This formulation helps to understand that positive values of $k_2 - k_1^2$ can be obtained inducing a large variability of T.

It is worthwhile to recall some notation from the previous sections. The compositional vector $\mathbf{X}$ can be split into two subvectors $\mathbf{X}_1 = (X_1, \ldots, X_k)^\intercal$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_D)^\intercal$ for some integer $k \geq 1$. X_1^+ and X_2^+ are the totals of $\mathbf{X}_1$ and $\mathbf{X}_2$; with the same notation it is possible to define the quantities $\boldsymbol{\alpha}_l, \boldsymbol{\tau}_l, \mathbf{p}_l, \alpha_l^+, \tau_l^+$ and p_l^+ $(l = 1, 2)$. Finally, $\mathbf{S}_1 = \mathcal{C}(\mathbf{X}_1)$ and $\mathbf{S}_2 = \mathcal{C}(\mathbf{X}_2)$ are the subcompositions originated by $\mathbf{X}_1$ and $\mathbf{X}_2$.

Proposition 14 (Marginal distributions)**.** *Let $\boldsymbol{X} \sim EFD(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$, then:*

$$\begin{aligned} \left(\boldsymbol{X}_1, 1 - X_1^+ \right) \sim p_1^+ EFD\left((\boldsymbol{\alpha}_1, \alpha_2)^\intercal, \boldsymbol{\tau}_1, \left(\frac{\boldsymbol{p}_1}{p_1^+}, 0 \right)^\intercal \right) + \\ + (1 - p_1^+) \sum_{i=k+1}^{D} \left(\frac{p_i}{p_2^+} \right) \mathcal{D}\left((\boldsymbol{\alpha}_1, \alpha_2^+ + \tau_i)^\intercal \right) \end{aligned} \tag{3.47}$$

In a compositional data analysis framework, the distribution of $\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2$ coincides, up to a scale transformation, with the one of $\mathbf{S}_1 | \mathbf{X}_2 = \mathbf{x}_2$ (as shown in Equation (3.7)). The latter is more interesting because it helps study the neutrality on the left.

Proposition 15 (Distribution of conditionals)**.** *Let $\boldsymbol{X}$ be a random vector distributed as $EFD(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$, then:*

$$\boldsymbol{S}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim w(\boldsymbol{x}_2)\, EFD(\boldsymbol{\alpha}_1, \boldsymbol{\tau}_1, \bar{\boldsymbol{p}}_1'(\boldsymbol{x}_2)) + (1 - w(\boldsymbol{x}_2)) \mathcal{D}(\boldsymbol{\alpha}_1), \tag{3.48}$$

where:

$$w(\boldsymbol{x}_2) = \frac{\sum_{i \leq k} p_i'(\boldsymbol{x}_2)}{\sum_{i \leq D} p_i'(\boldsymbol{x}_2)}, \tag{3.49}$$

$$p'_i(\boldsymbol{x}_2) = \begin{cases} p_i \dfrac{\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_1^+ + \tau_i)} (1 - x_2^+)^{\tau_i}, & i = 1, \ldots, k \\ p_i \dfrac{\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_1^+)} \dfrac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)} x_i^{\tau_i}, & i = k+1, \ldots, D \end{cases} \tag{3.50}$$

and

$$\vec{p}'_i(\boldsymbol{x}_2) = \frac{p'_i(\boldsymbol{x}_2)}{\sum_{i \leq k} p'_i(\boldsymbol{x}_2)}, \qquad i = 1, \ldots, k. \tag{3.51}$$

In order to keep the notation as simple as possible, let $p'_i = p'_i(\mathbf{x}_2)$ and $\vec{p}'_i = \vec{p}'_i(\mathbf{x}_2)$. In order to have left neutrality (i.e. $\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2$), at leat one of the following must hold (see Proposition 17):

- $\tau_1 = \cdots = \tau_k = 1$ and $\dfrac{p_1}{\alpha_1} = \cdots = \dfrac{p_k}{\alpha_k}$
- $p_1 = \cdots = p_k = 0$
- $\tau_1 = \cdots = \tau_k = \tau$ and $p_{k+1} = \cdots = p_D = 0$

From Proposition 15 it is possible to compute the conditional expectation $\mathbb{E}[S_{1i} | \mathbf{X}_2 = \mathbf{x}_2]$. Let $c_1 = \displaystyle\sum_{r=1}^{k} \frac{\vec{p}'_{1r}}{\alpha_1^+ + \tau_r}$; then:

$$\begin{aligned}
\mathbb{E}[S_{1i} | \mathbf{X}_2 = \mathbf{x}_2] &= w(\mathbf{x}_2) \left(\alpha_{1i}\, c_1 + \frac{\tau_{1i}\, \vec{p}'_{1i}}{\alpha_1^+ + \tau_{1i}} \right) + (1 - w(\mathbf{x}_2)) \frac{\alpha_{1i}}{\alpha_1^+} \\
&= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2) \underbrace{\frac{\tau_{1i}}{\alpha_1^+ + \tau_{1i}}}_{L_i} \left(\frac{\alpha_{1i}\, c_1}{\tau_{1i}} (\alpha_1^+ + \tau_{1i}) + \vec{p}'_{1i} \right) - w(\mathbf{x}_2) \frac{\alpha_{1i}}{\alpha_1^+} \\
\overset{(*)}{} &= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2)\, L_i \left(-\frac{c_1\, \alpha_{1i}}{\alpha_1^+\, L_i} (L_i - 1)(\alpha_1^+ + \tau_{1i}) + \vec{p}'_{1i} \right) - w(\mathbf{x}_2) \frac{\alpha_{1i}}{\alpha_1^+} \\
&= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2)\, L_i \left(\vec{p}'_{1i} - \frac{c_1\, \alpha_{1i}}{\alpha_1^+} (\alpha_1^+ + \tau_{1i}) \left(1 - \frac{1}{L_i} \right) \right) - w(\mathbf{x}_2) \frac{\alpha_{1i}}{\alpha_1^+} \\
&= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2)\, L_i \left(\vec{p}'_{1i} - \frac{c_1\, \alpha_{1i}}{\alpha_1^+} (\alpha_1^+ + \tau_{1i}) + \frac{c_1}{\alpha_1^+}\, \alpha_{1i} (\alpha_1^+ + \tau_{1i}) \frac{1}{L_i} \right) - w(\mathbf{x}_2) \frac{\alpha_{1i}}{\alpha_1^+} \\
&= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2)\, L_i \left(\vec{p}'_{1i} - \frac{c_1\, \alpha_{1i}}{\alpha_1^+} (\alpha_1^+ + \tau_{1i}) \right) + w(\mathbf{x}_2) \cancel{L_i} \frac{c_1\, \alpha_{1i}}{\alpha_1^+} \frac{(\alpha_1^+ + \tau_{1i})}{\cancel{L_i}} - w(\mathbf{x}_2) \frac{\alpha_{1i}^+}{\alpha_1^+} \\
&= \frac{\alpha_{1i}}{\alpha_1^+} + w(\mathbf{x}_2)\, L_i \left(\vec{p}'_{1i} - \frac{\alpha_{1i}\, c_1}{\alpha_1^+} (\alpha_1^+ + \tau_{1i}) \right) + \frac{\alpha_{1i}}{\alpha_1^+}\, w(\mathbf{x}_2) \left((c_1\, (\alpha_1^+ + \tau_{1i})) - 1 \right)
\end{aligned}$$

Step $^{(*)}$ is obtained thanks to the fact that $\tau_{1i} = \dfrac{-\alpha_1^+\, L_i}{L_i - 1}$.

Figure 3.10 shows that the conditional expectation of the EFD model allows for regression lines that are different from the S-shaped typical to the FD. These lines are not forced to be monotonic: they have more flexibility and therefore they can better fit the data points.

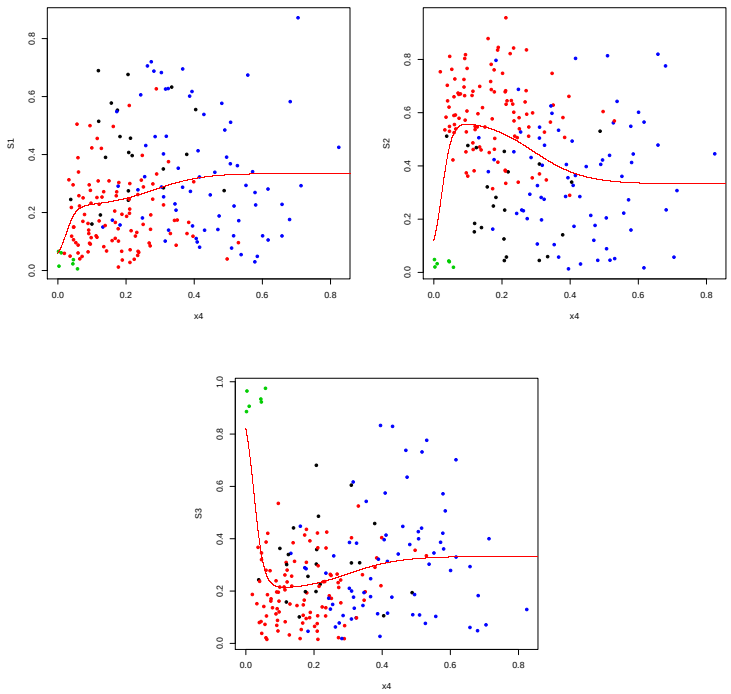

Fig. 3.10: EFD Conditional Expectation - $\boldsymbol{\alpha} = (2,2,2,2)^\intercal, \boldsymbol{\tau} = (1,4,70,3)^\intercal$ and $\mathbf{p} = (0.1,0.5,0.05,0.35)^\intercal$. Simulated data; each color defines a subpopulation (mixture component).

This expectation does not depend on $\mathbf{x}_2$ when the following jointly hold:

$$\begin{cases} c_1\left(\alpha_1^+ + \tau_{1i}\right) = 1 \\ \\ \bar{p}_i' = \dfrac{\alpha_i}{\alpha_1^+} \qquad i = 1,\ldots,k \end{cases} \tag{3.52}$$

If at least one of the following holds, the constraint 3.52 holds:

- $p_1 = \cdots = p_k = 0$ (it removes the dependence on $\mathbf{x}_2$ but it is not an acceptable set of values because it brings to $\bar{p}_i' = \frac{0}{0}$)
- $\tau_1 = \cdots = \tau_k$ and $p_{k+1} = \cdots = p_D = 0$
- $\tau_1 = \cdots = \tau_k$ and $\dfrac{\alpha_1}{p_1} = \cdots = \dfrac{\alpha_k}{p_k}$

Proposition 16 (Identifiability). *Let* $\boldsymbol{X} \sim EFD(\boldsymbol{\theta})$ *and* $\boldsymbol{X}' \sim EFD(\boldsymbol{\theta}')$*, where* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{p}, \boldsymbol{\tau})^\intercal$ *and* $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \boldsymbol{p}', \boldsymbol{\tau}')^\intercal$*. Then* $\boldsymbol{X} \sim \boldsymbol{X}'$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}'$*.*

Proof. It is obvious that if $\boldsymbol{\theta} = \boldsymbol{\theta}'$, then $\mathbf{X} \sim \mathbf{X}'$. In order to show the converse, one can focus on the marginal distribution of X_i. Let $g(x_i; \boldsymbol{\theta})$ be its density function, then:

$$g(x_i;\boldsymbol{\theta}) = x_i^{\alpha_i - 1}(1-x_i)^{\alpha^+ - \alpha_i - 1} \cdot \left\{ p_i \frac{\Gamma(\alpha^+ + \tau_i) x_i^{\tau_i}}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} + \right. \\ \left. + \sum_{l \neq i} p_l \frac{\Gamma(\alpha^+ + \tau_l)(1-x_i)^{\tau_l}}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l)} \right\}. \tag{3.53}$$

If $\mathbf{X} \sim \mathbf{X}'$, then $X_i \sim X_i'$ and therefore $g(x_i;\boldsymbol{\theta}) = g(x_i;\boldsymbol{\theta}')\ \forall\ x_i \in (0,1)$. Then, $\lim\limits_{x \to 0^+} \dfrac{g(x_i;\boldsymbol{\theta})}{x_i^{\alpha_i - 1}} = \lim\limits_{x \to 0^+} \dfrac{g(x_i;\boldsymbol{\theta}')}{x_i^{\alpha_i - 1}}$.

- $\lim\limits_{x_i \to 0^+} \dfrac{g(x_i;\boldsymbol{\theta})}{x_i^{\alpha_i - 1}} = \sum\limits_{l \neq i} p_l \dfrac{\Gamma(\alpha^+ + \tau_l)}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l)}$

- $\lim\limits_{x_i \to 0^+} \dfrac{g(x_i;\boldsymbol{\theta}')}{x_i^{\alpha_i - 1}} = \left(\lim\limits_{x_i \to 0^+} \dfrac{x_i^{\alpha_i' - 1}}{x_i^{\alpha_i - 1}} \right) \sum\limits_{l \neq i} p_l' \dfrac{\Gamma(\alpha'^+ + \tau_l')}{\Gamma(\alpha_i')\Gamma(\alpha'^+ - \alpha_i' + \tau_l')}$

In order to satisfy the equality of these two limits, the quantity $\left(\lim\limits_{x_i \to 0^+} \dfrac{x_i^{\alpha_i' - 1}}{x^{\alpha_i - 1}} \right)$ must be finite and different from 0:

$$\left(\lim_{x \to 0^+} \frac{x^{\alpha_k' - 1}}{x^{\alpha_k - 1}} \right) = \begin{cases} 0, \text{ if } \alpha_k' > \alpha_k \\ 1, \text{ if } \alpha_k' = \alpha_k \\ +\infty, \text{ if } \alpha_k' < \alpha_k \end{cases} \tag{3.54}$$

Then, it follows that $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$. This means that the equality $g(x_i; \boldsymbol{\theta}) = g(x_i; \boldsymbol{\theta}')$ can be re-written as:

$$\frac{p_i \Gamma(\alpha^+ + \tau_i) x_i^{\tau_i}}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} + \sum_{l \neq i} \frac{p_l \Gamma(\alpha^+ + \tau_l)(1 - x_i)^{\tau_l}}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l)} = \\ = \frac{p_i' \Gamma(\alpha^+ + \tau_i') x_i^{\tau_i'}}{\Gamma(\alpha_i + \tau_i')\Gamma(\alpha^+ - \alpha_i)} + \sum_{l \neq i} \frac{p_l' \Gamma(\alpha^+ + \tau_l')(1 - x_i)^{\tau_l'}}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l')}. \tag{3.55}$$

By taking the limits as $x_i \to 1^-$ on both sides, one can obtain:

$$p_i \frac{\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} = p_i' \frac{\Gamma(\alpha^+ + \tau_i')}{\Gamma(\alpha_i + \tau_i')\Gamma(\alpha^+ - \alpha_i)}. \tag{3.56}$$

Plugging it into the equality (3.55) and deriving both sides; the following equality must hold $\forall\, x_i \in (0, 1)$:

$$\frac{p_i \tau_i \Gamma(\alpha^+ + \tau_i) x_i^{\tau_i - 1}}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} - \sum_{l \neq i} \frac{p_l \tau_l \Gamma(\alpha^+ + \tau_l)(1 - x_i)^{\tau_l - 1}}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l)} = \\ = \frac{p_i \tau_i' \Gamma(\alpha^+ + \tau_i) x_i^{\tau_i' - 1}}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} - \sum_{l \neq i} \frac{p_l' \tau_l' \Gamma(\alpha^+ + \tau_l')(1 - x_i)^{\tau_l' - 1}}{\Gamma(\alpha_i)\Gamma(\alpha^+ - \alpha_i + \tau_l')} \tag{3.57}$$

Taking the limits as $x_i \to 1^-$ on both sides:

$$p_i \tau_i \frac{\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)} = p_i \tau_i' \frac{\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_i + \tau_i)\Gamma(\alpha^+ - \alpha_i)}. \tag{3.58}$$

It follows that $\tau_i = \tau_i' \; \forall\, i \Longrightarrow \boldsymbol{\tau} = \boldsymbol{\tau}'$. Finally, substituting this constraint in (3.56), it is possible to conclude that $\mathbf{p} = \mathbf{p}'$.

□

Proposition 17. *The EFD distribution allows for a variety of simplicial forms of independence. Let* $\boldsymbol{X} \sim EFD(\boldsymbol{\alpha}, \boldsymbol{p}, \tau)$*, then:*

- ***X*** *has left neutrality if at least one of the following holds:*
  - $\tau_1 = \cdots = \tau_k = 1$ *and* $\dfrac{p_1}{\alpha_1} = \cdots = \dfrac{p_k}{\alpha_k}$
  - $p_1 = \cdots = p_k = 0$
  - $\tau_1 = \cdots = \tau_k = \tau$ *and* $p_{k+1} = \cdots = p_D = 0$
- ***X*** *has right neutrality if at least one of the following holds:*
  - $\tau_{k+1} = \cdots = \tau_D = 1$ *and* $\dfrac{p_{k+1}}{\alpha_{k+1}} = \cdots = \dfrac{p_D}{\alpha_D}$
  - $p_{k+1} = \cdots = p_D = 0$
  - $\tau_{k+1} = \cdots = \tau_D = \tau$ *and* $p_1 = \cdots = p_k = 0$
- ***X*** *has subcompositional independence if* ***X*** *has right or left neutrality.*
- ***X*** *has complete left neutrality if at least one of the following holds:*
  - $\tau_1 = \cdots = \tau_{D-1} = 1$ *and* $\dfrac{p_1}{\alpha_1} = \cdots = \dfrac{p_{D-1}}{\alpha_{D-1}}$
  - $p_1 = \cdots = p_{D-1} = 0$
- ***X*** *has complete right neutrality if at least one among the following is satisfied:*
  - $\tau_2 = \cdots = \tau_D = 1$ *and* $\dfrac{p_2}{\alpha_2} = \cdots = \dfrac{p_D}{\alpha_D}$
  - $p_2 = \cdots = p_D = 0$

These results can be obtained noting that the EFD distribution can be expressed in term of a Generalized Liouville distribution of the second kind. According to Smith and Rayens [85], this distribution is characterized by the following density function:

$$g(\mathbf{x}; \boldsymbol{\alpha}, \beta_1, \ldots, \beta_D, q_1, \ldots, q_D) = A \prod_{r=1}^{D} x_r^{\alpha_r - 1} f\left(\sum_{r=1}^{D} \left(\frac{x_r}{q_r} \right)^{\beta_r} \right), \tag{3.59}$$

where $\alpha_r > 0$, $\beta_r > 0$ and $q_r > 0$ for every $r = 1, \ldots, D$, $f(x) = x$ and A is the normalization constant. Imposing

$$\begin{cases} \beta_r = \tau_r \\ q_r = \dfrac{\Gamma(\alpha_r + \tau_r)}{p_r \Gamma(\alpha_r)\Gamma(\alpha^+ + \tau_r} \end{cases}$$

one obtains the EFD density function and, therefore, take advantage of the independence properties listed in Smith and Rayens [85].

It has already been said that the Extended Flexible Dirichlet distribution allows for a finite mixture structure. This means that it is possible to study the behaviour of its components. A simple way is to compute some measures able to capture the degree of overlap among the mixture components. This can be done through the Kullback-Leibler divergence measure [38, 54, 60], that quantifies how much two probability distributions differ from one another. In particular, let $f_1(\mathbf{x}; \boldsymbol{\theta})$ and $f_2(\mathbf{x}; \boldsymbol{\theta}')$ be two probability density functions and, in order to make notation clearer, let $f_1 \equiv f_1(\mathbf{x}; \boldsymbol{\theta})$ and $f_2 \equiv f_2(\mathbf{x}; \boldsymbol{\theta}')$. Then,

$$d_{KL}(f_1, f_2) = \int f_1(\mathbf{x}; \boldsymbol{\theta}) \ln \frac{f_1(\mathbf{x}; \boldsymbol{\theta})}{f_2(\mathbf{x}; \boldsymbol{\theta}')} d\mathbf{x}. \tag{3.60}$$

Looking at (3.60) it is easy to show that the Kullback-Leibler divergence is not symmetric with respect to its arguments f_1 and f_2, since $d_{KL}(f_1, f_2) \neq d_{KL}(f_2, f_1)$. In their work, Kullback and Leibler defined "divergence" between f_1 and f_2 the sum of $d_{KL}(f_1, f_2)$ and $d_{KL}(f_2, f_1)$. Nowadays, this quantity is called "symmetrized Kullback-Leibler divergence":

$$d_{SKL}\left(f_1, f_2\right) = d_{KL}\left(f_1, f_2\right) + d_{KL}\left(f_2, f_1\right) \tag{3.61}$$

When this measure approaches the value 0, the densities f_1 and f_2 are very similar. Indeed, if $f_1 = f_2$ almost everywhere, then $d_{SKL}(f_1, f_2)$ is exactly equal to 0. In order to compute this measure of divergence in the EFD context, it is useful to remember that:

- The i-th mixture component follows a Dirichlet distribution: $f_i \equiv f_i(\mathbf{x}; \boldsymbol{\alpha}, \tau) = f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i)$.

- If $\mathbf{X} \sim \mathcal{D}(\boldsymbol{\alpha}) \Longrightarrow \mathbb{E}\left[\ln X_r\right] = \psi(\alpha_r) - \psi(\alpha^+)$, where $\psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x)$ is the digamma function.

In order to compute $d_{SKL}(f_i, f_h)$, it is convenient to compute the following fraction for every $i \neq h$ $(i, h = 1, \ldots, D)$:

$$\begin{aligned}
\frac{f_i}{f_h} &= \frac{f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i)}{f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau_h \, \mathbf{e}_h)} \\
&= \frac{\frac{\Gamma(\alpha^+ + \tau_i)}{\prod_{k \neq i} \Gamma(\alpha_k)} \frac{x_i^{\alpha_i + \tau_i - 1}}{\Gamma(\alpha_i + \tau_i)} \prod_{k \neq i} x_k^{\alpha_k - 1}}{\frac{\Gamma(\alpha^+ + \tau_h)}{\prod_{k \neq h} \Gamma(\alpha_k)} \frac{x_h^{\alpha_h + \tau_h - 1}}{\Gamma(\alpha_h + \tau_h)} \prod_{k \neq h} x_k^{\alpha_k - 1}} \\
&= \underbrace{\left(\frac{\Gamma(\alpha^+ + \tau_i)\Gamma(\alpha_i)\Gamma(\alpha_h + \tau_h)}{\Gamma(\alpha^+ + \tau_h)\Gamma(\alpha_h)\Gamma(\alpha_i + \tau_i)} \right)}_{C_{i,h}} \frac{x_i^{\tau_i}}{x_h^{\tau_h}} \\
&= C_{i,h} \frac{x_i^{\tau_i}}{x_h^{\tau_h}}
\end{aligned}$$

Then, the logarithm of the ratio (3.62) is:

$$\ln \frac{f_i}{f_h} = \ln C_{i,h} + \tau_i \ln x_i - \tau_h \ln x_h. \tag{3.62}$$

Note that $\ln C_{i,h} = -\ln C_{h,i}$.

$$\begin{aligned}
d_{KL}(f_i, f_h) &= \int f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha}_i) \ln \frac{f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha}_i)}{f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha}_h)} d\mathbf{x} \\
&= \int f_i \left(\ln C_{i,h} + \tau_i \ln x_i - \tau_h \ln x_h \right) d\mathbf{x} \\
&= \ln C_{i,h} + \tau_i \mathbb{E}\left[\ln X_i \right] - \tau_h \mathbb{E}\left[\ln X_h \right] \\
&= \ln C_{i,h} + \tau_i \left[\psi(\alpha_i + \tau_i) - \psi(\alpha^+ + \tau_i) \right] - \tau_h \left[\psi(\alpha_h) - \psi(\alpha^+ + \tau_i) \right]
\end{aligned}$$

where $\mathbb{E}\left[\cdot\right]$ is with respect to $\mathcal{D}(\mathbf{x}; \boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i)$.

$$\begin{aligned}
d_{SKL}(f_i, f_h) &= d_{KL}(f_i, f_h) + d_{KL}(f_h, f_i) \\
&= \cancel{\ln C_{i,h}} + \tau_i \left[\psi(\alpha_i + \tau_i) - \psi(\alpha^+ + \tau_i) \right] - \tau_h \left[\psi(\alpha_h) - \psi(\alpha^+ + \tau_i) \right] + \\
&\quad + \cancel{\ln C_{h,i}} + \tau_h \left[\psi(\alpha_h + \tau_h) - \psi(\alpha^+ + \tau_h) \right] - \tau_i \left[\psi(\alpha_i) - \psi(\alpha^+ + \tau_h) \right] \\
&= \tau_i \psi(\alpha_i + \tau_i) - \tau_i \psi(\alpha^+ + \tau_i) - \tau_h \psi(\alpha_h) + \tau_h \psi(\alpha^+ + \tau_i) +
\end{aligned}$$

$$
\begin{aligned}
& +\tau_h\psi(\alpha_h+\tau_h)-\tau_h\psi(\alpha^+ +\tau_h)-\tau_i\psi(\alpha_i)+\tau_i\psi(\alpha^+ +\tau_h) \\
= \;& \tau_i\left[\psi(\alpha_i+\tau_i)-\psi(\alpha_i)\right]+\tau_h\left[\psi(\alpha_h+\tau_h)-\psi(\alpha_h)\right]+ \\
& +(\tau_i-\tau_h)\left[\psi(\alpha^+ +\tau_h)-\psi(\alpha^+ +\tau_i)\right]
\end{aligned} \tag{3.63}
$$

Thanks to a graphical investigation of various EFD's contour plots, it is possible to say that values of d_{SKL} greater than 15 entail well-separated clusters.

3.4.2 Estimation procedure

Given the mixture structure of the EFD model, it is possible to propose an EM algorithm for maximizing the corresponding likelihood function [28]. Let us suppose to have a sample of n independent observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ from an EFD distribution; the observed log-likelihood can be thought of as originated from the following complete-data log-likelihood:

$$
\log L_C(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{s=1}^{n}\sum_{i=1}^{D} z_{s,i}\left\{\log p_i + \log f_{\mathcal{D}}(\mathbf{x}_s; \boldsymbol{\alpha}+\tau_i\,\mathbf{e}_i)\right\}, \tag{3.64}
$$

where $z_{s,i}$ is equal to 1 if the s-th observation has arisen from the i-th component of the mixture and 0 otherwise. We may apply the EM algorithm if one thinks of the vector $\mathbf{z}_j = (z_{s,1}, \ldots, z_{s,D})^\intercal$, $s = 1, \ldots, n$, as the missing component.

The EM algorithm is an iterative method whose generic $(m+1)$-th step can be described as follows:

- **E-step**: Given the parameter estimates obtained at step m, $\left(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\tau}^{(m)}, \mathbf{p}^{(m)}\right)^\intercal$, compute the conditional expectation of the complete-data log-likelihood given $\mathbf{x}$ as:

$$
\sum_{i=1}^{D}\sum_{s=1}^{n} \pi_i\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}, \boldsymbol{\tau}^{(m)}, \mathbf{p}^{(m)}\right)\left\{\log p_i^{(m)} + \log f_{\mathcal{D}}\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}+\tau_i^{(m)}\mathbf{e}_i\right)\right\}, \tag{3.65}
$$

  where

$$
\pi_i\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}, \boldsymbol{\tau}^{(m)}, \mathbf{p}^{(k)}\right) = \frac{p_i^{(m)} f_{\mathcal{D}}\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}+\tau_i^{(m)}\mathbf{e}_i\right)}{\sum_{h=1}^{D} p_h^{(m)} f_{\mathcal{D}}\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}+\tau_h^{(m)}\,\mathbf{e}_h\right)}, \quad i = 1, \ldots, D, \tag{3.66}
$$

  is the "posterior" probability that $\mathbf{x}_s$ belongs to the i-th component of the mixture given $\left(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\tau}^{(m)}, \mathbf{p}^{(m)}\right)^\intercal$.

- **M-step**: Maximize the conditional expectation (3.65) to update the parameter estimates. In order to obtain new values of $\hat{\boldsymbol{\alpha}}^{(m+1)}$ and $\hat{\boldsymbol{\tau}}^{(m+1)}$ a numeric maximization method (e.g. Newton-Raphson) is required, whereas a closed-form expression for $\hat{p_i}^{(m+1)}$ exists:

$$\hat{p_i}^{(m+1)} = \frac{1}{n}\sum_{s=1}^{n} \pi_i\left(\mathbf{x}_s; \boldsymbol{\alpha}^{(m)}, \boldsymbol{\tau}^{(m)}, \mathbf{p}^{(m)}\right), \qquad i = 1, \ldots, D-1. \quad (3.67)$$

E and M steps are alternated until a convergence criterion is reached (e.g. when there is a small difference between the log-likelihood of two consecutive steps and the distance between the estimates of parameters in two consecutive iterations is lower than a fixed threshold. Unfortunately, the EM algorithm typically leads to solutions that are highly dependent on the starting point; this means that the algorithm may get trapped in a local maxima close to the starting point. In order to weaken this dependence, a Stochastic EM (SEM) algorithm has been used [15, 21, 22]. SEM is a modified version of the classic EM which is likely to explore a wider region of the parametric space. The final estimation algorithm therefore implements a SEM phase followed by an EM one: the results of the SEM algorithm are used as starting points of a proper EM algorithm, which is very precise in finding maxima close to initial values. In SEM, after the E step, a new partition of data into D groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_D\}$ is generated, with a draw from a Multinomial distribution with parameters equal to the current estimates of the conditional probabilities $\pi_i(\cdot)$, $i = 1, \ldots, D$, given by (3.66). In this way, the algorithm has a chance to escape from a path of convergence to a local maximizer instead of a global one. Finally, the M step of the SEM consists in updating the weights $\mathbf{p}$ as the relative number of observations in each group. Updatings of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ are obtained maximizing the classified likelihood (i.e. the likelihood computed by assuming knowledge of the mixture component each observation comes from):

$$\prod_{i=1}^{D}\prod_{s\in A_i} f_{\mathcal{D}}(\mathbf{x}_s; \boldsymbol{\alpha} + \tau_i\, \mathbf{e}_i),$$

where $A_i = \{s : \mathbf{x}_s \in \mathcal{G}_i\}$. This maximization problem can be approached numerically (i.e. with Quasi-Newton optimization algorithms [20]).

Once an estimate for$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})^\intercal$ is obtained, it must be supplemented by the information on its sample variability. A well known result from statistical theory is that, under regularity conditions, the asymptotic covariance matrix of the ML Estimator $\hat{\boldsymbol{\theta}}$ can be approximated by the inverse of the observed information matrix

$\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{x})$. In order to compute the observed counterpart for this matrix, the second-order derivatives of the mixture log-likelihood, defined as:

$$\ln L_M(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \ln \left(\prod_{s=1}^{n} f_{EFD}(\mathbf{x}_s; \boldsymbol{\theta}) \right), \tag{3.68}$$

are required. Unfortunately, the evaluation is quite complicated, especially in this scenario. The method adopted by Migliorati, Ongaro and Monti [63] based on the work of Louis [56] can be adapted to the EFD: an evaluation of $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{x})$ can be obtained via decomposition of complete-data into observed and missing ones, so that the observed information matrix can be written as:

$$\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{x}) = \{\mathbb{E}_{\boldsymbol{\theta}} [\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}_c)|\mathbf{x}]\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \{\mathbb{E}_{\boldsymbol{\theta}} [\mathbf{S}_c(\boldsymbol{\theta}; \mathbf{X}_c)\mathbf{S}_c^{\intercal}(\boldsymbol{\theta}; \mathbf{X}_c)|\mathbf{x}]\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.69}$$

where $\mathbf{X}_c = (\mathbf{X}, \mathbf{Z})^{\intercal}$, $\mathbf{S}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ is the complete-data score statistics and $\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ is the negative Hessian matrix of the complete-data log-likelihood (3.64). Evaluation of the conditional expected value in (3.69) can be based on conditional bootstrap [29], noting that, conditionally on $\mathbf{x}$, the random vectors $\mathbf{Z}_s$, $s = 1, \ldots, n$ are distributed as independent multinomials with parameters $(p^*_{s,1}, \ldots, p^*_{s,D})^{\intercal}$, where $p^*_{s,i} = \pi_i(\mathbf{x}_s; \hat{\boldsymbol{\theta}})$ is given by (3.66). Consequently, the conditional expectations can be approximated the average of draws $\mathbf{z}_s^{(b)} \sim \mathbf{Z}_s$, over B independent bootstrap samples $(s = 1, \ldots, n; b = 1, \ldots, B)$ for a sufficiently high value of B.

3.4.3 An open problem: how to initialize the EM algorithm?

The convergence of the EM algorithm can be influenced by initial values required for the first steps of the algorithm. In general, this choice can even influence the ability to locate the global maximum of the log-likelihood function. To address these critical issues, some suitable ad hoc initialization strategies have been developed. Usually, the first step of the initialization consists in obtaining a partition of the n observations into D groups by means of a clustering method. The clustering algorithm proposed in this work is called "barycenter method", and it is based on the peculiar cluster structure of the EFD: observation $\mathbf{x}_s$ is assigned to group i if $x_{s,i}/x_{s,h} > B_i/B_h$, $\forall$ $h = 1, \ldots, D, h \neq i$, where $\mathbf{B} = (B_1, \ldots, B_D)^{\intercal}$ is a data barycenter (e.g. mean or median). Since any clustering algorithm assigns the group labels randomly, the groups have been relabelled on the basis of the structure imposed by the EFD model illustrated in section 3.4: group i will have the largest mean in component i. In the case a single group shows two or more components with maximum sample mean, the labelling procedure considers the permutations of labels compatible with the largest sample mean positions and choose the one that maximizes the corresponding likelihood. Then, given this partition, a possible initial value for p_i is the proportion of observations that are assigned to cluster i. The initialization of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ is a more

challenging problem. The initialization method used for the FD model [63] and two new ad hoc strategies have been considered here:

1) The procedure used for the FD in [63]. Initialization values for $\boldsymbol{\alpha}$ and τ (namely $\boldsymbol{\alpha}^*$ and τ^*) are obtained using the method of moments: in the EFD context an initialization for the vector $\boldsymbol{\tau}$ can be obtained imposing that each element τ_i is equal to τ^*: $\boldsymbol{\tau} = (\tau, \tau, \ldots, \tau)^\intercal$. This method assumes the FD's structure of the cluster means and, therefore, can be expected to produce inaccurate results if data do not show equal distance between the barycenter $\bar{\boldsymbol{\alpha}}$ and each cluster means.

2) Given a partition, one can compute the sample mean for each cluster: $\bar{\mathbf{x}}_h = (\bar{x}_{h,1}, \ldots, \bar{x}_{h,D})^\intercal$, $h = 1, \ldots, D$, where $\bar{x}_{h,i} = \sum_{s=1}^{n} z_{h,s} x_{s,i}$. Initial values of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ can then be obtained by minimizing the distance between $\bar{\mathbf{x}} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_D)^\intercal$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_D)^\intercal$:

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\tau}} \sum_{h=1}^{D} \delta_h (\boldsymbol{\mu}_h - \bar{\mathbf{x}}_h)^\intercal (\boldsymbol{\mu}_h - \bar{\mathbf{x}}_h), \tag{3.70}$$

where δ_h are suitable weights (e.g. the size of each group) and $\boldsymbol{\mu}_h$ is defined as in (3.41). Let $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$ and $\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau}/\alpha^+$ be the "relative" counterparts of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$, then:

$$\boldsymbol{\mu}_h = \frac{\bar{\boldsymbol{\alpha}}}{1 + \tilde{\tau}_h} + \frac{\tilde{\tau}_h}{1 + \tilde{\tau}_h} \mathbf{e}_h.$$

Since the constraints $\sum_{h=1}^{D} \bar{\alpha}_h = 1$, $\bar{\alpha}_h > 0$, $\tilde{\tau}_h > 0$, $h = 1. \ldots, D$ hold, this is a constrained minimization problem and it can be fulfilled numerically with a Quasi-Newton algorithm [20]. Since this approach requires a starting point, the FD's initialization method can be used assuming $\boldsymbol{\tau} = (\tau, \ldots, \tau)^\intercal$.

3) The above constrained minimization can also be approached analytically. Setting the partial derivatives of the target function (with respect to $\tilde{\alpha}_h$ and $\tilde{\tau}_h$) equal to zero, one obtains:

$$\tilde{\alpha}_h \left[\sum_{l=1}^{D} \frac{\delta_l}{(1 + \tilde{\tau}_l)^2} \right] = \sum_{l=1}^{D} \bar{x}_{h,l} \frac{\delta_l}{(1 + \tilde{\tau}_l)} - \frac{\tilde{\tau}_h \delta_h}{(1 + \tilde{\tau}_l)^2}, \tag{3.71}$$

$\bar{x}_{h,l}$ is the l-th element of $\bar{x}_h$ and

$$\tilde{\tau}_h = \begin{cases} \dfrac{b_h}{c_h} & \text{if } b_h > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.72}$$

where $b_h = (\bar{\mathbf{x}}_h - \bar{\boldsymbol{\alpha}})^\intercal(\mathbf{e}_h - \bar{\boldsymbol{\alpha}})$ and $c_h = \sum_{r=1}^{D} \bar{\alpha}_r \bar{x}_{h,r} + 1 - \bar{\alpha}_h - \bar{x}_{h,h}$ (note that c_h is always positive). It can be immediately observed that solutions for $\bar{\alpha}_h$'s depend on $\tilde{\tau}_h$'s and viceversa. The final algorithm is:

1. initialize the $\bar{\alpha}_h$'s (i.e. initializing $\boldsymbol{\alpha}^*$ with the FD method and computing $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^* / \sum_{k=1}^{D} \alpha_k^*$)

2. calculate the $\tilde{\tau}_h$'s on the basis of (3.72)

3. calculate the $\bar{\alpha}_h$'s on the basis of (3.71)

4. repeat step 2 and 3 until a convergence criterion is satisfied.

Methods 2 and 3 have some technical issues: whereas the constraint $\sum_{h=1}^{D} \bar{\alpha}_h = 1$ is automatically satisfied, other constraints could be violated:

- $\tilde{\tau}_h$ could be equal to 0. In this case $\tilde{\tau}_h$ is set equal to a very small positive quantity (e.g. 0.00001).

- $\bar{\alpha}_h$ could be negative: if this happens, $\tilde{\alpha}_h$ is set equal to a very small positive quantity and the remaining $\tilde{\alpha}_h$'s are re-normalized.

Another issue with methods 2 and 3 is that they only allow for initialization of the relative quantities $\bar{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\tau}}$: in order to initialize α^+ one can resort to the variances:

$$\text{Var}\,(X_k|\mathbf{Z} = \mathbf{e}_h) = \sigma_{h,k}^2 = \frac{\mu_{h,k}(1 - \mu_{h,k})}{\alpha^+ + \tau_h + 1}, \tag{3.73}$$

where $\mu_{h,k}$ is the k-th element of μ_h and and h ($h = 1, \ldots, D$) indices clusters. The estimate of each $\sigma_{h,k}^2$ is $s_{h,k}^2$, the sample variance of component k among group h. With some algebra, one can obtain:

$$\widehat{\alpha^+ + \tau_h} = \frac{1 - \sum_{k=1}^{D} \bar{x}_{h,k}^2}{\sum_{k=1}^{D} s_{h,k}^2} - 1, \qquad h = 1, \ldots, D.$$

The sum of variances in the denominator permits to have stable estimates whenever some $s_{h,k}^2$ is close to zero. These estimates can be used to obtain several estimates of α^+:

$$\widehat{\alpha^+}_{(h)} = \frac{\widehat{\alpha^+ + \tau_h}}{1 + \tilde{\tau}_h}, \qquad h = 1, \ldots, D,$$

where $\tilde{\tau}_h$ was obtained with one of the methods above. Finally, one can aggregate the $\widehat{\alpha^+}_{(h)}$'s using a weighted mean.

3.4.4 A Simulation Study

Two simulation studies have been set up. The first one is aimed at investigating the behavior of the initialization procedures proposed in section 3.4.3, whereas the second one is aimed at evaluating the performance of the EM algorithm and the variance estimator (section 3.4.2). In order to implement these studies, 21 parameter configurations have been investigated. In this section it is possible to find the results for five of them; in Table 8.8 (appendix 8.3) it is possible to find all the configurations. The chosen parametric configurations allow to cover a great variety of cases, including well-separated as well as overlapping clusters, according to the Symmetrized Kullback-Leibler divergence.

ID	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3	$d_{SKL}(f_1, f_2)$	$d_{SKL}(f_1, f_3)$	$d_{SKL}(f_2, f_3)$
1	15	15	15	20	20	20	1/3	1/3	1/3	34.666	34.666	34.666
11	10	40	80	5	30	25	1/3	1/3	1/3	14.801	6.176	23.627
13	50	50	50	5	30	25	0.2	0.6	0.2	10.945	8.267	24.292
20	15	15	15	10	20	15	1/3	1/3	1/3	20.892	15.456	27.581
21	5	30	70	10	25	15	0.1	0.75	0.15	25.180	14.400	17.472

Tab. 3.3: A subset of parameter configurations. See Table 8.8 for the complete list.

On the choice of the initialization method

This simulation study considers only the five configurations in Table 3.3. To evaluate which of the three methods described in subsection 3.4.3 provides the best initialization, $K = 100$ datasets have been simulated for each configuration and the clustering methods described in subection 3.4.3 have been applied to them. The one that provided the best performance has been selected. Given the resulting data partition, an initial estimate for $\mathbf{p}$ is obtained as in (3.67). Then, the three methods built for initializing $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ have been applied to each dataset. These initializations stand for the starting point of a SEM+EM procedure that provides the final estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$. Table 3.4 shows the results of these simulations for each initialization method (rows):

- the first column "%" reports the proportion of simulations where the EFD likelihood evaluated at the initial values is the highest one; the second column "%" reports the proportion of simulations where the final estimates maximize the likelihood function
- columns "Mean $\hat{l}$" represent the mean of the likelihoods evaluated at the initial values and at the final estimates

- columns "Mean d_2" represent the mean of the euclidean distances between the initial values (or the final estimates) and the true parameter values.

ID 1: $\boldsymbol{\alpha} = (15, 15, 15)^\intercal$, $\boldsymbol{\tau} = (20, 20, 20)^\intercal$, $\mathbf{p} = (1/3, 1/3, 1/3)^\intercal$

	Initial Values			Final Estimates		
Meth.	%	Mean $\hat{l}$	Mean d_2	%	Mean $\hat{l}$	Mean d_2
1	0.02	209.76901	5.1489	0.32	211.06737	5.6106
2	0.38	210.86665	6.1022	0.35	211.06731	5.6159
3	0.60	210.90500	5.5194	0.33	211.06734	5.6110

ID 11: $\boldsymbol{\alpha} = (10, 40, 80)^\intercal$, $\boldsymbol{\tau} = (5, 30, 25)^\intercal$, $\mathbf{p} = (1/3, 1/3, 1/3)^\intercal$

	Initial Values			Final Estimates		
Meth.	%	Mean $\hat{l}$	Mean d_2	%	Mean $\hat{l}$	Mean d_2
1	0.00	285.82046	23.3435	0.38	349.49176	13.4146
2	0.61	347.33590	12.4744	0.33	349.49008	13.1921
3	0.39	347.26455	11.9301	0.29	349.49127	13.0897

ID 13: $\boldsymbol{\alpha} = (50, 50, 50)^\intercal$, $\boldsymbol{\tau} = (5, 30, 25)^\intercal$, $\mathbf{p} = (0.2, 0.6, 0.2)^\intercal$

	Initial Values			Final Estimates		
Meth.	%	Mean $\hat{l}$	Mean d_2	%	Mean $\hat{l}$	Mean d_2
1	0.00	300.47035	27.2624	0.34	325.76015	11.9057
2	0.00	316.19300	19.0152	0.38	325.76017	11.9268
3	1.00	316.81200	15.5242	0.28	325.76004	11.7613

ID 20: $\boldsymbol{\alpha} = (15, 15, 15)^\intercal$, $\boldsymbol{\tau} = (10, 20, 15)^\intercal$, $\mathbf{p} = (1/3, 1/3, 1/3)^\intercal$

	Initial Values			Final Estimates		
Meth.	%	Mean $\hat{l}$	Mean d_2	%	Mean $\hat{l}$	Mean d_2
1	0.00	194.93455	9.0557	0.34	204.83934	5.1533
2	0.11	204.18551	5.8822	0.35	204.83930	5.1493
3	0.89	204.29076	5.3072	0.31	204.83936	5.1125

ID 21: $\boldsymbol{\alpha} = (5, 30, 70)^\intercal$, $\boldsymbol{\tau} = (10, 25, 15)^\intercal$, $\mathbf{p} = (0.1, 0.75, 0.15)^\intercal$

	Initial Values			Final Estimates		
Meth.	%	Mean $\hat{l}$	Mean d_2	%	Mean $\hat{l}$	Mean d_2
1	0	323.9595	23.9439	0.25	376.7433	11.6179
2	0.05	337.6969	23.4810	0.37	376.7447	11.4295
3	0.95	339.0964	21.7384	0.38	376.7442	11.5162

Tab. 3.4: Simulation results: initialization.

Method 3 generally provides the best starting points, with method 2 displaying only slightly worse performances. On the contrary, method 1 behaves rather poorly compared to the other two, except in the symmetric scenario 1, as expected. Remarkably, after the SEM+EM step, the differences between the three methods are not significant. This evidentiates a strong robustness of the SEM phase with respect to the choice of the initial value. In the following, method 3 will be considered, since it is also the one with the fastest convergence.

EM algorithm performance

The aim of the second simulation study is the evaluation of the performance of the EM algorithm and of the estimated variance. For each of the 21 parameter configurations used in subsection 3.4.4, $K = 1000$ samples of size $n = 100$ have been generated. After every estimation procedure, a conditional bootstrap algorithm has been launched (with $B = 3000$ bootstrap samples), to produce an estimate of the standard errors that can be used to compute confidence intervals as well (based on the asymptotic normal distribution of the ML estimator). Table 3.5 shows the results of the simulations for the configurations reported in Table 3.3 ; results for all the configurations can be found in appendix 8.3. Rows "MLE mean" and "MLE sd" represent the simulated mean and standard deviation of the ML estimator (namely, the Monte Carlo approximation of its expected value and standard error). The quantity "SE mean" shows the mean of the bootstrap based simulated standard errors and the row "arb" represents its absolute relative bias (i.e. the mean of the absolute deviations between such standard errors' estimates and the simulated standard deviation - row "MLE sd" - divided by this last quantity). Lastly, "Coverage" reports the simulated coverage levels of confidence intervals for a nominal level $1 - \alpha = 0.95$.

Despite the mixture nature of the EFD model and the relatively small sample size, the performance of the MLE appears rather satisfactory: in most of the scenarios we have considered, small bias and standard deviation are obtained. Furthermore, the bootstrap estimates of the standard errors are remarkably close to the Monte Carlo approximations (here considered as the gold standard) and the coverage levels of the confidence intervals are fairly precise. It is also worth noting that the results relative to the other parameter configurations included in appendix are similar to the reported ones. As a consequence, it is possible to conclude that the proposed estimation procedure appears to be both accurate and reliable.

Case 1	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	1/3	1/3	15	15	15	20	20	20
MLE Mean	0.333	0.332	15.561	15.549	15.6	20.757	20.841	20.828
MLE sd	0.047	0.047	1.655	1.674	1.673	2.793	2.784	2.798
SE mean	0.047	0.047	1.618	1.617	1.623	2.744	2.753	2.749
arb	0.028	0.029	0.080	0.083	0.083	0.080	0.079	0.080
Coverage	0.951	0.952	0.946	0.943	0.942	0.944	0.947	0.951

Case 2	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	1/3	1/3	10	40	80	5	30	25
MLE Mean	0.336	0.331	10.491	41.982	83.890	5.388	31.506	26.901
MLE sd	0.073	0.051	1.259	5.166	10.007	1.461	4.545	6.742
SE mean	0.07	0.05	1.228	5.006	9.834	1.376	4.500	6.491
arb	0.168	0.048	0.083	0.080	0.078	0.135	0.075	0.105
Coverage	0.919	0.936	0.939	0.942	0.950	0.935	0.947	0.931

Case 3	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.2	0.6	50	50	50	5	30	25
MLE Mean	0.201	0.595	52.496	52.643	52.555	5.808	31.499	26.607
MLE sd	0.058	0.057	6.091	6.458	6.211	3.814	4.539	5.080
SE mean	0.056	0.054	5.900	6.155	5.907	3.750	4.437	4.966
arb	0.166	0.076	0.085	0.088	0.088	0.198	0.087	0.112
Coverage	0.922	0.933	0.946	0.943	0.937	0.962	0.941	0.930

Case 4	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	1/3	1/3	15	15	15	10	20	15
MLE Mean	0.333	0.333	15.626	15.624	15.626	10.408	20.935	15.700
MLE sd	0.050	0.047	1.702	1.695	1.718	1.991	2.914	2.454
SE mean	0.050	0.048	1.691	1.700	1.694	1.941	2.876	2.411
arb	0.044	0.038	0.078	0.075	0.079	0.078	0.077	0.076
Coverage	0.950	0.944	0.958	0.949	0.943	0.952	0.950	0.95

Case 5	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	10	25	15
MLE Mean	0.100	0.750	5.212	31.598	73.177	10.531	25.942	16.764
MLE sd	0.031	0.045	0.582	4.021	8.177	2.193	3.820	8.670
SE mean	0.032	0.045	0.555	3.827	7.878	2.048	3.761	8.232
arb	0.132	0.059	0.090	0.098	0.086	0.164	0.086	0.147
Coverage	0.936	0.936	0.943	0.940	0.945	0.928	0.957	0.941

Tab. 3.5: Simulation results: performance of parameter estimates and confidence intervals.

4 The Double Flexible Dirichlet

In Section 3 several models suitable for compositional data have been illustrated. Among them, the more promising is the Extended Flexible Dirichlet, thanks to its flexibility in clusters' location on the simplex. A common drawback of all the models illustrated regards the number of clusters available: indeed, the Dirichlet and the Additive Logistic-Normal distributions do not consider that a sample can be generated by several subpopulations, whereas both the FD and the EFD distributions allow for up to D clusters (and, consequently, modes). These clusters must have their barycenter placed in a triangle shape on the simplex. This means that it is not possible either to consider more components than the length of the composition $\mathbf{X}$ or to place the cluster means in a more elaborated configuration. To overcome such an issue, in this section a new generalization of the Flexible Dirichlet is proposed: the Double Flexible Dirichlet (DFD). Thanks to its particular mixture structure it allows for $\frac{D(D+1)}{2}$ clusters, located in a very precise way in the simplex. Furthermore, it allows for a more general covariance structure than the one induced by the FD, thanks to the larger number of mixture components. Indeed, the number of parameters of this model is $(D+1)+\frac{D(D+1)}{2}$, it is strictly connected to the number of distinct elements in the covariance matrix. This makes the model for the dependence structure more flexible. Some theoretical properties are reported and an estimation procedure is also proposed. To test the reliability of this estimation algorithm, a simulation study has been conducted.

4.1 The basis

4.1.1 Constructing the basis

Let us assume that the vectors $\mathbf{W}=(W_1,\dots,W_D)^\intercal$, $\mathbf{U}=(U_1,U_2)^\intercal$ and $\mathbf{Z}=(\mathbf{Z}_1,\mathbf{Z}_2)^\intercal$ are jointly independent and that:

- the vector $\mathbf{W}$ has independent Gamma components with scale parameter equal to 1: $W_r\sim\text{Gamma}(\alpha_r,1)$, $\alpha_r>0$, $r=1,\dots,D$

- U_1 and U_2 are independent and they have the same Gamma distribution: $U_l \sim \text{Gamma}(\tau, 1)$, $\tau > 0$, $l = 1, 2$

- $\mathbf{Z}_1 \sim \text{Multinomial}(1, \mathbf{p})$

- $\mathbf{Z}_2 \sim \text{Multinomial}(1, \boldsymbol{\eta})$

where $\mathbf{p} = (p_{1\cdot}, p_{2\cdot}, \ldots, p_{D\cdot})^\intercal$ and $\boldsymbol{\eta} = (p_{\cdot 1}, p_{\cdot 2}, \ldots, p_{\cdot D})^\intercal$ are two vectors belonging to the D-part simplex. Supposing that $\mathbf{P}$ is the matrix with generic element $\mathbf{P}_{(i,j)} = p_{i,j} = P(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j)$, where $\mathbf{e}_i$ is the i-th element of the usual canonical basis (a vector with elements equal to 0 except for the i-th that is equal to 1), then it is possible to compute the vectors $\mathbf{p}$ and $\boldsymbol{\eta}$ as the row sum and the column sum of $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,D} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ p_{D,1} & p_{D,2} & \cdots & p_{D,D} \end{bmatrix} \begin{matrix} p_{1\cdot} \\ p_{2\cdot} \\ \vdots \\ p_{D\cdot} \end{matrix}$$
$$\begin{matrix} p_{\cdot 1} & p_{\cdot 2} & \cdots & p_{\cdot D} & 1 \end{matrix}$$

Since each $p_{i,j}$ represents a probability, the following constraints must hold:

$$\sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} = 1, \qquad p_{i,j} > 0. \tag{4.1}$$

It is now possible to construct a basis whose elements are:

$$Y_r = W_r + Z_{1,r}\, U_1 + Z_{2,r}\, U_2, \qquad r = 1, \ldots, D. \tag{4.2}$$

The basis $\mathbf{Y} = (Y_1, \ldots, Y_D)^\intercal$ has a distribution called Double Flexible Gamma (DFG), because it duplicates the constructing scheme of the FD basis in (3.14). This distribution is parametrized by $\boldsymbol{\alpha}, \tau$ and $\mathbf{P}$. By conditioning on $(\mathbf{Z}_1, \mathbf{Z}_2)^\intercal$ it is possible to derive a finite mixture representation. It follows that its density function can be expressed as:

$$f_{DFG}(\mathbf{y}; \boldsymbol{\alpha}, \tau, \mathbf{P}) = \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)), \tag{4.3}$$

where $f_G(\mathbf{y}; \boldsymbol{\alpha})$ is the density function of a random vector with independent gamma components with shape parameter α_h and common rate parameter equal to 1.

Proposition 18. *(Non-identifiability) Let $\boldsymbol{P}^*$ be a $D \times D$ matrix with generic element $p_{i,j}$ such that $0 \leq p_{i,j} < 1$ and $\sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} = 1$. Then the DFG model with $\boldsymbol{P} = \boldsymbol{P}^*$ is non-identifiable.*

Proof. Let us assume that the model is identifiable. If we suppose, for simplicity, that $D = 2$, then the density function of $\mathbf{Y}$ is:

$$\begin{aligned} f_{DFG}(\mathbf{y}; \boldsymbol{\alpha}, \tau, \mathbf{P}) &= \sum_{i=1}^{2}\sum_{j=1}^{2} p_{i,j} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)) \\ &= p_{1,1} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_1 + \mathbf{e}_1)) + p_{1,2} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_1 + \mathbf{e}_2)) + \\ &\qquad + p_{2,1} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_2 + \mathbf{e}_1)) + p_{2,2} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_2 + \mathbf{e}_2)) \\ &= p_{1,1} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_1 + \mathbf{e}_1)) + p_{2,2} f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_2 + \mathbf{e}_2)) + \\ &\qquad + [p_{1,2} + p_{2,1}]\, f_G(\mathbf{y}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_1 + \mathbf{e}_2)) \end{aligned}$$

The distribution is identified by $\boldsymbol{\alpha}$, τ, $p_{1,1}, p_{2,2}$ and by the sum $(p_{1,2} + p_{2,1})$. It is obvious that the same value of $(p_{1,2} + p_{2,1})$ can be obtained with different values of $(p_{1,2}, p_{2,1})^\intercal$. This issue can hold also for $D \neq 2$ and thus the model cannot be identifiable. □

In order to make the model identifiable, a possibility is to impose the symmetry of $\mathbf{P}$. Imposing $\mathbf{P}$ symmetric implies also $\mathbf{p} = \boldsymbol{\eta}$. Note that, even if $\mathbf{P}$ is assumed to be symmetric, it does not imply that $\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2$, since $p_{i,j} \neq p_{i\cdot} p_{\cdot j}$, in general.

4.1.2 Properties of $\mathbf{Y}$

This new basis has more interesting properties compared to the Flexible Gamma defined in 3.3. In order to explore them, it can be useful to remember that if $Y \sim \text{Gamma}(\alpha, \beta = 1)$ then $\mathbb{E}[Y] = \alpha$, $\mathbb{E}[Y^2] = \alpha(\alpha + 1)$ and $\text{Var}(Y) = \alpha$.

Proposition 19 (Moments of the basis)**.** *Let $\mathbf{Y} \sim DFG(\boldsymbol{\alpha}, \tau, \boldsymbol{P})$ where $\boldsymbol{P}$ is a symmetric matrix. Then the first two moments are:*

$$\mathbb{E}[Y_r] = \alpha_r + 2\tau\, p_{r\cdot}, \qquad r = 1, \ldots, D. \tag{4.4}$$

$$Var(Y_r) = \alpha_r + 2\tau p_{r\cdot} + 2\tau^2 \left(p_{r\cdot} - 2p_{r\cdot}^2 + p_{r,r}\right), \qquad r = 1, \ldots, D. \tag{4.5}$$

$$Cov\left(Y_r, Y_h\right) = 2\tau^2(p_{r,h} - 2p_{r\cdot}p_{h\cdot}), \qquad r, h = 1, \ldots, D,\ r \neq h. \tag{4.6}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[Y_r\right] &= \mathbb{E}\left[W_r\right] + \mathbb{E}\left[Z_{1,r}\, U_1\right] + \mathbb{E}\left[Z_{2,r}\, U_2\right] \\
&= \alpha_r + \tau\left(p_{r\cdot} + p_{\cdot r}\right) \\
&= \alpha_r + 2\tau\, p_{r\cdot}
\end{aligned}$$

where the last equality holds if $\mathbf{P}$ is symmetric.

$$\begin{aligned}
\text{Var}\left(Y_r\right) &= \text{Var}\left(W_r\right) + \text{Var}\left(Z_{1,r}\, U_1\right) + \text{Var}\left(Z_{2,r}\, U_2\right) + 2\text{Cov}\left(Z_{1,r}\, U_1, Z_{2,r}\, U_2\right) \\
&= \alpha_r + \mathbb{E}\left[Z_{1,r}\, U_1^2\right] - \left(\mathbb{E}\left[Z_{1,r}\, U_1\right]\right)^2 + \mathbb{E}\left[Z_{2,r}\, U_2^2\right] - \left(\mathbb{E}\left[Z_{2,r}\, U_2\right]\right)^2 + \\
&\quad + 2\left\{\mathbb{E}\left[Z_{1,r}Z_{2,r}U_1U_2\right] - \mathbb{E}\left[Z_{1,r}U_1\right]\mathbb{E}\left[Z_{2,r}U_2\right]\right\} \\
&= \alpha_r + \mathbb{E}\left[Z_{1,r}\right]\mathbb{E}\left[U_1^2\right] - \left(\mathbb{E}\left[Z_{1,r}\right]\mathbb{E}\left[U_1\right]\right)^2 + \mathbb{E}\left[Z_{2,r}\right]\mathbb{E}\left[U_2^2\right] - \left(\mathbb{E}\left[Z_{2,r}\right]\mathbb{E}\left[U_2\right]\right)^2 + \\
&\quad + 2\left\{\mathbb{E}\left[Z_{1,r}Z_{2,r}\right]\mathbb{E}\left[U_1\right]\mathbb{E}\left[U_2\right] - \mathbb{E}\left[Z_{1,r}\right]\mathbb{E}\left[Z_{2,r}\right]\mathbb{E}\left[U_1\right]\mathbb{E}\left[U_2\right]\right\} \\
&= \alpha_r + p_{r\cdot}\tau(\tau+1) - p_{r\cdot}^2\tau^2 + p_{\cdot r}\tau(\tau+1) - p_{\cdot r}^2\tau^2 + 2\left\{p_{r,r}\tau^2 - p_{r\cdot}p_{\cdot r}\tau^2\right\} \\
&= \underbrace{\alpha_r}_{\text{Original Gamma}} + \underbrace{p_{r\cdot}\tau + p_{r\cdot}(1-p_{r\cdot})\tau^2}_{1^{st}\text{ Gamma}} + \underbrace{p_{\cdot r}\tau + p_{\cdot r}(1-p_{\cdot r})\tau^2}_{2^{nd}\text{ Gamma}} + \underbrace{2\tau^2(p_{r,r} - p_{r\cdot}p_{\cdot r})}_{\text{Gammas' dependence}} \\
&= \alpha_r + 2\tau p_{r\cdot} + 2\tau^2 p_{r\cdot}(1-p_{r\cdot}) + 2\tau^2(p_{r,r} - p_{r\cdot}^2) \\
&= \alpha_r + 2\tau p_{r\cdot} + 2\tau^2\left(p_{r\cdot} - 2p_{r\cdot}^2 + p_{r,r}\right)
\end{aligned}$$

The above result has been obtained considering that $\mathbb{E}\left[Z_{1,r}Z_{2,r}\right] = 1 \cdot P(\mathbf{Z}_1 = \mathbf{e}_r, \mathbf{Z}_2 = \mathbf{e}_r) + 0 \cdot [1 - P(\mathbf{Z}_1 = \mathbf{e}_r, \mathbf{Z}_2 = \mathbf{e}_r)] = p_{r,r}$.

$$\begin{aligned}
\text{Cov}\left(Y_r, Y_h\right)_{r\neq h} &= \text{Cov}\left(W_r + Z_{1,r}\, U_1 + Z_{2,r}\, U_2, W_h + Z_{1,h}\, U_1 + Z_{2,h}\, U_2\right) \\
&= \text{Cov}\left(Z_{1,r}\, U_1 + Z_{2,r}\, U_2, Z_{1,h}\, U_1\right) + \text{Cov}\left(Z_{1,r}\, U_1 + Z_{2,r}\, U_2, Z_{2,h}\, U_2\right) \\
&= \text{Cov}\left(Z_{1,r}\, U_1, Z_{1,h}\, U_1\right) + \text{Cov}\left(Z_{2,r}\, U_2, Z_{1,h}\, U_1\right) + \\
&\quad + \text{Cov}\left(Z_{1,r}\, U_1, Z_{2,h}\, U_2\right) + \text{Cov}\left(Z_{2,r}\, U_2, Z_{2,h}\, U_2\right) \\
&= \mathbb{E}\left[Z_{1,r}Z_{1,h}U_1^2\right] - \mathbb{E}\left[Z_{1,r}U_1\right]\mathbb{E}\left[Z_{1,h}U_1\right] + \mathbb{E}\left[Z_{2,r}Z_{1,h}U_1U_2\right] + \\
&\quad - \mathbb{E}\left[Z_{2,r}U_2\right]\mathbb{E}\left[Z_{1,h}U_1\right] + \mathbb{E}\left[Z_{1,r}U_1Z_{2,h}U_2\right] + \\
&\quad - \mathbb{E}\left[Z_{1,r}U_1\right]\mathbb{E}\left[Z_{2,h}U_2\right] + \mathbb{E}\left[Z_{2,r}U_2Z_{2,h}U_2\right] - \mathbb{E}\left[Z_{2,r}U_2\right]\mathbb{E}\left[Z_{2,h}U_2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \cancel{\left[0 \cdot \tau(\tau+1) - p_{r\cdot}\tau \cdot p_{h\cdot}\tau\right]} + [p_{h,r}\tau \cdot \tau - p_{\cdot r}\tau \cdot p_{h\cdot}\tau] + \\
&\qquad + [p_{r,h}\tau \cdot \tau - p_{r\cdot}\tau \cdot p_{\cdot h}\tau] + \cancel{\left[0 \cdot \tau(\tau+1) - p_{\cdot r}\tau \cdot p_{\cdot h}\tau\right]} \\
&= \underbrace{-\tau^2 p_{r\cdot}p_{h\cdot}}_{1^{st}\text{Gamma}} \underbrace{+\tau^2\left(p_{h,r} - p_{h\cdot}p_{\cdot r}\right) + \tau^2\left(p_{r,h} - p_{r\cdot}p_{\cdot h}\right)}_{\text{Gammas' dependence}} \underbrace{-\tau^2 p_{\cdot r}p_{\cdot h}}_{2^{nd}\text{Gamma}} \\
&= -\tau^2 p_{r\cdot}p_{h\cdot} + 2\tau^2(p_{h,r} - p_{h\cdot}p_{r\cdot}) \\
&= 2\tau^2(p_{r,h} - 2p_{r\cdot}p_{h\cdot}) \qquad (4.7)
\end{aligned}
$$

□

Proposition 20. *(Positivity of Covariance) Let* $\boldsymbol{Y} \sim DFG(\boldsymbol{\alpha}, \tau, \boldsymbol{P})$ *and* $\boldsymbol{P}$ *be a* $D \times D$ *symmetric matrix; then the covariance between two components* Y_r *and* Y_h $(r \neq h)$ *can assume positive values.*

Proof. Remember from (4.6) that $\text{Cov}\,(Y_r, Y_h) = 2\tau^2(p_{r,h} - 2p_{r\cdot}p_{h\cdot})$. Then:

$$
\begin{aligned}
\text{Cov}\,(Y_r, Y_h) \geq 0 \quad &\iff \quad 2\tau^2(p_{r,h} - 2p_{r\cdot}p_{h\cdot}) \geq 0 \\
&\iff \quad 4\tau^2 p_{r\cdot}p_{h\cdot} \leq 2\tau^2 p_{r,h} \\
&\iff \quad \frac{p_{r\cdot}p_{\cdot h}}{p_{r,h}} \leq \frac{2\tau^2}{(2\tau)^2} \\
&\iff \quad \boxed{\frac{p_{r\cdot}p_{\cdot h}}{p_{r,h}} \leq \frac{1}{2}}
\end{aligned}
$$

□

With similar steps, it is easy to derive the more general condition:

$$
\text{Cov}\,(Y_r, Y_h) \geq c \quad \iff \quad \boxed{2p_{r\cdot}p_{h\cdot} - p_{r,h} \leq -\frac{c}{2\tau^2}}
$$

Proposition 21. *(Covariance's infimum and supremum) Let* $\boldsymbol{P}$ *be a symmetric matrix. Then, given* $\boldsymbol{\alpha}$ *and* τ*, the infimum over* $\boldsymbol{P}$ *of the covariance is* $-\tau^2$ *and the supremum is* $\frac{\tau^2}{4}$.

Proof. From equation (4.6) it is easy to see that $\text{Cov}\,(Y_r, Y_h)$ is minimum when $p_{r,h} = 0$ and $(p_{r\cdot}p_{h\cdot})$ is maximum, under the obvious constraint $p_{r\cdot} + p_{h\cdot} \leq 1$. This means that minimizing the covariance is equivalent to maximize the function $f(x, y) = x \cdot y$ with the constraint $x + y \leq 1$. In order to solve this maximization problem it is necessary to define the Lagrangian function: $\mathcal{L}(x, y, \lambda) = xy - \lambda(x + y - 1)$. Then the optimality conditions are:

$$
\begin{cases}
\dfrac{\partial \mathcal{L}(x,y,\lambda)}{\partial x} = 0 \\
\dfrac{\partial \mathcal{L}(x,y,\lambda)}{\partial y} = 0 \\
\dfrac{\partial \mathcal{L}(x,y,\lambda)}{\partial \lambda} \geq 0 \quad \wedge \quad \lambda \dfrac{\partial \mathcal{L}(x,y,\lambda)}{\partial \lambda} = 0 \\
\lambda \geq 0
\end{cases}
$$

$$
\implies
\begin{cases}
y - \lambda = 0 \\
x - \lambda = 0 \\
-(x + y - 1) \geq 0 \quad \wedge \quad -\lambda(x + y - 1) = 0 \\
\lambda \geq 0
\end{cases}
$$

$$
\implies
\begin{cases}
y = \lambda \\
x = \lambda \\
-2\lambda + 1 \geq 0 \quad \wedge \quad -\lambda(2\lambda - 1) = 0 \\
\lambda \geq 0
\end{cases}
$$

Considering the two cases $\lambda = 0$ and $\lambda \neq 0$ separately:

- $\lambda = 0 \implies \begin{cases} x = 0 \\ y = 0 \end{cases} \implies f(0,0) = 0$

- $\lambda \neq 0 \implies -\lambda(2\lambda - 1) = 0 \implies \lambda = \dfrac{1}{2}$
  $\implies \begin{cases} x = \frac{1}{2} \\ y = \frac{1}{2} \end{cases} \implies f\left(\dfrac{1}{2}, \dfrac{1}{2}\right) = \dfrac{1}{4}$

It follows that the maximizer of $(p_{r\cdot}p_{h\cdot})$ under the constraint $p_{r\cdot} + p_{h\cdot} \leq 1$ is $(p_{r\cdot}, p_{h\cdot})^\intercal = \left(\frac{1}{2}, \frac{1}{2}\right)^\intercal$. It is possible to find this solution also studying the contour plot of the function $g(x,y) = x \cdot y$ in the region $\{0 \leq x < 1; 0 \leq y \leq 1 - x\}$ (Figure 4.1).

It follows that $(p_{r\cdot}p_{h\cdot})$ is maximum when $p_{r\cdot} = p_{h\cdot} = 0.5$. In conclusion, having $p_{r,h} = 0$ and $p_{r\cdot} = p_{h\cdot} = 0.5$ implies that:

$$
\begin{aligned}
\min\left[\mathrm{Cov}\left(Y_r, Y_h\right)\right] &= \min\left[-p_{r\cdot}p_{h\cdot}(2\tau)^2 + 0\right] \\
&= -0.5 \cdot 0.5(2\tau)^2 \\
&= -\frac{(2\tau)^2}{4} = -\tau^2
\end{aligned}
$$

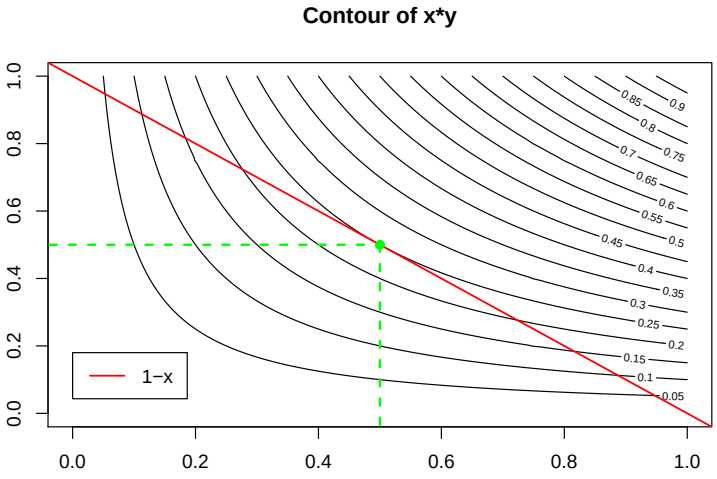

Fig. 4.1: Contour plot of $g(x, y) = x \cdot y$.

It is remarkable to note that, due to the symmetry of **P**, the conditions $\begin{cases} p_{r,h} = p_{h,r} = 0 \\ p_{r\cdot} = p_{h\cdot} = 0.5 \end{cases}$ imply that **P** has only two elements different from 0: $p_{r,r} = p_{h,h} = 0.5$.

Inspecting (4.6), it is possible to note that the highest covariance is obtained when $(p_{r\cdot}p_{h\cdot})$ reaches its lowest value and $p_{r,h}$ reaches its highest one, under the following constraints:

$$\begin{cases} p_{r\cdot} \geq p_{r,h} \\ p_{h\cdot} \geq p_{r,h} \\ p_{r\cdot} + p_{h\cdot} \leq 1 \end{cases}$$

Assuming that $p_{r,h}$ is fixed, the product $(p_{r\cdot}p_{h\cdot})$ is minimum when $p_{r\cdot}$ and $p_{h\cdot}$ are minimum. Then $p_{r\cdot} = p_{h\cdot} = p_{r,h} = p_{h,r}$ (and this implies that $p_{r,r} = p_{h,h} = 0$). Now it is easy to find the maxima, since the covariance assumes the following form:

$$\begin{aligned} \text{Cov}\,(Y_r, Y_h) &= -p_{r\cdot}p_{h\cdot}(2\tau)^2 + 2\tau^2 p_{r,h} \\ &= -4p_{r,h}p_{r,h}\tau^2 + 2\tau^2 p_{r,h} \\ &= -4p_{r,h}^2\tau^2 + 2\tau^2 p_{r,h} \end{aligned}$$

Applying the usual maximization approach based on derivative, one obtains:

$$\begin{aligned}\frac{\partial}{\partial p_{r,h}}\left\{-4p_{r,h}^2\tau^2+2\tau^2 p_{r,h}\right\} &= 0\\ -2p_{r,h}4\tau^2+2\tau^2 &= 0\\ p_{r,h}^* &= \tfrac{1}{4}\end{aligned} \tag{4.8}$$

In order to check if the stationary point (4.8) is a maximizer, the second-order derivative is computed:

$$\frac{\partial^2}{\partial p_{r,h}^2}\left\{-2p_{r,h}4\tau^2+2\tau^2\right\} = -8\tau^2 < 0. \tag{4.9}$$

It follows that $p_{r,h}^* = p_{r\cdot}^* = p_{h\cdot}^* = \dfrac{1}{4}$ is a maxima.

$$\begin{aligned}\max\left[\text{Cov}\left(Y_r, Y_h\right)\right] &= -4p_{r\cdot}^* p_{h\cdot}^* \tau^2 + 2\tau^2 p_{r,h}^*\\ &= -\left(\frac{1}{4}\right)^2 4\tau^2 + 2\tau^2\frac{1}{4}\\ &= -\frac{\tau^2}{4}+2\frac{\tau^2}{4}\\ &= \frac{\tau^2}{4}\end{aligned}$$

□

In conclusion, the parameter configuration of **P** leading to the minimum value of $\text{Cov}\left(Y_r, Y_h\right)$ given the values of $\boldsymbol{\alpha}$ and τ is:

$$\begin{cases} p_{r,h} = p_{h,r} = 0\\ p_{r\cdot} = p_{h\cdot} = 0.5\\ p_{r,r} = p_{h,h} = 0.5\end{cases} \tag{4.10}$$

and the one that leads to the maximum value of $\text{Cov}(Y_r, Y_h)$ is:

$$\begin{cases} p_{r,h} = p_{h,r} = \frac{1}{4} \\ p_{r\cdot} = p_{h\cdot} = \frac{1}{4} \\ p_{r,r} = p_{h,h} = 0 \end{cases} \tag{4.11}$$

The following definition introduces a matrix that will be used in some of the following properties:

Definition 21. *Let $\boldsymbol{a} = (a_0, a_1, \ldots, a_{C-1}, a_C)^\intercal$ be a vector of non negative integers such that $0 = a_0 < a_1 < \cdots < a_{C-1} < a_C = D$, a **double collapser** matrix $\boldsymbol{M_a}$ is the unique $C \times D$ amalgamation matrix such that in the r-th row ($r = 1, \ldots, C$) there are $(a_r - a_{r-1})$ 1's and they are in position $(a_{r-1} + 1, \ldots, a_r)$.*

Proposition 22. *Let $\boldsymbol{M_a}$ be a double collapser matrix. If $\boldsymbol{B} = \boldsymbol{M_a} \cdot \boldsymbol{D} \cdot \boldsymbol{M_a}^\intercal$, then $b_{i,j} = \sum_{r=a_{i-1}+1}^{a_i} \sum_{l=a_{j-1}+1}^{a_j} d_{r,l}$, where $b_{i,j}$ and $d_{r,l}$ are the generic element of $\boldsymbol{B}$ and $\boldsymbol{D}$, respectively. In particular, this means that $\boldsymbol{B}$ is equal to the matrix $\boldsymbol{D}$ with blocks of rows and columns summed over.*

Example 8. *Let $\boldsymbol{D}$ be the following 6×6 matrix:*

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 2 & 1 & 3 & 7 \\ 0 & 5 & 3 & 1 & 0 & 1 \\ 4 & 1 & 1 & 3 & 4 & 0 \\ 3 & 1 & 0 & 1 & 3 & 2 \\ 0 & 2 & 0 & 2 & 0 & 3 \\ 1 & 0 & 2 & 1 & 0 & 1 \end{bmatrix}.$$

*Supposing that a statistician wants to obtain a new matrix with rows **and** columns 2 and 3 and rows/columns 5 and 6 summed up. This means that the desired matrix is formed by the sum of the elements belonging to each block of the following matrix:*

$$\left[\begin{array}{c|cc|c|cc} 1 & 0 & 2 & 1 & 3 & 7 \\ \hline 0 & 5 & 3 & 1 & 0 & 1 \\ 4 & 1 & 1 & 3 & 4 & 0 \\ \hline 3 & 1 & 0 & 1 & 3 & 2 \\ \hline 0 & 2 & 0 & 2 & 0 & 3 \\ 1 & 0 & 2 & 1 & 0 & 1 \end{array}\right].$$

In order to obtain this matrix, he defines the vector $\boldsymbol{a} = (0, 1, 3, 4, 6)^\intercal$ *and the corresponding double collapser matrix:*

$$\boldsymbol{M_a} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

Then, computing the matrix multiplication $\mathbf{M_a} \cdot \mathbf{D} \cdot \mathbf{M_a^\intercal}$, he obtains the matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 1 & 10 \\ 4 & 10 & 4 & 5 \\ 3 & 1 & 1 & 5 \\ 1 & 4 & 3 & 4 \end{bmatrix},$$

that is the desired matrix.

Proposition 23 (Closure under Amalgamation)**.** *Let* $\boldsymbol{Y} \sim DFG(\boldsymbol{\alpha}, \tau, \boldsymbol{P})$ *and let* C *be a positive integer such that* $C \leq D$*. Then,* $\boldsymbol{Y}^+ = \left(Y_1^+, \ldots, Y_C^+\right)^\intercal$*, namely the amalgamation induced by the* C*-dimensional vector* $\boldsymbol{a}$*, follows a* $DFG(\boldsymbol{\alpha}^+, \tau, \boldsymbol{P}^+)$ *distribution, where* $\boldsymbol{\alpha}^+ = \left(\sum_{i=a_0+1}^{a_1} \alpha_i, \ldots, \sum_{i=a_{C-1}+1}^{a_C} \alpha_i\right)^\intercal$*,* $\boldsymbol{P}^+$ *is a* $C \times C$ *matrix such that:* $\boldsymbol{P}^+ = \boldsymbol{M_a} \cdot \boldsymbol{P} \cdot \boldsymbol{M_a^\intercal}$ *and* $\boldsymbol{M_a}$ *is a double collapser matrix.*

Proof. Given the basis defined in (4.2), elements of $\mathbf{Y}^+ = (Y_1^+, Y_2^+, \ldots, Y_C^+)^\intercal = \left(\sum_{i=a_0+1}^{a_1} Y_i, \sum_{i=a_1+1}^{a_2} Y_i, \ldots, \sum_{i=a_{C-1}+1}^{a_C} Y_i\right)^\intercal$ can be defined as:

$$\begin{aligned} Y_h^+ &= \sum_{i=a_{h-1}+1}^{a_h} Y_i = \sum_{i=a_{h-1}+1}^{a_h} W_i + U_1 \left(\sum_{i=a_{h-1}+1}^{a_h} Z_{1,i}\right) + U_2 \left(\sum_{i=a_{h-1}+1}^{a_h} Z_{2,i}\right) \\ &= \sum_{i=a_{h-1}+1}^{a_h} W_i + U_1 \cdot Z_{1,h}^* + U_2 \cdot Z_{2,h}^* \end{aligned}$$

Thanks to well-known properties of Gamma and Multinomial distributions, the following results are immediately obtained:

- $\left(\sum_{i=a_{h-1}+1}^{a_h} W_i\right) \sim \text{Gamma}\left(\sum_{i=a_{h-1}+1}^{a_h} \alpha_i, 1\right)$

- $\mathbf{Z}_1^* = \left(Z_{1,1}^*, \ldots, Z_{1,C}^*\right)^\intercal \sim \text{Multinomial}(1, \mathbf{p}^+)$
- $\mathbf{Z}_2^* = \left(Z_{2,1}^*, \ldots, Z_{2,C}^*\right)^\intercal \sim \text{Multinomial}(1, \mathbf{p}^+)$

where $\mathbf{p}^+ = (p_1^+, \ldots, p_C^+)^\intercal$ and $p_h^+ = \left(\sum_{i=a_{h-1}+1}^{a_h} p_{i\cdot}\right)$.

In order to show that $\mathbf{P}^+ = \mathbf{M_a} \cdot \mathbf{P} \cdot \mathbf{M_a}^\intercal$ one can rely on the following algebra:

$$
\begin{aligned}
P\left(\mathbf{Z}_1^* = \mathbf{e}_i, \mathbf{Z}_2^* = \mathbf{e}_j\right) &= P\left(\mathbf{Z}_1^* = \mathbf{e}_i | \mathbf{Z}_2^* = \mathbf{e}_j\right) P\left(\mathbf{Z}_2^* = \mathbf{e}_j\right) \\
&= P\left(\mathbf{Z}_1 \in \{\mathbf{e}_{1+a_{i-1}}, \ldots, \mathbf{e}_{a_i}\} | \mathbf{Z}_2 \in \{\mathbf{e}_{1+a_{j-1}}, \ldots, \mathbf{e}_{a_j}\}\right) \cdot \\
&\qquad \cdot P\left(\mathbf{Z}_2 \in \{\mathbf{e}_{a_{j-1}}, \ldots, \mathbf{e}_{a_j}\}\right) \\
&= \frac{\sum_{r=1+a_{i-1}}^{a_i} \sum_{l=1+a_{j-1}}^{a_j} p_{r,l}}{\sum_{l=1+a_{j-1}}^{a_j} p_{j\cdot}} \sum_{l=1+a_{j-1}}^{a_j} p_{j\cdot} \\
&= \sum_{r=1+a_{i-1}}^{a_i} \sum_{l=1+a_{j-1}}^{a_j} p_{r,l} = \left(\mathbf{M_a} \cdot \mathbf{P} \cdot \mathbf{M_a}^\intercal\right)_{i,j}
\end{aligned}
$$

With these random elements it is possible to define the new basis $\mathbf{Y}^+ = (Y_1^+, \ldots, Y_C^+)^\intercal$ as in (4.2). Then, $\mathbf{Y}^+ \sim DFG(\boldsymbol{\alpha}^*, \tau, \mathbf{P}^*)$. □

Proposition 24 (Compositional Invariance)**.** *The basis defined in (4.2) is compositionally invariant (this means that* $\boldsymbol{X} = \boldsymbol{Y}/Y^+ \perp\!\!\!\perp Y^+$*).*

Proof. Thanks to the mixture structure of the DFG distribution, it is easy to show that $\mathbf{X}|(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j) \sim \text{Dir}(\boldsymbol{\alpha} + \tau\left(\mathbf{e}_i + \mathbf{e}_j\right))$. Due to the compositional invariance property of the Dirichlet distribution, $\mathbf{X}|(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j)$ is independent of $Y^+|(\mathbf{Z}_1 = \mathbf{z}_1, \mathbf{Z}_2 = \mathbf{z}_2) \sim \text{Gamma}(\alpha^+ + 2\tau)$. Then it is possible to write:

$$
\begin{aligned}
F_{\mathbf{X},Y^+}(\mathbf{x}, y^+) &= \sum_{i=1}^{D}\sum_{j=1}^{D} F_{\mathbf{X},Y^+|(\mathbf{Z}_1=\mathbf{e}_i,\mathbf{Z}_2=\mathbf{e}_j)}(\mathbf{x}, y^+) \cdot P(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j) \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} F_{\mathbf{X},Y^+|(\mathbf{Z}_1=\mathbf{e}_i,\mathbf{Z}_2=\mathbf{e}_j)}(\mathbf{x}, y^+) \cdot p_{i,j} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} F_{\mathbf{X}|(\mathbf{Z}_1=\mathbf{e}_i,\mathbf{Z}_2=\mathbf{e}_j)}(\mathbf{x}) \cdot F_{Y^+|(\mathbf{Z}_1=\mathbf{e}_i,\mathbf{Z}_2=\mathbf{e}_j)}(y^+)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} F_{\mathbf{X}|(\mathbf{Z}_1=\mathbf{e}_i,\mathbf{Z}_2=\mathbf{e}_j)}(\mathbf{x}) \cdot F_{Y^+}(y^+) \\
&= \text{Gamma}(y^+;\alpha^+ + 2\tau, 1) \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \text{Dir}(\mathbf{x};\boldsymbol{\alpha} + \tau(\mathbf{e}_i + \mathbf{e}_j))
\end{aligned} \tag{4.12}
$$

From this result it is immediate to obtain also the marginal distribution of Y^+ and $\mathbf{X}$ (the latter can be saw as a finite mixture of a Dirichlet components with the same parametric configurations of the components in (4.3)). □

4.1.3 Correlation

Let $\mathbf{Y} \sim DFG(\boldsymbol{\alpha}, \tau, \mathbf{P})$. Then the correlation coefficient of Y_r and Y_h $(r \neq h)$ is:

$$
\begin{aligned}
\rho_{Y_r,Y_h} &= \frac{\text{Cov}(Y_r, Y_h)}{\sqrt{\text{Var}(Y_r)\,\text{Var}(Y_h)}} \\
&= \frac{-2\tau^2(2p_{r\cdot}p_{h\cdot} - p_{r,h})}{\sqrt{(\alpha_r + 2\tau p_{r\cdot} + 2\tau^2(p_{r\cdot} - 2p_{r\cdot}^2 + p_{r,r})) \cdot (\alpha_h + 2\tau p_{h\cdot} + 2\tau^2(p_{h\cdot} - 2p_{h\cdot}^2 + p_{h,h}))}}
\end{aligned} \tag{4.13}
$$

The DFG model allows for even strong positive linear relationships (that means high positive correlation coefficients) between its components.

Example 9. *A way to obtain high values of ρ_{Y_r,Y_h} given $\boldsymbol{\alpha}$ and τ,is to maximize Cov (Y_r, Y_h) and setting the remaining parameters in order to minimize both Var (Y_r) and Var (Y_h). From Proposition (21) it follows that Cov (Y_r, Y_h) is maximized if $p_{r\cdot} = p_{h\cdot} = p_{r,h} = p_{h,r} = \frac{1}{4}$. This choice of parameters implies that:*

$$
\begin{aligned}
\textit{Var}(Y_r) &= \alpha_r + 2\tau p_{r\cdot} + 2\tau^2 p_{r\cdot}(1 - p_{r\cdot}) + 2\tau^2\left(p_{r,r} - p_{r\cdot}^2\right) \\
&= \alpha_r + 2\tau\frac{1}{4} + 2\tau^2\frac{1}{4}\cdot\frac{3}{4} + 2\tau^2\left(p_{r,r} - \frac{1}{16}\right) \\
&= \alpha_r + \frac{1}{2}\tau + \tau^2\left(\frac{3}{8} + 2p_{r,r} - \frac{1}{8}\right) \\
&= \alpha_r + \frac{1}{2}\tau + \tau^2\left(2p_{r,r} + \frac{1}{4}\right).
\end{aligned}
$$

With the same arguments one can show that Var $(Y_h) = \alpha_h + \frac{1}{2}\tau + \tau^2\left(2p_{h,h} + \frac{1}{4}\right)$. These variances are minimized if $p_{r,r} = 0$ and $p_{h,h} = 0$. Then:

$$
\begin{aligned}
\rho_{Y_r,Y_h} &= \frac{\tau^2}{4\cdot\sqrt{\left(\alpha_r+\frac{\tau}{2}+\frac{\tau^2}{4}\right)\cdot\left(\alpha_h+\frac{\tau}{2}+\frac{\tau^2}{4}\right)}} \\
&= \frac{\tau^2}{4\cdot\sqrt{\tau^2\left(\frac{\alpha_r}{\tau^2}+\frac{1}{2\tau}+\frac{1}{4}\right)\cdot\tau^2\left(\frac{\alpha_h}{\tau^2}+\frac{1}{2\tau}+\frac{1}{4}\right)}} \\
&= \frac{1}{4\cdot\sqrt{\left(\frac{\alpha_r}{\tau^2}+\frac{1}{2\tau}+\frac{1}{4}\right)\cdot\left(\frac{\alpha_h}{\tau^2}+\frac{1}{2\tau}+\frac{1}{4}\right)}} \underset{\tau\to+\infty}{\longrightarrow} 1
\end{aligned}
$$

Suppose, without loss of generality, that $D=3$; in order to fulfill the above conditions, the matrix $\boldsymbol{P}$ can be set equal to $\boldsymbol{P}=\begin{bmatrix} 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & 0 & 0 \end{bmatrix}$. *Figure 4.2 shows how the correlation coefficient among Y_1 and Y_3 changes in function of $\boldsymbol{\alpha}$ and τ.*

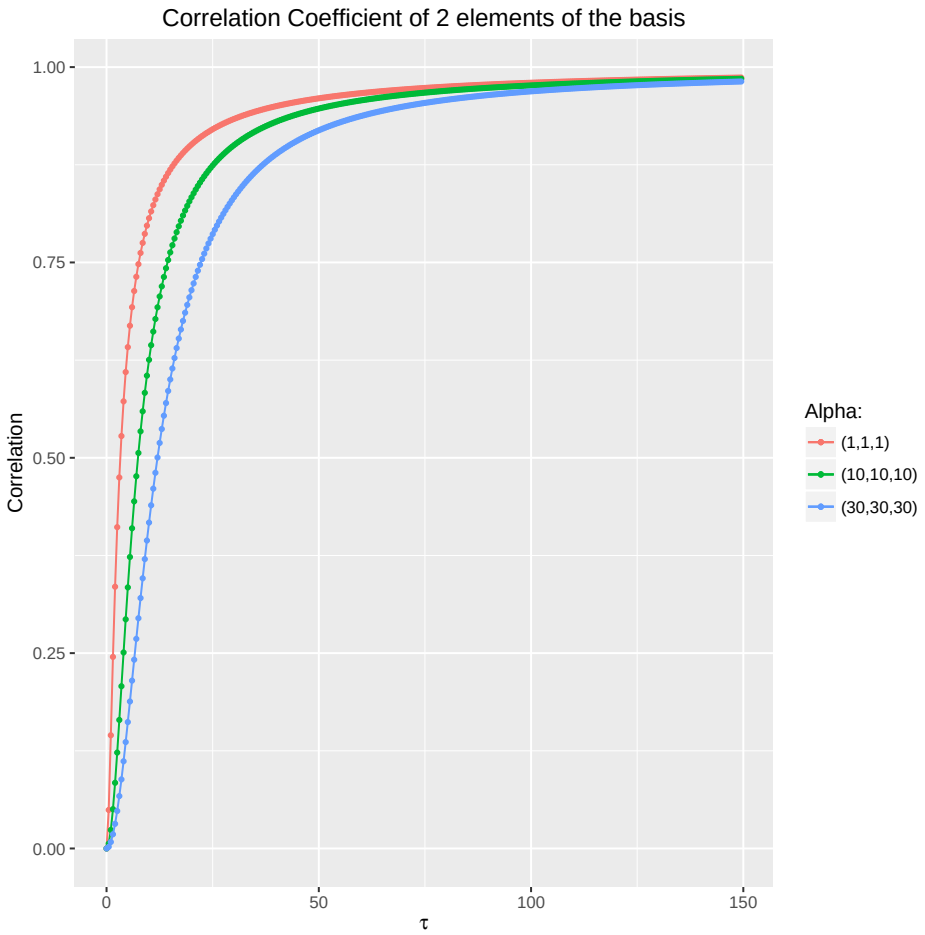

Fig. 4.2: Correlation between Y_1 and Y_3.

Of course, the correlation coefficient among Y_r and Y_h can assume also extreme negative values. An example consists in using the configuration of parameters leading to the minimum covariance (Equation 4.10) $\begin{cases} p_{r,h} = p_{h,r} = 0 \\ p_{r\cdot} = p_{h\cdot} = p_{r,r} = p_{h,h} = 0.5 \end{cases}$, the following computations can be derived:

$$\begin{aligned} \rho_{Y_r,Y_h} &= \frac{-\tau^2}{\sqrt{(\alpha_r + \tau + \tau^2) \cdot (\alpha_h + \tau + \tau^2)}} \\ &= \frac{-\tau^2}{\sqrt{\tau^2\left(\frac{\alpha_r}{\tau^2} + \frac{1}{\tau} + 1\right) \cdot \tau^2\left(\frac{\alpha_h}{\tau^2} + \frac{1}{\tau} + 1\right)}} \\ &= \frac{-1}{\sqrt{\left(\frac{\alpha_r}{\tau^2} + \frac{1}{\tau} + 1\right)\left(\frac{\alpha_h}{\tau^2} + \frac{1}{\tau} + 1\right)}} \underset{\tau \to +\infty}{\longrightarrow} -1. \end{aligned}$$

4.2 The DFD model

Let $\mathbf{Y} \sim DFG(\boldsymbol{\alpha}, \tau, \mathbf{P})$, then the composition obtained closing $\mathbf{Y}$, $\mathbf{X} = \mathcal{C}(\mathbf{Y})$, follows a new distribution, called *Double Flexible Dirichlet* and denoted by $DFD(\boldsymbol{\alpha}, \tau, \mathbf{P})$. Due to the DFG's generation mechanism (4.2), the parametric space of this random vector is:

$$\Theta_{\text{DFD}} = \left\{ (\boldsymbol{\alpha}, \tau, \mathbf{P}) : \boldsymbol{\alpha} \in \mathbb{R}_+^D, \tau \in \mathbb{R}^+, 0 \leq p_{i,j} < 1,\ i,j = 1, \ldots, D, \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} = 1 \right\}.$$

Thanks to the well-know relationship between Gamma and Dirichlet random variables, by conditioning on $\mathbf{Z}_1$ and $\mathbf{Z}_2$ each component follows a particular Dirichlet distribution. This allows to derive a mixture representation of the DFD model, already noted in (4.12):

$$DFD(\mathbf{x}; \boldsymbol{\alpha}, \tau, \mathbf{P}) = \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \mathcal{D}(\mathbf{x}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)),$$

where $\mathbf{x} \in \mathcal{S}^D$ and $\mathcal{D}(\mathbf{x}; \boldsymbol{\alpha})$ denotes the distribution function of a Dirichlet random vector. Given this representation, it is easy to write the density function characterizing the DFD:

$$
\begin{aligned}
f_{DFD}(\mathbf{x};\boldsymbol{\alpha},\tau,\mathbf{P}) &= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} f_{\mathcal{D}}(\mathbf{x};\boldsymbol{\alpha}+\tau\left(\mathbf{e}_i+\mathbf{e}_j\right)) \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\Gamma(\alpha^{+}+2\tau)}{\prod_{r=1}^{D}\Gamma(\alpha_r+\tau\left(e_{i_r}+e_{j_r}\right))} \prod_{r=1}^{D} x_r^{\alpha_r+\tau\left(e_{i_r}+e_{j_r}\right)-1} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\Gamma(\alpha^{+}+2\tau)}{\prod_{r=1}^{D}\Gamma(\alpha_r+\tau\left(e_{i_r}+e_{j_r}\right))} \prod_{r\neq i,j}\left(x_r^{\alpha_r-1}\right) x_i^{\alpha_i+\tau-1} x_j^{\alpha_j+\tau-1} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\Gamma(\alpha^{+}+2\tau)}{\prod_{r=1}^{D}\Gamma(\alpha_r+\tau\left(e_{i_r}+e_{j_r}\right))} \prod_{r=1}^{D}\left(x_r^{\alpha_r-1}\right)(x_i x_j)^{\tau} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\Gamma(\alpha^{+}+2\tau)\left(\prod_{r=1}^{D} x_r^{\alpha_r-1}\right)(x_i x_j)^{\tau}}{\Gamma(\alpha_i+\tau)\Gamma(\alpha_j+\tau)\prod_{r\neq i,j}\Gamma(\alpha_r)} + \\
&\qquad + \sum_{i=1}^{D} p_{i,i} \frac{\Gamma(\alpha^{+}+2\tau)\left(\prod_{r=1}^{D} x_r^{\alpha_r-1}\right)}{\Gamma(\alpha_i+2\tau)\prod_{r\neq i}\Gamma(\alpha_r)} x_i^{2\tau} \\
&= \frac{\Gamma(\alpha^{+}+2\tau)}{\prod_{r=1}^{D}\Gamma(\alpha_r)}\left(\prod_{r=1}^{D} x_r^{\alpha_r-1}\right)\left[\sum_{\substack{i=1\\ i\neq j}}^{D}\sum_{j=1}^{D} p_{i,j}\frac{\Gamma(\alpha_i)\Gamma(\alpha_j)(x_i x_j)^{\tau}}{\Gamma(\alpha_i+\tau)\Gamma(\alpha_j+\tau)} + \sum_{i=1}^{D} p_{i,i}\frac{\Gamma(\alpha_i)x_i^{2\tau}}{\Gamma(\alpha_i+2\tau)}\right]
\end{aligned}
$$

Proposition 25 (Closure under Amalgamation)**.** *Let* $\boldsymbol{X}=(X_1,\ldots,X_D)^{\intercal}\sim DFD(\boldsymbol{\alpha},\tau,\boldsymbol{P})$*,* $\boldsymbol{a}=(a_0,a_1,\ldots,a_{C-1},a_C)^{\intercal}$ *be a vector of non negative integers such that* $0=a_0<a_1<\cdots<a_{C-1}<a_C=D$ *and* $\boldsymbol{X}^{+}=\left(X_1^{+},\ldots,X_C^{+}\right)^{\intercal}=\left(\sum_{j=a_0+1}^{a_1}X_j,\ldots,\sum_{j=a_{C-1}+1}^{a_C}X_j\right)^{\intercal}$*. Then* $\boldsymbol{X}^{+}\sim DFD(\boldsymbol{\alpha}^{+},\tau,\boldsymbol{M_a}\cdot\boldsymbol{P}\cdot\boldsymbol{M_a}^{\intercal})$*, where* $\boldsymbol{\alpha}^{+}=\left(\sum_{j=a_0+1}^{a_1}\alpha_j,\ldots,\sum_{j=a_{C-1}+1}^{a_C}\alpha_j\right)^{\intercal}$ *and* $\boldsymbol{M_a}$ *is the double collapser matrix associated to* $\boldsymbol{a}$*.*

Proof. $\left(X_1^{+},\ldots,X_C^{+}\right)^{\intercal}=\dfrac{\left(Y_1^{+},\ldots,Y_C^{+}\right)^{\intercal}}{Y^{+}}$. Thanks to Proposition 23, it is possible to show that the numerator is DFG-distributed with parameters $\boldsymbol{\alpha}^{+}$, τ and $\mathbf{M_a}\cdot\mathbf{P}\cdot\mathbf{M_a}^{\intercal}$. This means that $\mathbf{X}^{+}$ is the composition obtained closing a DFG basis and then:

$$\mathbf{X}^{+}\sim \mathrm{DFD}(\boldsymbol{\alpha}^{+},\tau,\mathbf{M_a}\cdot\mathbf{P}\cdot\mathbf{M_a}^{\intercal}).$$

□

Proposition 26 (Marginals). *As a consequence of Proposition 25, by setting* $\boldsymbol{a} = (0, 1, \ldots, k-1, k, D)^\intercal$*, for any* $1 \leq k < D$*, it is possible to get the distribution of* $\left(\boldsymbol{X}_1, 1 - X_1^+\right)^\intercal$*, where* $\boldsymbol{X}_1 = (X_1, \ldots, X_k)^\intercal$*:*

$$\left(\boldsymbol{X}_1, 1 - X_1^+\right)^\intercal \sim DFD\left(\boldsymbol{\alpha}_1, \tau, \boldsymbol{M_a} \cdot \boldsymbol{P} \cdot \boldsymbol{M_a}^\intercal\right), \tag{4.14}$$

where $\boldsymbol{\alpha}_1 = \left(\alpha_1, \alpha_2, \ldots, \alpha_k, \sum_{h=k+1}^{D} \alpha_h\right)^\intercal$*. In particular, the one-dimensional marginals can be expressed in the following finite mixture structure:*

$$\begin{aligned} X_r \sim p_{r,r} Beta\left(\alpha_r + 2\tau, \alpha^+ - \alpha_r\right) + 2\left(\sum_{i \neq r} p_{r,i}\right) Beta\left(\alpha_r + \tau, \alpha^+ - \alpha_r + \tau\right) + \\ + \left(\sum_{i \neq r}\sum_{j \neq r} p_{i,j}\right) Beta\left(\alpha_k, \alpha^+ - \alpha_r + 2\tau\right) \end{aligned} \tag{4.15}$$

Note that if $\mathbf{P}$ is a diagonal matrix such as:

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & 0 & \ldots & 0 \\ 0 & p_{2,2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{D,D} \end{bmatrix},$$

the DFD reduces to a FD with $\mathbf{p}_{FD} = \text{diag}(\mathbf{P}_{DFD})$ and $\tau_{\text{FD}} = 2\tau_{\text{DFD}}$. Allowing for this scenario an identification issue arises: for example, the following two parametric configurations provide the same distribution (the Dirichlet one):

$$\begin{cases} \tau = 0.5 \\ p_{i,j} = \begin{cases} \dfrac{\alpha_i}{\alpha^+}, & i = j \\ 0, & i \neq j \end{cases} \end{cases} \qquad \begin{cases} \tau = 1 \\ p_{i,j} = \begin{cases} \dfrac{\alpha_i \alpha_j}{\alpha^+(\alpha^+ + 1)}, & i \neq j \\ \dfrac{\alpha_i(\alpha_i + 1)}{\alpha^+(\alpha^+ + 1)}, & i = j \end{cases} \end{cases}$$

Theorem 1 (Identifiability of the DFD model). *Let* $\boldsymbol{X} \sim DFD(\boldsymbol{\theta})$*,* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \tau, \boldsymbol{P})^\intercal$ *and* $\boldsymbol{X}' \sim DFD(\boldsymbol{\theta}')$*,* $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \tau', \boldsymbol{P}')^\intercal$*. Then, if* $\boldsymbol{P}$ *and* $\boldsymbol{P}'$ *are not diagonal matricies,* $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta}')$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}'$*.*

The proof of the above theorem is quite long and, therefore, it is reported in Appendix 8.4.

Note that a diagonal matrix $\mathbf{P}^*$ can be obtained amalgamating a DFD-distributed vector characterized by a non-diagonal matrix $\mathbf{P}$, as showed in 10.

Example 10. *Supposing that* $\boldsymbol{X} \sim DFD(\boldsymbol{\alpha}, \tau, \boldsymbol{P})$*, with* $\boldsymbol{P} = \begin{bmatrix} p_{1,1} & 0 & 0 & 0 \\ 0 & p_{2,2} & p_{2,3} & 0 \\ 0 & p_{3,2} & p_{3,3} & 0 \\ 0 & 0 & 0 & p_{4,4} \end{bmatrix}$. *Then, from Proposition 25, the vector* $(X_1, X_2 + X_3, X_4)^\intercal$ *is distributed according to an Extended Flexible Dirichlet with parameters* $\boldsymbol{\alpha}^* = (\alpha_1, \alpha_2 + \alpha_3, \alpha_4)^\intercal$*,* τ *and* $\boldsymbol{P}^* = \boldsymbol{M_a} \cdot \boldsymbol{P} \cdot \boldsymbol{M_a}^\intercal$*, where* $\boldsymbol{M_a}$ *is the double-collapser matrix associated to the vector* $\boldsymbol{a} = (0, 1, 3, 4)^\intercal$*. Then:*

$$
\begin{aligned}
\boldsymbol{P}^* &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} p_{1,1} & 0 & 0 & 0 \\ 0 & p_{2,2} & p_{2,3} & 0 \\ 0 & p_{3,2} & p_{3,3} & 0 \\ 0 & 0 & 0 & p_{4,4} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}^\intercal . \\
&= \begin{bmatrix} p_{1,1} & 0 & 0 \\ 0 & p_{2,2} + 2p_{2,3} + p_{3,3} & 0 \\ 0 & 0 & p_{4,4} \end{bmatrix}
\end{aligned}
$$

4.2.1 Mixture components and cluster means

An advantage of the Double Flexible Dirichlet is the high number of potential clusters and their position on the simplex. Indeed, if the matrix $\mathbf{P}$ is symmetric and with each $p_{i,j} > 0$, $\frac{D(D+1)}{2}$ clusters are potentially present. These clusters have a rigid allocation scheme on the simplex. From Equation (4.12) it is easy to see that the generic component is distributed according to $\mathcal{D}(\boldsymbol{\alpha} + \tau\ (\mathbf{e}_i + \mathbf{e}_j))$, $i, j = 1, \ldots, D$. Then, the mean vectors of these components are:

$$
\begin{aligned}
\boldsymbol{\mu}_{i,j}^{DFD} &= \frac{\boldsymbol{\alpha} + \tau\ (\mathbf{e}_i + \mathbf{e}_j)}{\alpha^+ + 2\tau} \\
&= \left(\frac{\alpha^+}{\alpha^+ + 2\,\tau}\right) \bar{\boldsymbol{\alpha}} + \left(\frac{\tau}{\alpha^+ + 2\,\tau}\right) \mathbf{e}_i + \left(\frac{\tau}{\alpha^+ + 2\,\tau}\right) \mathbf{e}_j,
\end{aligned}
\tag{4.16}
$$

where $i, j = 1, \ldots, D$, $i \leq j$ and $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$. The constraint $i \leq j$ is due to the symmetry of $\mathbf{P}$.

Please note that $\boldsymbol{\mu}_{i,j}^{DFD}$ does not represent the generic element of a matrix (indeed $\boldsymbol{\mu}_{i,j}^{DFD}$ is a vector!): it is rather associated to the realizations of the random vector

$\mathbf{Z}_1$ and $\mathbf{Z}_2$ ($\mathbf{Z}_1 = \mathbf{e}_i$ and $\mathbf{Z}_2 = \mathbf{e}_j$). These cluster means (and, consequently, the corresponding mixture components) are located in a very rigid scheme on the simplex, as can be seen in Figure 4.3, where the blue triangles represent the cluster means and the green one represents $\bar{\boldsymbol{\alpha}}$ in a scenario with $D = 3$. It is possible to compare Figures 3.4 and 4.3. They are very similar: if we connect the cluster means, we obtain a "mini-simplex" with edges parallel to the simplex ones is formed. In the FD case, the vertices of this scaled simplex are $\boldsymbol{\mu}_1^{FD}$, $\boldsymbol{\mu}_2^{FD}$ and $\boldsymbol{\mu}_3^{FD}$ whereas in the DFD case they are $\boldsymbol{\mu}_{1,1}^{DFD}$, $\boldsymbol{\mu}_{2,2}^{DFD}$ and $\boldsymbol{\mu}_{3,3}^{DFD}$. Vectors $\boldsymbol{\mu}_{i,j}^{DFD}$ with $i \neq j$ are situated at the midpoint of the segment joining $\boldsymbol{\mu}_{i,i}^{DFD}$ and $\boldsymbol{\mu}_{j,j}^{DFD}$.

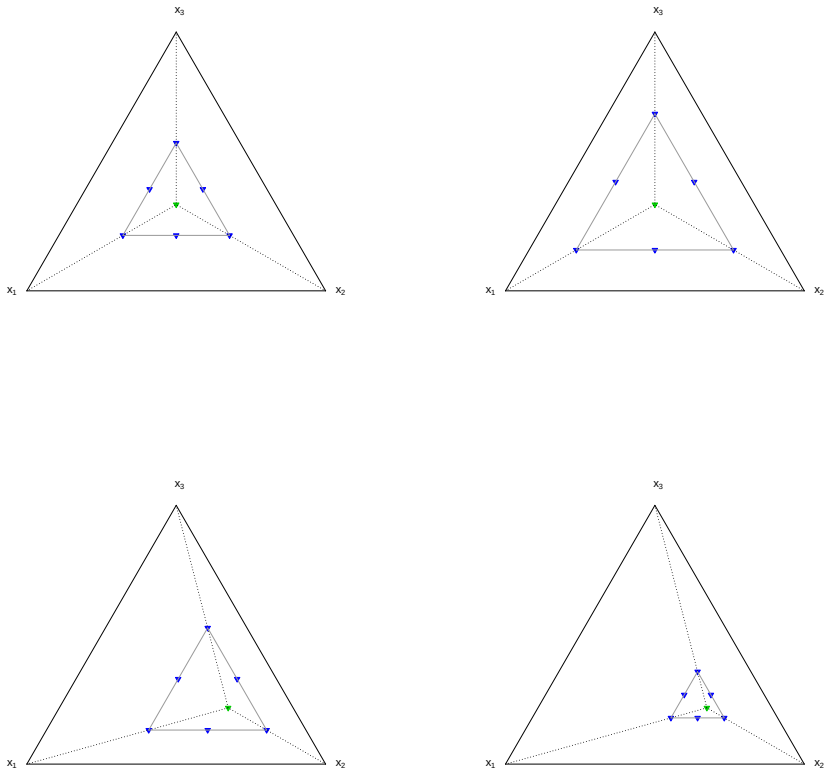

Fig. 4.3: DFD cluster means structure. *Top-Left:* $\boldsymbol{\alpha} = (3, 3, 3)^\intercal$, $\tau = 5$. *Top-Right:* $\boldsymbol{\alpha} = (3, 3, 3)^\intercal$, $\tau = 15$. *Bottom-Left:* $\boldsymbol{\alpha} = (5, 13, 5)^\intercal$, $\tau = 5$. *Bottom-Right:* $\boldsymbol{\alpha} = (5, 13, 5)^\intercal$, $\tau = 15$.

While this structure is quite rigid, it is similar to the one of the FD distribution but allowing for more clusters. Furthermore, thanks to the fact that some $p_{i,j}$ can be equal to 0, this model allows for a variety of cluster that cannot be defined by the FD and the EFD models. For example, in Figure 4.4 it is possible to see some cluster configurations that can not be reached by simpler models. Note that joining the cluster means in the top-left and top-right panels produces a diamond and an inverse triangle, respectively.

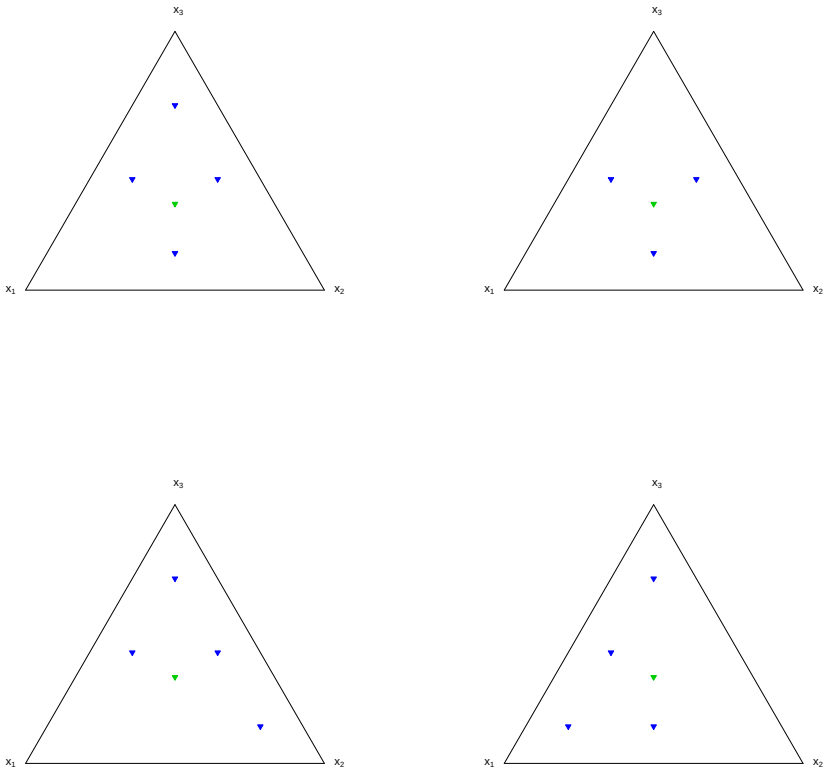

Fig. 4.4: Particular DFD cluster means with $alphav = (5,5,5)^\intercal$ and $\tau = 10$. *Top-Left:* $p_{1,1} = p_{2,2} = 0$. *Top-Right:* $p_{1,1} = p_{2,2} = p_{3,3} = 0$. *Bottom-Left:* $p_{1,1} = p_{1,2} = 0$. *Bottom-Right:* $p_{2,2} = p_{2,3} = 0$.

4.3 Properties

4.3.1 Moments

Thanks to the mixture representation, it is also easy computing joint moments:

Proposition 27 (Joint Moments)**.** *The joint moments of* $\boldsymbol{X} \sim DFD(\boldsymbol{\alpha}, \tau, \boldsymbol{P})$ *can be expressed in the following form:*

$$\begin{aligned}
\mathbb{E}\left[\prod_{r=1}^{D} X_r^{\gamma_r}\right] &= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j}\left[\frac{\Gamma(\alpha^+ + 2\tau)}{\Gamma(\alpha^+ + 2\tau + \gamma^+)}\prod_{r=1}^{D}\frac{\Gamma(\alpha_r + \tau(e_{i_r} + e_{j_r}) + \gamma_r)}{\Gamma(\alpha_r + \tau(e_{i_r} + e_{j_r}))}\right] \\
&= \frac{\Gamma(\alpha^+ + 2\tau)}{\Gamma(\alpha^+ + 2\tau + \gamma^+)}\left[\sum_{i=1}^{D}\sum_{\substack{j=1 \\ i\neq j}}^{D} p_{i,j}\prod_{r=1}^{D}\frac{\Gamma(\alpha_r + \tau(e_{i_r} + e_{j_r}) + \gamma_r)}{\Gamma(\alpha_r + \tau(e_{i_r} + e_{j_r}))} + \right.
\end{aligned}$$

$$+\sum_{i=1}^{D} p_{i,i} \prod_{r=1}^{D} \frac{\Gamma(\alpha_r + 2\tau\, e_{i_r} + \gamma_r)}{\Gamma(\alpha_r + 2\tau\, e_{i_r})}\Bigg]$$

where $\gamma_r \geq 0$ *are non-negative integer and* $\gamma^+ = \sum_{r=1}^{D} \gamma_r$.

In particular, the first two orders moments are:

$$\begin{aligned}
\mathbb{E}[\mathbf{X}] &= \int_{\mathcal{S}^D} \mathbf{x} \left(\sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)) \right) d\mathbf{x} \\
&= \int_{\mathcal{S}^D} \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \mathbf{x} f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)) d\mathbf{x} \\
\overset{(*)}{} &= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \int_{\mathcal{S}^D} \mathbf{x} f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)) d\mathbf{x} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j)}{\alpha^+ + 2\tau} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\boldsymbol{\alpha}}{\alpha^+ + 2\tau} + \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\tau\,\mathbf{e}_i}{\alpha^+ + 2\tau} + \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j} \frac{\tau\,\mathbf{e}_j}{\alpha^+ + 2\tau} \\
&= \frac{\boldsymbol{\alpha}}{\alpha^+ + 2\tau} + \sum_{i=1}^{D} p_{i\cdot} \frac{\tau\,\mathbf{e}_i}{\alpha^+ + 2\tau} + \sum_{j=1}^{D} p_{\cdot j} \frac{\tau\,\mathbf{e}_j}{\alpha^+ + 2\tau} \\
&= \frac{\boldsymbol{\alpha}}{\alpha^+ + 2\tau} + \frac{\tau\,\mathbf{p}}{\alpha^+ + 2\tau} + \frac{\tau\,\mathbf{p}}{\alpha^+ + 2\tau} \\
&= \frac{\boldsymbol{\alpha} + 2\tau\,\mathbf{p}}{\alpha^+ + 2\tau}
\end{aligned} \tag{4.17}$$

where the integral in $^{(*)}$ is the expectation of $\mathcal{D}(\boldsymbol{\alpha} + \tau(\mathbf{e}_i + \mathbf{e}_j))$. One can note that $\dfrac{\boldsymbol{\alpha} + 2\tau\mathbf{p}}{\alpha^+ + 2\tau}$ coincides with the expected value of a Flexible Dirichlet model with $\tau_{\text{FD}} = 2\tau_{\text{DFD}}$.

From Proposition 27 it follows that:

$$\mathbb{E}[X_r \cdot X_h] = \frac{\alpha_r\,\alpha_h + 2\tau(\alpha_r p_{h\cdot} + \alpha_h p_{r\cdot}) + 2\tau^2 p_{rh}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)}. \tag{4.18}$$

$$\mathbb{E}\left[X_r^2\right] = \frac{\alpha_r^2 + \alpha_r + 4\tau\alpha_h p_{r\cdot} + 2\tau^2 p_{r\cdot} + 2\tau^2 p_{r,r} + 2\tau p_{r\cdot}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)}. \tag{4.19}$$

$$
\begin{aligned}
\text{Var}\left(X_r\right) &= \mathbb{E}\left[X_r^2\right] - \left(\mathbb{E}\left[X_r\right]\right)^2 \\
&= \frac{\alpha_r^2 + \alpha_r + 4\tau\alpha_h p_{r\cdot} + 2\tau^2 p_{r\cdot} + 2\tau^2 p_{r,r} + 2\tau p_{r\cdot}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} - \left(\frac{\alpha_r + 2\tau p_{r\cdot}}{\alpha^+ + 2\tau}\right)^2 \\
&= \frac{\mathbb{E}\left[X_r\right]\left(1 - \mathbb{E}\left[X_r\right]\right)}{\alpha^+ + 2\tau + 1} + \frac{2\tau^2\left(p_{r\cdot}(1 - 2p_{r\cdot}) + p_{r,r}\right)}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)}
\end{aligned}
\tag{4.20}
$$

$$
\begin{aligned}
\text{Cov}\left(X_r, X_h\right)_{r\neq h} &= \mathbb{E}\left[X_r \cdot X_h\right] - \left(\mathbb{E}\left[X_r\right]\mathbb{E}\left[X_h\right]\right) \\
&= \frac{\alpha_r\,\alpha_h + 2\tau(\alpha_r p_{h\cdot} + \alpha_h p_{r\cdot}) + 2\tau^2 p_{r,h}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} - \left(\frac{\alpha_r + 2\tau p_{r\cdot}}{\alpha^+ + 2\tau}\right)\left(\frac{\alpha_h + 2\tau p_{h\cdot}}{\alpha^+ + 2\tau}\right) \\
&= -\frac{\mathbb{E}\left[X_r\right]\mathbb{E}\left[X_h\right]}{\alpha^+ + 2\tau + 1} + \frac{2\tau^2(p_{r,h} - 2p_{r\cdot}p_{h\cdot})}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)}
\end{aligned}
\tag{4.21}
$$

In general, closing a basis leads to unclear covariance structure, due to the sum-to-1 constraint. In other words, the covariance matrix of a basis is rarely connected to the covariance matrix of a composition in a simple way. The most simple but effective example to show is the closure of a basis formed by independent Gamma random variables, as in Example 6. Therefore, no automatic relation between basis and composition dependence structure exists; rather it depends on the underlying distribution. An interesting feature of the DFD model is that the dependence induced in the basis appears in the composition. Indeed, the covariance among two elements of a DFD-distributed vector can be written as:

$$
\text{Cov}\left(X_r, X_h\right) = -\frac{\mathbb{E}\left[X_r\right]\mathbb{E}\left[X_h\right]}{\alpha^+ + 2\tau + 1} + \frac{\text{Cov}\left(Y_r, Y_h\right)}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)}. \tag{4.22}
$$

The first element (always negative) is due to the closure of a Gamma-related basis, whereas the second is exactly the covariance of the corresponding basis' elements multiplied by a constant. Since this last part can assume both positive and negative values, according to the difference $(p_{r,h} - 2p_{r\cdot}p_{h\cdot})$, it influences the negative linear dependence which is typical to the Dirichlet. In particular, thanks to this new term, the covariance among two components can assume values greater than zero, allowing for positive dependence. This is a noteworthy aspect, because most distributions on the simplex do not have such a coherent dependence structure between the basis and the composition. Nonetheless, the FD has a similar structure, as it can be seen by looking at Equations (3.15) and (3.21).

The analytical expression for the correlation coefficient of two arbitrary components X_r and X_h $(r, h = 1, \ldots, D,\ r \neq h)$ of $\mathbf{X}$

$$\rho_{r,h} = \frac{\text{Cov}\left(X_r, X_h\right)}{\sqrt{\text{Var}\left(X_r\right) \cdot \text{Var}\left(X_h\right)}}.$$

is heavy and hardly tractable; however it is easy to show that it may take high positive values. For example, setting the matrix $\mathbf{P}$ as in (4.11) leads to the following quantities:

$$\begin{aligned}
\mathbb{E}\left[X_h\right] &= \frac{\alpha_h + \frac{\tau}{2}}{\alpha^+ + 2\tau} \underset{\tau \to +\infty}{\longrightarrow} \frac{1}{4} \\
\text{Var}\left(X_h\right) &= \frac{\mathbb{E}\left[X_r\right]\left(1 - \mathbb{E}\left[X_r\right]\right)}{\alpha^+ + 2\tau + 1} + \frac{2\tau^2 \frac{1}{8}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} \underset{\tau \to +\infty}{\longrightarrow} \frac{1}{16} \\
\text{Cov}\left(X_r, X_h\right) &= -\frac{\mathbb{E}\left[X_r\right]\mathbb{E}\left[X_h\right]}{\alpha^+ + 2\tau + 1} + \frac{2\tau^2 \frac{1}{8}}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} \underset{\tau \to +\infty}{\longrightarrow} -0 + \frac{1}{16} = \frac{1}{16}
\end{aligned}$$

$$\Longrightarrow \rho_{r,h} = \frac{\text{Cov}\left(X_r, X_h\right)}{\sqrt{\text{Var}\left(X_h\right) \cdot \text{Var}\left(X_r\right)}} \underset{\tau \to +\infty}{\longrightarrow} \frac{\frac{1}{16}}{\sqrt{\frac{1}{16} \cdot \frac{1}{16}}} = 1$$

It can also be shown that the DFD model allows for negative dependence. This is not surprising, since the unit-sum constraint naturally induces a negative dependence. Setting the elements of the matrix $\mathbf{P}$ as in (4.10) and taking limits as $\tau \to +\infty$:

$$\begin{aligned}
\mathbb{E}\left[X_h\right] &= \frac{\alpha_h + \tau}{\alpha^+ + 2\tau} \underset{\tau \to +\infty}{\longrightarrow} \frac{1}{2} \\
\text{Var}\left(X_h\right) &= \frac{\mathbb{E}\left[X_h\right]\left(1 - \mathbb{E}\left[X_h\right]\right)}{\alpha^+ + 2\tau + 1} + \frac{\tau^2}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} \underset{\tau \to +\infty}{\longrightarrow} 0 + \frac{1}{4} = \frac{1}{4} \\
\text{Cov}\left(X_r, X_h\right) &= -\frac{\mathbb{E}\left[X_r\right]\mathbb{E}\left[X_h\right]}{\alpha^+ + 2\tau + 1} + \frac{-\tau^2}{(\alpha^+ + 2\tau + 1)(\alpha^+ + 2\tau)} \underset{\tau \to +\infty}{\longrightarrow} 0 - \frac{1}{4} = -\frac{1}{4}
\end{aligned}$$

$$\Longrightarrow \rho_{r,h} = \frac{\text{Cov}\left(X_r, X_h\right)}{\sqrt{\text{Var}\left(X_h\right) \cdot \text{Var}\left(X_r\right)}} \underset{\tau \to +\infty}{\longrightarrow} -\frac{\frac{1}{4}}{\sqrt{\frac{1}{4} \cdot \frac{1}{4}}} = -1$$

Example 11. *Let $\boldsymbol{P}$ be a symmetric matrix equal to the one defined in Example (9). Figure (4.5) shows how the correlation coefficient among X_1 and X_3 changes according to $\boldsymbol{\alpha}$ and τ.*

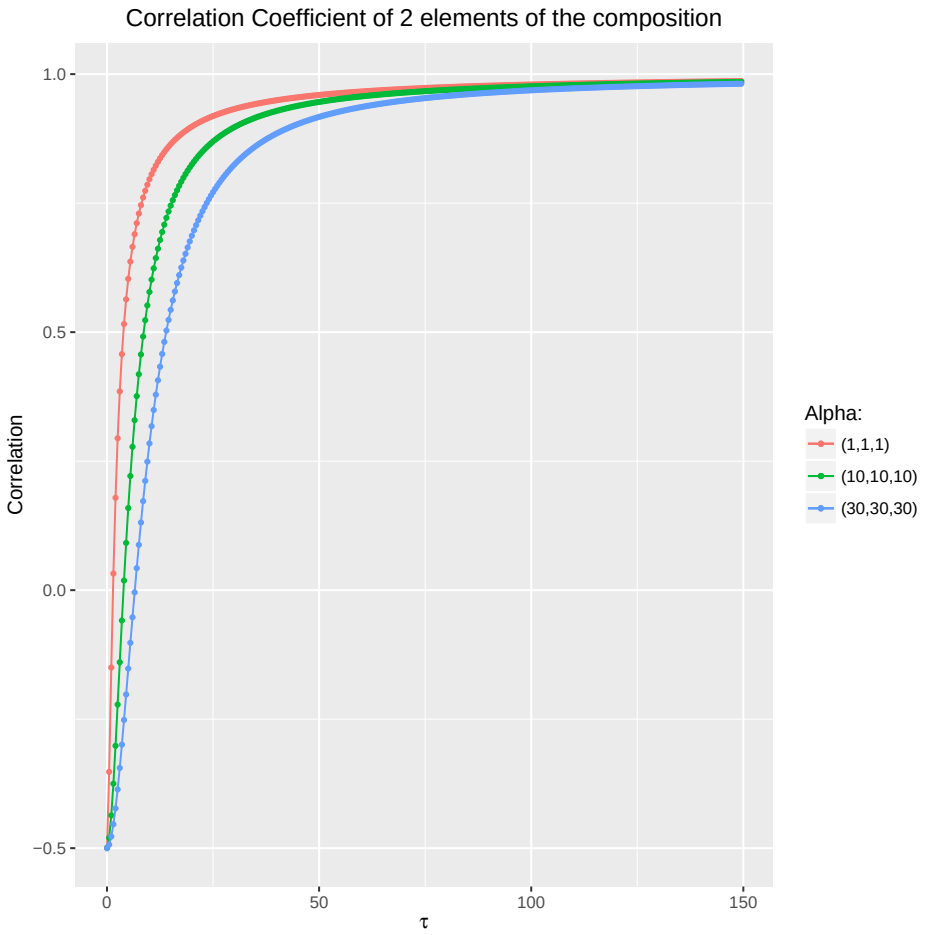

Fig. 4.5: Correlation between X_1 and X_3.

4.3.2 Conditional distributions

Let $\mathbf{S}_1 = (Y_1, \dots, Y_k)^\intercal / Y_1^+$ be a $k-$dimensional subcomposition. It could be of interest to compute the distribution of $\mathbf{S}_1 \,|\mathbf{X}_2 = \mathbf{x}_2$. Let Z_1^+ and Z_2^+ be two random variables such that $Z_1^+ = \sum_{h=1}^k Z_{1,h}$ and $Z_2^+ = \sum_{h=1}^k Z_{2,h}$.

- Condition on $Z_1^+ = 1, Z_2^+ = 1$:

$$\mathbf{S}_1 \Big| (Z_1^+ = 1, Z_2^+ = 1) \sim \text{DFD}^k \left(\boldsymbol{\alpha}_1, \tau, \frac{\mathbf{P}_{11}^*}{P_{11}^{*+}} \right),$$

where $\mathbf{P}_{11}^* = \mathbf{M_a} \cdot \mathbf{P} \cdot \mathbf{M_a}^\intercal$ and $\mathbf{M_a}$ is the double-collapser matrix associated to $\mathbf{a} = (0, k, D)$. By compositional invariance of the DFD model, $\mathbf{S}_1$ is independent of $Y_1^+ \Big| (Z_1^+ = 1, Z_2^+ = 1)$. Given $Z_1^+ = 1, Z_2^+ = 1$, $\mathbf{Y}_1 = (Y_1, \dots, Y_k)^\intercal \sim \text{DFG} \left(\boldsymbol{\alpha}_1, \tau, \frac{\mathbf{P}_{11}^*}{P_{11}^{*+}} \right)$, $\mathbf{Y}_2 = (Y_{k+1}, \dots, Y_D)^\intercal$ is a random vector with independent gamma components and $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2$. Due to compositional invariance of both Dirichlet and DFD, $(\mathbf{S}_1, \mathbf{S}_2, Y_1^+, Y_2^+)^\intercal$ is a vector with independent components:

then $\mathbf{S}_1 \perp\!\!\!\perp g(\mathbf{S}_2, Y_1^+, Y_2^+)$. Since $\mathbf{X}_2 = g(\mathbf{S}_2, Y_1^+, Y_2^+) = \dfrac{\mathbf{S}_2 \cdot Y_2^+}{Y_1^+ + Y_2^+}$, $\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2 \left| (Z_1^+ = 1, Z_2^+ = 1) \right.$.

- Condition on $Z_1^+ = 1, Z_2^+ = 0$:

$$\mathbf{S}_1 \left| (Z_1^+ = 1, Z_2^+ = 0) \sim \text{FD}^k \left(\boldsymbol{\alpha}_1, \tau, \frac{\left(\sum_{j=k+1}^{D} p_{1j}, \sum_{j=k+1}^{D} p_{2j}, \ldots, \sum_{j=k+1}^{D} p_{kj} \right)^\intercal}{P_{12}^{*+}} \right) \right..$$

With similat arguments it is possible to show that $\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2 \left| (Z_1^+ = 1, Z_2^+ = 0) \right.$.

- Let us condition on $Z_1^+ = 0, Z_2^+ = 1$.

$$\mathbf{S}_1 \left| (Z_1^+ = 0, Z_2^+ = 1) \sim \text{FD}^k \left(\boldsymbol{\alpha}_1, \tau, \frac{\left(\sum_{i=k+1}^{D} p_{i1}, \sum_{i=k+1}^{D} p_{i2}, \ldots, \sum_{i=k+1}^{D} p_{ik} \right)^\intercal}{P_{21}^{*+}} \right) \right..$$

Also in this case $\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2 \left| (Z_1^+ = 0, Z_2^+ = 1) \right.$ (same arguments).

Thanks to the symmetry of $\mathbf{P}$, $\mathbf{S}_1 \left| (Z_1^+ = 1, Z_2^+ = 0) \sim \mathbf{S}_1 \right| (Z_1^+ = 0, Z_2^+ = 1)$. Then it is possible to write:

$$\mathbf{S}_1 \left| (Z_1^+ = 0, Z_2^+ = 1) \sim \text{FD}^k \left(\boldsymbol{\alpha}_1, \tau, \frac{\left(\sum_{j=k+1}^{D} p_{1j}, \sum_{j=k+1}^{D} p_{2j}, \ldots, \sum_{j=k+1}^{D} p_{kj} \right)^\intercal}{P_{12}^{*+}} \right) \right..$$

- Condition on $Z_1^+ = 0, Z_2^+ = 0$:

$$\mathbf{S}_1 \left| (Z_1^+ = 0, Z_2^+ = 0) \sim \text{Dir}^k (\boldsymbol{\alpha}_1) \right..$$

Once again, with the same arguments it is possible to show that $\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2 \left| (Z_1^+ = 0, Z_2^+ = 0) \right.$.

One can show that:

z_1^+	z_2^+	$P\left(Z_1^+ = z_1^+, Z_2^+ = z_2^+\right)$
1	1	P_{11}^{*+}
1	0	P_{12}^{*+}
0	1	P_{21}^{*+}
0	0	P_{22}^{*+}

Since $\mathbf{P}$ is symmetric, $P_{12}^{*+} = P_{21}^{*+}$. Finally, from $(\mathbf{S}_1 \perp\!\!\!\perp \mathbf{X}_2)|(Z_1^+, Z_2^+)$, the distribution function of $\mathbf{S}_1\,|\mathbf{X}_2 = \mathbf{x}_2$ is easily obtained with basic probability theory:

$$
\begin{aligned}
F_{\mathbf{S}_1|\mathbf{X}_2=\mathbf{x}_2}(\mathbf{s}_1) &= \sum_{z_1^+=0}^{1}\sum_{z_2^+=0}^{1} F_{\mathbf{S}_1|\mathbf{X}_2=\mathbf{x}_2, Z_1^+=z_1^+, Z_2^+=z_2^+}(\mathbf{s}_1)\cdot P(Z_1^+ = z_1^+, Z_2^+ = z_2^+|\mathbf{X}_2=\mathbf{x}_2) \\
&= \sum_{z_1^+=0}^{1}\sum_{z_2^+=0}^{1} F_{\mathbf{s}_1|Z_1^+=z_1^+, Z_2^+=z_2^+}(\mathbf{S}_1)\cdot P(Z_1^+ = z_1^+, Z_2^+ = z_2^+|\mathbf{X}_2=\mathbf{x}_2) \\
&= \text{DFD}^k\left(\boldsymbol{\alpha}_1, \tau, \frac{\mathbf{P}_{11}^*}{P_{11}^{*+}}\right)\cdot P\left(Z_1^+ = 1, Z_2^+ = 1|\mathbf{X}_2=\mathbf{x}_2\right) + \\
&+ \text{FD}^k\left(\boldsymbol{\alpha}_1, \tau, \frac{\left(\sum_{j=k+1}^{D} p_{1j}, \sum_{j=k+1}^{D} p_{2j}, \ldots, \sum_{j=k+1}^{D} p_{kj}\right)^\intercal}{P_{12}^{*+}}\right)\cdot P\left(Z_1^+ = 1, Z_2^+ = 0|\mathbf{X}_2=\mathbf{x}_2\right) + \\
&+ \text{FD}^k\left(\boldsymbol{\alpha}_1, \tau, \frac{\left(\sum_{j=k+1}^{D} p_{1j}, \sum_{j=k+1}^{D} p_{2j}, \ldots, \sum_{j=k+1}^{D} p_{kj}\right)^\intercal}{P_{12}^{*+}}\right)\cdot P\left(Z_1^+ = 0, Z_2^+ = 1|\mathbf{X}_2=\mathbf{x}_2\right) + \\
&+ \text{Dir}^k(\boldsymbol{\alpha}_1)\cdot P\left(Z_1^+ = 0, Z_2^+ = 0|\mathbf{X}_2=\mathbf{x}_2\right)
\end{aligned}
$$

This means that the conditional distribution of $\mathbf{S}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ can be expressed as a finite mixture with Dirichlet, FD and EFD components and weights that depend on the value of $\mathbf{x}_2$ and that can be computed by the following formula:

$$
P\left(Z_1^+ = z_1^+, Z_2^+ = z_2^+\,|\mathbf{X}_2=\mathbf{x}_2\right) = \frac{P\left(Z_1^+ = z_1^+, Z_2^+ = z_2^+\right) f_{\mathbf{X}_2|Z_1^+=z_1^+, Z_2^+=z_2^+}(\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)} \tag{4.23}
$$

Thanks to the result on marginals, the denominator $f_{\mathbf{X}_2}(\mathbf{x}_2)$ is equal to the density function of:

$$(1-X_2^+,\mathbf{X}_2)^\intercal \sim \text{DFD}^{1+D-k}\left((\alpha^+-\alpha_2^+,\boldsymbol{\alpha}_2)^\intercal,\tau,\left[\begin{array}{c|ccc} P_{11}^{*+} & \sum_{i=1}^{k} p_{i,k+1} & \sum_{i=1}^{k} p_{i,k+2} & \cdots \quad \sum_{i=1}^{k} p_{i,D} \\ \hline \sum_{j=1}^{k} p_{k+1,j} & & & \\ \sum_{j=1}^{k} p_{k+2,j} & & \mathbf{P}^*_{22} & \\ \vdots & & & \\ \sum_{j=1}^{k} p_{D,j} & & & \end{array}\right]\right), \tag{4.24}$$

where $\mathbf{P}^*_{22} = \mathbf{M_a}\cdot\mathbf{P}\cdot\mathbf{M_a^\intercal}$ is the result of a matrix multiplication involving a double collapser matrix. In order to compute $f_{\mathbf{X}_2|Z_1^+=z_1^+,Z_2^+=z_2^+}(\mathbf{x}_2)$, it is necessary to note that:

$$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \left|(Z_1^+=1,Z_2^+=1) \sim \text{Dir}^{1+D-k}\left(\left(\alpha^+-\alpha_2^++2\tau,\boldsymbol{\alpha}_2\right)^\intercal\right)\right. \tag{4.25}$$

$$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \left|(Z_1^+=1,Z_2^+=0) \sim\right.$$
$$\sim \text{FD}^{1+D-k}\left((\alpha^+-\alpha_2^++\tau,\boldsymbol{\alpha}_2)^\intercal,\tau,\left(0,\frac{\sum_{i=1}^{k} p_{i,k+1},\sum_{i=1}^{k} p_{i,k+2},\ldots,\sum_{i=1}^{k} p_{i,D}}{P_{12}^{*+}}\right)^\intercal\right) \tag{4.26}$$

$$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \left|(Z_1^+=0,Z_2^+=1) \sim\right.$$
$$\sim \text{FD}^{1+D-k}\left((\alpha^+-\alpha_2^++\tau,\boldsymbol{\alpha}_2)^\intercal,\tau,\left(0,\frac{\sum_{j=1}^{k} p_{k+1,j},\sum_{j=1}^{k} p_{k+2,j},\ldots,\sum_{j=1}^{k} p_{D,j}}{P_{21}^{*+}}\right)^\intercal\right) \tag{4.27}$$

Thanks to the symmetry of $\mathbf{P}$, it can be shown that:

$$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \Big|(Z_1^+=0,Z_2^+=1) \sim \\ \sim \text{FD}^{1+D-k}\left((\alpha^+-\alpha_2^++\tau,\boldsymbol{\alpha}_2)^\intercal,\tau,\left(0,\frac{\sum_{i=1}^k p_{i,k+1},\sum_{i=1}^k p_{i,k+2},\ldots,\sum_{i=1}^k p_{i,D}}{P_{12}^{*+}}\right)^\intercal\right). \tag{4.28}$$

The above result means that:
$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \Big|(Z_1^+=1,Z_2^+=0) \sim \left(1-X_2^+,\mathbf{X}_2\right)^\intercal \Big|(Z_1^+=0,Z_2^+=1)$

The last conditional distribution needed is:

$$\left(1-X_2^+,\mathbf{X}_2\right)^\intercal \Big|(Z_1^+=0,Z_2^+=0) \sim \text{DFD}^{1+D-k}\left((\alpha^+-\alpha_2^+,\boldsymbol{\alpha}_2)^\intercal,\tau,\begin{bmatrix} 0 & 0 \\ 1\times 1 & 1\times(D-k) \\ 0 & \mathbf{P}_{22}^*/P_{22}^{*+} \\ (D-k)\times 1 & (D-k)\times(D-k)\end{bmatrix}\right) \tag{4.29}$$

In order to simplify future notation, it is possible to introduce the following functions:

- $\gamma(\mathbf{x}_2;\phi,\varphi)=\sum_{i=k+1}^{D}\sum_{\substack{j=k+1\\ i\neq j}}^{D}\frac{p_{i,j}\Gamma(\alpha_i)\Gamma(\alpha_j)(x_ix_j)^\phi}{\Gamma(\alpha_i+\tau)\Gamma(\alpha_j+\tau)(1-x_2^+)^\varphi}$
- $\delta(\mathbf{x}_2;\phi,\varphi)=\sum_{i=k+1}^{D}\frac{p_{i,i}\Gamma(\alpha_i)x_i^\phi}{\Gamma(\alpha_i+2\tau)(1-x_2^+)^\varphi}$
- $\nu(\mathbf{x}_2;\phi,\varphi)=\sum_{h=k+1}^{D}\frac{(\sum_{r=1}^k p_{h,r})\Gamma(\alpha_1^+)\Gamma(\alpha_h)x_h^\phi}{\Gamma(\alpha_1^++\tau)\Gamma(\alpha_h+\tau)(1-x_2^+)^\varphi}$

Then it is possible to write:

$$f_{\mathbf{X}_2}(\mathbf{x}_2)=\frac{\Gamma(\alpha^++2\tau)\left(\prod_{r=k+1}^{D}x_r^{\alpha_r-1}\right)(1-x_2^+)^{\alpha_1^+-1}}{\Gamma(\alpha_1^+)\prod_{r=k+1}^{D}\Gamma(\alpha_r)}\cdot\left[\gamma(\mathbf{x}_2;\tau,0)+\delta(\mathbf{x}_2;2\tau,0)+\right. \\ \left.+\frac{P_{11}^{*+}\Gamma(\alpha_1^+)(1-x_2^+)^{2\tau}}{\Gamma(\alpha_1^++2\tau)}+2\nu(\mathbf{x}_2;\tau,0)\right]$$

Finally, thanks to Equation (4.23) it is possible to compute the weights of the mixture:

$$P\left(Z_1^+ = 1, Z_2^+ = 1 \,|\mathbf{X}_2 = \mathbf{x}_2\right) = \frac{P\left(Z_1^+ = 1, Z_2^+ = 1\right) f_{\mathbf{X}_2|Z_1^+=1,Z_2^+=1}(\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)}$$
$$= \frac{P_{11}^{*+}}{P_{11}^{*+} + \frac{\Gamma(\alpha_1^+ + 2\tau)}{\Gamma(\alpha_1^+)}\left[\gamma(\mathbf{x}_2;\tau,2\tau) + \delta(\mathbf{x}_2;2\tau,2\tau) + 2\nu(\mathbf{x}_2;\tau,\tau)\right]}$$

$$P\left(Z_1^+ = 0, Z_2^+ = 0 \,|\mathbf{X}_2 = \mathbf{x}_2\right) = \frac{P\left(Z_1^+ = 0, Z_2^+ = 0\right) f_{\mathbf{X}_2|Z_1^+=0,Z_2^+=0}(\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)}$$
$$= \frac{1}{1 + \frac{1}{q(\mathbf{x}_2)}\left[P_{11}^{*+}\frac{\Gamma(\alpha_1^+)(1-x_2^+)^{2\tau}}{\Gamma(\alpha_1^+ + 2\tau)} + 2\nu(\mathbf{x}_2;\tau,-\tau)\right]}$$

where $q(\mathbf{x}_2) = \gamma(\mathbf{x}_2;\tau,0) + \delta(\mathbf{x}_2;2\tau,0)$.

$$P\left(Z_1^+ = 1, Z_2^+ = 0 \,|\mathbf{X}_2 = \mathbf{x}_2\right) = P\left(Z_1^+ = 0, Z_2^+ = 1 \,|\mathbf{X}_2 = \mathbf{x}_2\right) = \frac{P\left(Z_1^+ = 1, Z_2^+ = 0\right) f_{\mathbf{X}_2|Z_1^+=1,Z_2^+=0}(\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)}$$
$$= \frac{1}{2 + \frac{1}{\nu(\mathbf{x}_2;\tau,0)}\left[\gamma(\mathbf{x}_2;\tau,\tau) + \delta(\mathbf{x}_2;2\tau,\tau) + \frac{P_{11}^{*+}\cdot\Gamma(\alpha_1^+)(1-x_2^+)^{\tau}}{\Gamma(\alpha_1^+ + 2\tau)}\right]}$$

4.3.3 Symmetrized Kullback-Leibler Divergence

In order to compute the symmetrized Kullback-Leibler divergence, it is useful to remember that:

- Each mixture component follows a Dirichlet distribution: $f_{i,j} \equiv f_{i,j}(\mathbf{x}; \boldsymbol{\alpha}, \tau) = f_{\mathcal{D}}(\mathbf{x}; \boldsymbol{\alpha} + \tau\,(\mathbf{e}_i + \mathbf{e}_j))$.
- If $\mathbf{X} \sim \text{Dir}(\boldsymbol{\alpha}) \implies \mathbb{E}\left[\ln X_r\right] = \psi(\alpha_r) - \psi(\alpha^+)$, where $\psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x)$ is the digamma function.

Let $i \neq j \neq r \neq h$ be four generic indices assuming value in $\{1, 2, \ldots, D\}$ and $f_{i,j}(\mathbf{x}; \boldsymbol{\alpha}, \tau)$ is the density function of a Dirichlet with parameters $(\boldsymbol{\alpha} + \tau(\mathbf{e}_i + \mathbf{e}_j))$. Then, the following quantities are of interest:

- $f_{i,j}(\mathbf{x}; \boldsymbol{\alpha}, \tau) = \dfrac{\Gamma(\alpha^+ + 2\tau)\Gamma(\alpha_i)\Gamma(\alpha_j)}{\left[\prod_{r=1}^{D} \Gamma(\alpha_r)\right] \Gamma(\alpha_i + \tau)\Gamma(\alpha_j + \tau)} x_i^{\tau} x_j^{\tau} \prod_{r=1}^{D} x_r^{\alpha_r - 1}$
- $f_{i,i}(\mathbf{x}; \boldsymbol{\alpha}, \tau) = \dfrac{\Gamma(\alpha^+ + 2\tau)\Gamma(\alpha_i)}{\left[\prod_{r=1}^{D} \Gamma(\alpha_r)\right] \Gamma(\alpha_i + 2\tau)} x_i^{2\tau} \prod_{r=1}^{D} x_r^{\alpha_r - 1}$
- $\dfrac{f_{i,j}}{f_{i,i}} = \left(\dfrac{x_j}{x_i}\right)^{\tau} \dfrac{\Gamma(\alpha_j)\Gamma(\alpha_i + 2\tau)}{\Gamma(\alpha_j + \tau)\Gamma(\alpha_i + \tau)} = \left(\dfrac{x_j}{x_i}\right)^{\tau} C_{i,j;i,i}$
- $\dfrac{f_{i,j}}{f_{r,r}} = \left(\dfrac{x_i x_j}{x_r}\right)^{\tau} \dfrac{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_r + 2\tau)}{\Gamma(\alpha_i + \tau)\Gamma(\alpha_j + \tau)\Gamma(\alpha_r)} = \left(\dfrac{x_i x_j}{x_r}\right)^{\tau} C_{i,j;r,r}$
- $\dfrac{f_{i,j}}{f_{r,j}} = \left(\dfrac{x_i x_j}{x_r x_h}\right)^{\tau} \dfrac{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_r + \tau)\Gamma(\alpha_h + \tau)}{\Gamma(\alpha_i + \tau)\Gamma(\alpha_j + \tau)\Gamma(\alpha_r)\Gamma(\alpha_h)} = \left(\dfrac{x_i x_j}{x_r x_h}\right)^{\tau} C_{i,j;r,h}$
- $\dfrac{f_{i,i}}{f_{r,r}} = \left(\dfrac{x_i}{x_r}\right)^{\tau} \dfrac{\Gamma(\alpha_i)\Gamma(\alpha_r + 2\tau)}{\Gamma(\alpha_i + 2\tau)\Gamma(\alpha_r)} = \left(\dfrac{x_i}{x_r}\right)^{\tau} C_{i,i;r,r}$
- $\dfrac{f_{i,j}}{f_{i,h}} = \left(\dfrac{x_j}{x_r}\right)^{\tau} \dfrac{\Gamma(\alpha_j)\Gamma(\alpha_r + \tau)}{\Gamma(\alpha_j + \tau)\Gamma(\alpha_r)} = \left(\dfrac{x_j}{x_r}\right)^{\tau} C_{i,j;i,h}$
- $\dfrac{f_{i,i}}{f_{i,j}} = \left(\dfrac{f_{i,j}}{f_{i,i}}\right)^{-1}$; $\dfrac{f_{r,r}}{f_{i,j}} = \left(\dfrac{f_{i,j}}{f_{r,r}}\right)^{-1}$; $\dfrac{f_{r,h}}{f_{i,j}} = \left(\dfrac{f_{i,j}}{f_{r,h}}\right)^{-1}$; $\dfrac{f_{r,r}}{f_{i,i}} = \left(\dfrac{f_{i,i}}{f_{r,r}}\right)^{-1}$; $\dfrac{f_{i,h}}{f_{i,j}} = \left(\dfrac{f_{i,j}}{f_{i,h}}\right)^{-1}$

$$
\begin{aligned}
d_{KL}(f_{i,j}, f_{i,i}) &= \int f_{i,j} \cdot \ln \frac{f_{i,j}}{f_{i,i}} d\mathbf{x} = \int f_{i,j} \cdot [\ln C_{i,j;i,i} + \tau \ln x_j - \tau \ln x_i] \, d\mathbf{x} \\
&= \ln C_{i,j;i,i} + \tau \, \mathbb{E}\left[\ln X_j\right] - \tau \, \mathbb{E}\left[\ln X_i\right] \\
&= \ln C_{i,j;i,i} + \tau \left\{ \psi(\alpha_j + \tau) - \psi(\alpha^+ + 2\,\tau) - \left[\psi(\alpha_i + \tau) - \psi(\alpha^+ + 2\,\tau) \right] \right\} \\
&= \ln C_{i,j;i,i} + \tau \, \left\{ \psi(\alpha_j + \tau) - \psi(\alpha_i + \tau) \right\}
\end{aligned}
$$

where $\mathbb{E}\left[\cdot\right]$ is with respect to $\mathcal{D}(\boldsymbol{\alpha} + \tau(\mathbf{e}_i + \mathbf{e}_j))$.

$$
\begin{aligned}
d_{KL}(f_{i,i}, f_{i,j}) &= \int f_{i,i} \cdot \ln \frac{f_{i,i}}{f_{i,j}} d\mathbf{x} = -\int f_{i,i} \cdot \ln \frac{f_{i,j}}{f_{i,i}} d\mathbf{x} = -\int f_{i,i} \cdot [\ln C_{i,j;i,i} + \tau \ln x_j - \tau \ln x_i] \, d\mathbf{x} \\
&= -\left\{\ln C_{i,j;i,i} + \tau \, \mathbb{E}\left[\ln X_j\right] - \tau \, \mathbb{E}\left[\ln X_i\right]\right\} \\
&= -\ln C_{i,j;i,i} - \tau \left\{ \psi(\alpha_j) - \psi(\alpha^+ + 2\,\tau) - \left[\psi(\alpha_i + 2\,\tau) - \psi(\alpha^+ + 2\,\tau) \right] \right\} \\
&= -\ln C_{i,j;i,i} - \tau \, \left\{ \psi(\alpha_j) - \psi(\alpha_i + 2\,\tau) \right\}
\end{aligned}
$$

where $\mathbb{E}\left[\cdot\right]$ is with respect to $\mathcal{D}(\boldsymbol{\alpha} + 2\,\tau\,\mathbf{e}_i)$.

$$
\begin{aligned}
d_{SKL}(f_{i,j}, f_{i,i}) &= d_{KL}(f_{i,j}, f_{i,i}) + d_{KL}(f_{i,i}, f_{i,j}) \\
&= \cancel{\ln C_{i,j;i,i}} + \tau \, \left(\psi(\alpha_j + \tau) - \psi(\alpha_i + \tau) \right) - \cancel{\ln C_{i,j;i,i}} - \tau \, \left(\psi(\alpha_j) - \psi(\alpha_i + 2\,\tau) \right) \\
&= \tau \, \left[\psi(\alpha_j + \tau) - \psi(\alpha_i + \tau) - \psi(\alpha_j) + \psi(\alpha_i + 2\,\tau) \right]
\end{aligned}
$$

With the same arguments it is easy to show that:

- $d_{SKL}(f_{i,j}, f_{i,i}) = \tau \, \left[\psi(\alpha_j + \tau) - \psi(\alpha_j) + \psi(\alpha_i + 2\,\tau) - \psi(\alpha_i + \tau) \right]$
- $d_{SKL}(f_{i,j}, f_{r,r}) = \tau \, \left[\psi(\alpha_i + \tau) - \psi(\alpha_i) + \psi(\alpha_j + \tau) - \psi(\alpha_j) + \psi(\alpha_r + 2\,\tau) - \psi(\alpha_r) \right]$
- $d_{SKL}(f_{i,j}, f_{i,h}) = \tau \, \left[\psi(\alpha_j + \tau) - \psi(\alpha_j) + \psi(\alpha_h + \tau) - \psi(\alpha_h) \right]$
- $d_{SKL}(f_{i,i}, f_{r,r}) = \tau \, \left[\psi(\alpha_i + 2\,\tau) - \psi(\alpha_i) + \psi(\alpha_r + 2\,\tau) - \psi(\alpha_r) \right]$
- $d_{SKL}(f_{i,j}, f_{r,h}) = \tau \, \left[\psi(\alpha_i + \tau) - \psi(\alpha_i) + \psi(\alpha_j + \tau) - \psi(\alpha_j) + \right.$
  $\left. + \, \psi(\alpha_r + \tau) - \psi(\alpha_r) + \psi(\alpha_h + \tau) - \psi(\alpha_h) \right]$

A graphical investigation of several ternary plots of different scenarios shows that values of $d_{SKL}(\cdot, \cdot)$ greater than 20 characterize two well separated clusters.

4.4 Computational issues

In the previous sections the DFD distribution has been introduced and some theoretical properties have been listed. In this section the interest is in providing an estimation procedure for the parameters $\boldsymbol{\alpha}$, τ and $\mathbf{P}$.

4.4.1 Cluster-code matrix

During this work, the necessity of finding a way to identify clusters has arisen. In the previous Flexible models (Sections 3.3 and 3.4) clusters are identified thanks to a precise cluster means structure. Let $\boldsymbol{\mu}_k$ be the mean vector of the k-th mixture component ($k = 1, \ldots, D$). Then, in both FD and EFD models, cluster k is characterized by a value of the k-th element of $\boldsymbol{\mu}_k$ greater than the corresponding element of $\boldsymbol{\mu}_{k'}$, $k' \neq k$. This peculiarity does not hold anymore for the DFD distributions, since the number of clusters is greater than the dimension of each $\boldsymbol{\mu}_k$. Indeed, the number of mixture components in a DFD model is $D^* = \frac{D(D+1)}{2}$ and this quantity is a quadratic function of D. The next definition provides a matrix useful in identifying clusters:

Definition 22. *A **cluster-code matrix** of order D, $\boldsymbol{C}_D \in \mathcal{M}(D, D)$ is an upper triangular matrix such that:*

- *the main diagonal is composed by the first (ordered) D integers:*

$$\boldsymbol{C}_5 = \begin{bmatrix} 1 & \cdot & \cdot & \cdot & \cdot \\ & 2 & \cdot & \cdot & \cdot \\ & & 3 & \cdot & \cdot \\ & & & 4 & \cdot \\ & & & & 5 \end{bmatrix}$$

- *the remaining elements are equal to the (ordered) integers from $D+1$ to $\dfrac{D(D+1)}{2}$ allocated by row:*

$$\boldsymbol{C}_5 = \begin{bmatrix} 1 & 6 & 7 & 8 & 9 \\ & 2 & \cdot & \cdot & \cdot \\ & & 3 & \cdot & \cdot \\ & & & 4 & \cdot \\ & & & & 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 7 & 8 & 9 \\ & 2 & 10 & 11 & 12 \\ & & 3 & \cdot & \cdot \\ & & & 4 & \cdot \\ & & & & 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 6 & 7 & 8 & 9 \\ & 2 & 10 & 11 & 12 \\ & & 3 & 13 & 14 \\ & & & 4 & 15 \\ & & & & 5 \end{bmatrix}$$

Example 12. *The following are cluster-code matrices of order D = 2, 3, 4 and 5 (0 entries below the main diagonal are omitted):*

$$\mathbf{C}_2 = \begin{bmatrix} 1 & 3 \\ & 2 \end{bmatrix} \qquad \mathbf{C}_3 = \begin{bmatrix} 1 & 4 & 5 \\ & 2 & 6 \\ & & 3 \end{bmatrix}$$

$$\mathbf{C}_4 = \begin{bmatrix} 1 & 5 & 6 & 7 \\ & 2 & 8 & 9 \\ & & 3 & 10 \\ & & & 4 \end{bmatrix} \qquad \mathbf{C}_5 = \begin{bmatrix} 1 & 6 & 7 & 8 & 9 \\ & 2 & 10 & 11 & 12 \\ & & 3 & 13 & 14 \\ & & & 4 & 15 \\ & & & & 5 \end{bmatrix}$$

Given a particular value of D, a cluster-code matrix allows us to identify a particular cluster uniquely. Let k $\left(k = 1, 2, \ldots, \frac{D(D+1)}{2}\right)$ be the cluster label; if $c_{i,j} = k$, then cluster k is the one with parameters $\boldsymbol{\alpha} + \tau(\mathbf{e}_i + \mathbf{e}_j)$ (of course i can be equal to j).

With this cluster structure, it is possibe to rewrite the DFD model as:

$$DFD(\mathbf{x}; \boldsymbol{\alpha}, \tau, \boldsymbol{\pi}) = \sum_{k=1}^{D^*} \pi_k Dir(\mathbf{x}; \boldsymbol{\alpha} + \tau \cdot e(k)),$$

where $D^* = \frac{D(D+1)}{2}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{D^*})$, $e(k) = \sum_{i=1}^{D} \sum_{j \geq i} (\mathbf{e}_i + \mathbf{e}_j) \cdot I(c_{i,j} = k)$ and

$$\pi_k = \begin{cases} p_{k,k} & \text{if } k = 1, \ldots, D \\ 2 \cdot p_{\{i,j\,:\,c_{ij}=k\}} & \text{if } k = D+1, \ldots, D^* \end{cases} .$$

This new notation makes the definition of $\boldsymbol{\mu}_k$ easy:

$$\boldsymbol{\mu}_k = \frac{\boldsymbol{\alpha}}{\alpha^+ + 2\,\tau} + \frac{1}{\alpha^+ + 2\,\tau} \cdot \tau e(k), \qquad (k = 1, \ldots, D^*). \tag{4.30}$$

4.4.2 Parameter estimation: the EM algorithm

Let us assume that a random sample $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\intercal$ of size n has been collected, where each $\mathbf{x}_s$ $(s = 1, \ldots, n)$ is a realization of $\mathbf{X} \sim \text{DFD}(\boldsymbol{\alpha}, \tau, \boldsymbol{\pi})$. To compute the ML estimates of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \tau, \boldsymbol{\pi})^\intercal$ we may use the Expectation-Maximization (EM) algorithm, formalized by Dempster et al. in 1977 [28]. In this context, the EM algorithm is defined to maximize the conditional expectation of the **Complete-data log Likelihood** function:

$$\log L_C(\boldsymbol{\alpha}, \tau, \boldsymbol{\pi} | \mathbf{x}) = \sum_{s=1}^{n} \sum_{k=1}^{D^*} z_{s,k} \left\{ \log \pi_k + \log f_D(\mathbf{x}_s; \boldsymbol{\alpha} + \tau \cdot e(k) \right\}, \tag{4.31}$$

where) $z_{s,k}$ is a component indicator that is equal to 1 if observation $\mathbf{x}_s$ has arisen from the cluster k.

- **E-Step: m-th iteration**

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\log L_c|\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\intercal\right] &= \int \log L_C \cdot f_{\mathbf{Z}|\mathbf{X}} d\mathbf{Z} \\ &= \sum_{k=1}^{D^*}\sum_{s=1}^{n} \pi_k\left(\mathbf{x}_s;\boldsymbol{\theta}^{(m)}\right)\cdot\left\{\log \pi_k^{(m)} + \log f_D\left(\mathbf{x}_s;\boldsymbol{\alpha}^{(m)}+\tau^{(m)}\cdot e(k)\right)\right\}\end{aligned}$$

  In the above expression, the quantity:

$$\pi_k\left(\mathbf{x}_s;\boldsymbol{\theta}\right) = \frac{\pi_k \cdot f_D(\mathbf{x}_s;\boldsymbol{\alpha}+\tau\cdot e(k))}{\sum_{h=1}^{D^*}\pi_h \cdot f_D(\mathbf{x}_s;\boldsymbol{\alpha}+\tau\cdot e(h))}, \qquad k=1,\ldots,D^* \tag{4.32}$$

  represents the posterior probabilities for observation $\mathbf{x}_s$ in component k.

- **M-Step: m-th iteration**

  It consists in maximizing $\mathbb{E}_{\mathbf{Z}}\left[\log L_c|\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\intercal\right]$. It can be shown that $\hat{\pi}_k^{(m+1)} = \frac{1}{n}\sum_{s=1}^{n}\pi_k\left(\mathbf{x}_s;\boldsymbol{\theta}^{(m)}\right)$. Values of $\hat{\boldsymbol{\alpha}}^{(m+1)}$ and $\hat{\tau}^{(m+1)}$ can be found maximizing the classification likelihood with a Newton-Raphson method.

It is well known that the EM algorithm is not robust with respect to the choice of the initial values (in this case w.r.t the choice of $\boldsymbol{\alpha}^{(0)}$, $\tau^{(0)}$ and $\boldsymbol{\pi}^{(0)}$) [15, 29, 70]. For this reason, two initialization procedures have been implemented and compared. Both require a partition of the sample $\mathbf{x}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\intercal$ into $\frac{D(D+1)}{2}$ groups and thus a clustering method. The following algorithms have been compared:

1. k-means clustering based on the entire composition (D components)
2. k-means clustering based on $D-1$ components, varying the left-out dimension
3. k-means clustering based on additive log-ratio tranformation, varying the baseline dimension
4. k-means clustering based on centered log-ratio tranformation
5. hierarchical clustering based on the Aitchison metric defined in section 2.

An exploratory simulation study has highlighted that method 1. works better in most parameter configurations. Although in the DFD context there exist a clear cluster structure, the k-means algorithm (as any clustering method) labels clusters in a

random way. Then, a labelling scheme has been constructed ad hoc to assign the "correct" label to each cluster. Suppose, without loss of generality, that $D = 3$ so that $D^* = 6$. Remembering that the component specific distribution is a $Dir(\boldsymbol{\alpha} + \tau \cdot e(k))$, then the mean vector for each cluster can be expressed as in Table 4.1:

	Component		
Cluster k	$\mu_{k,1}$	$\mu_{k,2}$	$\mu_{k,3}$
1	$\frac{\alpha_1 + 2\tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_2}{\alpha^+ + 2\tau}$	$\frac{\alpha_3}{\alpha^+ + 2\tau}$
2	$\frac{\alpha_1}{\alpha^+ + 2\tau}$	$\frac{\alpha_2 + 2\tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_3}{\alpha^+ + 2\tau}$
3	$\frac{\alpha_1}{\alpha^+ + 2\tau}$	$\frac{\alpha_2}{\alpha^+ + 2\tau}$	$\frac{\alpha_3 + 2\tau}{\alpha^+ + 2\tau}$
4	$\frac{\alpha_1 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_2 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_3}{\alpha^+ + 2\tau}$
5	$\frac{\alpha_1 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_2}{\alpha^+ + 2\tau}$	$\frac{\alpha_3 + \tau}{\alpha^+ + 2\tau}$
6	$\frac{\alpha_1}{\alpha^+ + 2\tau}$	$\frac{\alpha_2 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_3 + \tau}{\alpha^+ + 2\tau}$

Tab. 4.1: Mean Vectors stratified by cluster.

It is easy to note that the highest value of $\mu_{k,i}$ is reached when $k = i$. Hence, the cluster associated to the greatest sample mean $\bar{x}_k$ can be labelled as cluster k. In order to label the remaining clusters, two methods have been proposed. The first one resorts on the stratified covariances (Table 4.2).

	Covariance		
Cluster k	Cov (X_1, X_2)	Cov (X_1, X_3)	Cov (X_2, X_3)
1	$\frac{-(\alpha_1 + 2\tau)\alpha_2}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_1 + 2\tau)\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_2\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$
2	$\frac{-\alpha_1(\alpha_2 + 2\tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_1\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_2 + 2\tau)\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$
3	$\frac{-\alpha_1\alpha_2}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_1(\alpha_3 + 2\tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_2(\alpha_3 + 2\tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$
4	$\frac{-(\alpha_1 + \tau)(\alpha_2 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_1 + \tau)\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_2 + \tau)\alpha_3}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$
5	$\frac{-(\alpha_1 + \tau)\alpha_2}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_1 + \tau)(\alpha_3 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_2(\alpha_3 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$
6	$\frac{-\alpha_1(\alpha_2 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-\alpha_1(\alpha_3 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$	$\frac{-(\alpha_2 + \tau)(\alpha_3 + \tau)}{(\alpha^+ + 2\tau)^2(\alpha^+ + 2\tau + 1)}$

Tab. 4.2: Covariances stratified by cluster.

The focus here is in clusters 4, 5 and 6 ($k = D + 1, \ldots, D^*$): cluster k minimizes the covariance between X_i and X_j, where $i, j \;:\; c_{i,j} = k$. Hence, excluding

clusters $1, \dots, D$ already labelled, it is sufficient to label the group associated with the minimum sample covariance between X_i and X_j as $k = c_{i,j}$.

If a cluster maximizes 2 or more sample means or minimizes 2 or more sample covariance, all label permutations compatible with the observed structure will be considered. This labelling scheme can be generalized also for $D \neq 3$.

This method has a very clear shortcoming: if a cluster contains few observations, the group-specific covariances can be unstable and unreliable. This can be avoided with the second method to provide labels for clusters $D+1, \dots, D^*$ based entirely on the mean vectors. Let $\mathcal{U}_i$ be the following set of indices:

$$\mathcal{U}_i = \{k : c_{j,i} = k,\ j = 1, \dots, i \qquad \& \qquad k : c_{i,j} = k,\ j = i+1, \dots, D\}\,. \tag{4.33}$$

If $k \in \mathcal{U}_i$ then index k is on the i-th row or the i-th column of the cluster code matrix $\mathbf{C}_D$.

Conditioning on cluster $k > D$, it is easy to note that $\mu_{k,i}$ is maximized by those $k \in \mathcal{U}_i \setminus \{i\}$.

	Component		
Cluster k	$\mu_{k,1}$	$\mu_{k,2}$	$\mu_{k,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
4	$\frac{\alpha_1 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_2 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_3}{\alpha^+ + 2\tau}$
5	$\frac{\alpha_1 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_2}{\alpha^+ + 2\tau}$	$\frac{\alpha_3 + \tau}{\alpha^+ + 2\tau}$
6	$\frac{\alpha_1}{\alpha^+ + 2\tau}$	$\frac{\alpha_2 + \tau}{\alpha^+ + 2\tau}$	$\frac{\alpha_3 + \tau}{\alpha^+ + 2\tau}$

Tab. 4.3: Mean Vectors stratified by cluster (clusters $D+1, \dots, D^*$).

Then, the cluster that maximizes $\mu_{\cdot,i}$ and $\mu_{\cdot,j}$ is labelled as $k = c_{i,j}$. Once again, if multiple labelled schemes occour, the estimation procedure is applied to every single label permutation compatible with the observed structure. Also when each cluster has lot of observations, the two approaches provide very similar results. Then the one entirely based on the means has been chosen, due to its simplicity and reliability in a few observations-scenario.

Given a data partition obtained with the above method, an initialization for $\boldsymbol{\pi}$ is the percentage of data points allocated in each cluster:

$$\boldsymbol{\pi}^{(0)} = \left(\pi_1^{(0)}, \ldots, \pi_{D^*}^{(0)}\right)^\intercal, \qquad \text{where } \pi_k^{(0)} = \frac{1}{n}\sum_{s=1}^{n} z_{s,k}.$$

In order to obtain $\boldsymbol{\alpha}^{(0)}$ and $\tau^{(0)}$ the method of moments based on the first D clusters or based on all D^* clusters can be used. Let $\bar{x}_{h,i}$ and $s^2_{h,i}$ be the sample mean and the sample variance of component i among cluster h:

1. D **clusters:** Each component is characterized by a Dirichlet density. Components $1, \ldots, D$ have a similar parametric structure ($\boldsymbol{\alpha} + 2\tau\, \mathbf{e}_k$ for some k). Let $X_{h,i}$ be the $i-$th component of cluster h. Then:

$$\mathbb{E}\left[X_{h,i}\right] = \frac{\alpha_i + 2\tau\, e_{i_h}}{\alpha^+ + 2\tau} \quad \Longrightarrow \quad \widehat{\left(\frac{\alpha_i}{\alpha^+ + 2\tau}\right)} = \frac{\sum_{\substack{h=1 \\ h\neq i}}^{D} \bar{x}_{h,i} \cdot \hat{\pi}_h}{\sum_{\substack{h=1 \\ h\neq i}}^{D} \hat{\pi}_h}.$$

   Since $\dfrac{2\tau}{\alpha^+ + 2\tau} = \mathbb{E}\left[X_{h,h}\right] - \dfrac{\alpha_h}{\alpha^+ + 2\tau}$, then $\widehat{\dfrac{2\tau}{\alpha^+ + 2\tau}}$ can be computed as the weighted mean of the D estimates:

$$\bar{x}_{i,i} - \frac{\sum_{\substack{h=1 \\ h\neq i}}^{D} \bar{x}_{h,i}\hat{\pi}_h}{\sum_{\substack{h=1 \\ h\neq i}}^{D} \hat{\pi}_h}, \qquad i = 1, \ldots, D.$$

   Finally, remembering that $\text{Var}\left(X_{h,i}\right) = \dfrac{\mathbb{E}\left[X_{h,i}\right]\left(1 - \mathbb{E}\left[X_{h,i}\right]\right)}{\alpha^+ + 2\tau + 1}$, initialization of the common denominator $(\alpha^+ + 2\tau)$ can be obtained as the weighted mean of:

$$\frac{1 - \sum_{i=1}^{D} \bar{x}^2_{h,i}}{\sum_{i=1}^{D} s^2_{h,i}} - 1, \qquad h = 1, \ldots, D.$$

   This approach uses only data points assigned to D groups instead of $\frac{D(D+1)}{2}$: the lack of information can be huge! It is possible to modify this algorithm in order to consider all the data points.

2. D^* **clusters:**

Remembering that $D^* = \dfrac{D(D+1)}{2}$, then the algorithm can be expressed as:

a) Initialize $\widehat{\left(\dfrac{\alpha_i}{\alpha^+ + 2\tau}\right)} = \dfrac{\sum_{\substack{h=1 \\ h \notin \mathcal{U}_i}}^{D^*} \bar{x}_{h,i} \cdot \hat{\pi}_h}{\sum_{\substack{h=1 \\ h \notin \mathcal{U}_i}}^{D^*} \hat{\pi}_h}$, where $\mathcal{U}_i$ is defined in (4.33)

b) Given that

$$\left(\frac{2\tau}{\alpha^+ + 2\tau}\right) = \begin{cases} \left(\dfrac{\alpha_h + 2\tau}{\alpha^+ + 2\tau}\right) - \left(\dfrac{\alpha_h}{\alpha^+ + 2\tau}\right), & \text{if } h = 1, \ldots, D \\ \left(\dfrac{\alpha_l + \tau}{\alpha^+ + 2\tau}\right) - \left(\dfrac{\alpha_l}{\alpha^+ + 2\tau}\right) + \left(\dfrac{\alpha_w + \tau}{\alpha^+ + 2\tau}\right) - \left(\dfrac{\alpha_w}{\alpha^+ + 2\tau}\right), & \text{if } h = D+1, \ldots, D^* \end{cases}$$

where l and w are two indices such that $c_{l,w} = h$ or $c_{w,l} = h$, then the inizialization $\widehat{\left(\dfrac{2\tau}{\alpha^+ + 2\tau}\right)}$ is the weighted mean of the $\dfrac{D(D+1)}{2}$ quantities:

$$\begin{cases} \bar{x}_{h,h} - \widehat{\left(\dfrac{\alpha_h}{\alpha^+ + 2\tau}\right)}, & \text{if } h = 1, \ldots, D \\ \bar{x}_{h,l} - \widehat{\left(\dfrac{\alpha_l}{\alpha^+ + 2\tau}\right)} + \bar{x}_{h,w} - \widehat{\left(\dfrac{\alpha_w}{\alpha^+ + 2\tau}\right)}, & \text{if } h = D+1, \ldots, D^* \end{cases}$$

c) The quantity $(\alpha^+ + 2\tau)$ can be estimated as the weighted mean of:

$$\frac{1 - \sum_{i=1}^{D} \bar{x}_{h,i}^2}{\sum_{i=1}^{D} s_{h,i}^2} - 1, \qquad h = 1, \ldots, \frac{D(D+1)}{2}.$$

This second method is preferred, since it uses all the collected data. Table 4.4 reports the means of 500 initializations for $\boldsymbol{\alpha}$ and τ. These initializations have been obtained with samples of size 300 generated from the parameters configuration represented by column ID (see Table 4.5).

4.4.3 Simulation study

In order to study the characteristics of the DFD and the performances of the proposed EM algorithm estimation procedure, three simulation studies have been implemented. Each of these simulations refers to the nine parametric configurations reported in Table 4.5:

	α_1	α_2	α_3	τ
True	10	10	10	10
Init.	9.427	9.286	9.526	9.703
True	10	10	10	40
Init.	9.706	9.397	9.874	37.745
True	2	23	12	17
Init.	1.888	20.615	10.861	15.836
True	100	40	40	15
Init.	92.450	37.019	37.261	13.209
True	10	100	14	8
Init.	10.034	98.212	13.570	7.800
True	12	0.900	30	20
Init.	12.478	0.892	32.036	21.476

Tab. 4.4: Mean of 500 initialializations for $\boldsymbol{\alpha}$ and τ in different configurations of parameters.

ID	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
1	10	10	10	15	0.11	0.11	0.11	0.22	0.22	0.22
2	10	10	10	40	0.11	0.11	0.11	0.22	0.22	0.22
3	2	23	12	17	0.08	0.16	0.18	0.10	0.40	0.08
4	40	20	30	25	0.00	0.16	0.26	0.40	0.00	0.18
5	40	20	30	50	0.00	0.16	0.26	0.40	0.00	0.18
6	100	40	40	15	0.22	0.17	0.15	0.15	0.10	0.20
7	40	20	30	18	0.00	0.00	0.00	0.30	0.19	0.51
8	10	100	14	8	0.10	0.15	0.15	0.10	0.40	0.10
9	12	0.90	30	20	0.08	0.16	0.18	0.10	0.40	0.08

Tab. 4.5: Parameter configurations for all the DFD simulations.

These configurations allow to cover several scenarios: well separated as well as overlapping clusters, clusters very closed to one edge of the simplex, positive and negative correlations and configurations where not all the components are present. Some of these features can be inspected graphically by looking at Figures 4.6 - 4.14.

Alternative EM methods

The EM algorithm is one of the most popular algorithm used to obtain Maximum Likelihood estimates in a missing data scenario. It is very popular but it has an important disadvantage: it is very sensitive to initial values. For this reason several authors proposed alternative versions of the EM algorithm [15, 21, 22]: two of them are the Classification EM (CEM) and the Stochastic EM (SEM). The CEM has a further classification step: at step m, each observation is allocated to the group k maximizing $\hat{\pi}_k^{(m)}(\mathbf{x}_s; \boldsymbol{\theta})$. The SEM algorithm has a similar approach: at each step,

new partition for observation $\mathbf{x}_s$ is generated according to a Multinomial distribution with probabilities $(\hat{\pi}_1(\mathbf{x}_s;\boldsymbol{\theta}),\ldots,\hat{\pi}_{D^*}(\mathbf{x}_s;\boldsymbol{\theta}))^\intercal$, defined in (4.32).

The simulation study is composed by the following steps:

- For each configuration of parameters, 100 samples of size 100 have been generated.
- For each sample, initialization of the parameters have been obtained according to the method described previously.
- ML estimates have been obtained with these algorithms:

  1) EM
  2) CEM
  3) SEM
  4) CEM + EM
  5) SEM + EM

Table 4.6 reports the proportion of simulations where each method provided the highest log-likelihood and the mean of the log-likelihoods evaluated at the obtained initial parameters. From these results one can conclude that the SEM + EM combination is the one providing the best values in most cases. The presence of a EM step is fundamental: the CEM and the SEM are not able to find the global maximizer by themselves (look at columns "%" for the CEM and SEM methods).

	EM		CEM		SEM		CEM+EM		SEM+EM	
ID	%	Mean $\hat{l}$	%	Mean $\hat{l}$	%	Mean $\hat{l}$	%	Mean $\hat{l}$	%	Mean $\hat{l}$
1	0.287	128.6434	0	104.5987	0	104.5987	0.330	128.6434	0.383	129.5926
2	0.005	191.0407	0	189.1948	0	189.1948	0.000	191.0407	0.995	191.0408
3	0.285	169.1768	0	166.2929	0	166.2929	0.278	169.1768	0.437	169.1559
4	0.280	220.7419	0	211.3169	0	211.3169	0.330	220.7419	0.390	220.7137
5	0.018	261.2633	0	251.4802	0	251.4802	0.028	261.2633	0.953	261.2633
6	0.358	307.6959	0	279.7145	0	279.7145	0.325	307.6959	0.317	307.6959
7	0.337	216.5530	0	184.9596	0	184.9596	0.447	218.6098	0.217	216.2881
8	0.318	344.0424	0	305.5102	0	305.5102	0.330	344.0424	0.352	344.0383
9	0.295	260.0346	0	258.3536	0	258.3536	0.270	260.0346	0.435	260.0347

Tab. 4.6: DFD initialization simulation results.

Models fitting

The Double Flexible Dirichlet has a very particular clusters structure. This simulation study is aimed at evaluating if features of the DFD model can be catched by simpler models. For this reason, for each of the nine configurations showed in Table 4.5, 50 samples of size 150 have been generated and, for each of them, the Dirichlet, ALN, FD, EFD and the DFD's parameters have been estimated.
Table 4.7 shows the mean of the resulting AIC (Akaike Information Criterion) [6] and BIC (Bayesian Information Criterion)[82]. In 7 scenarios out of 9 the DFD results the best model (i.e. the one with the best penalized fit to simulated data). Configurations 6 and 8 seem to favor simpler models. The reason why this happens relies in the cluster structure: looking at Figures 4.11 and 4.13 it is possible to note that these configurations are characterized by very closed and overlapped clusters. Increasing the sample size to 500 (Table 4.8), the superiority of the DFD model is confirmed in every scenario.

ID	Crit.	Dir	ALN	FD	EFD	DFD
1	AIC	-382.322	-392.230	-393.002	-389.306	-414.572
	BIC	-373.289	-377.177	-374.938	-365.221	-387.476
2	AIC	-206.711	-236.967	-215.968	-220.995	-665.674
	BIC	-197.679	-221.914	-197.904	-196.910	-638.579
3	AIC	-364.447	-371.644	-442.615	-490.375	-607.278
	BIC	-355.415	-356.591	-424.551	-466.290	-580.183
4	AIC	-463.938	-541.371	-553.536	-604.246	-746.003
	BIC	-454.906	-526.318	-535.472	-580.161	-718.907
5	AIC	-295.221	-396.091	-400.039	-473.547	-862.729
	BIC	-286.189	-381.038	-381.975	-449.462	-835.634
6	AIC	-887.023	-889.846	-916.720	-918.477	-908.828
	BIC	-877.992	-874.793	-898.656	-894.392	-881.732
7	AIC	-720.475	-707.645	-718.769	-715.057	-801.165
	BIC	-711.443	-692.592	-700.706	-690.972	-774.069
8	AIC	-1032.602	-1023.258	-1027.488	-1030.226	-1013.702
	BIC	-1023.570	-1008.205	-1009.424	-1006.141	-986.606
9	AIC	-621.435	-715.025	-747.292	-790.366	-905.478
	BIC	-612.404	-699.972	-729.228	-766.281	-878.382

Tab. 4.7: Mean of the AIC and BIC for the simulation with $n = 150$.

The proposed models have also been compared in situations where the data generating process is not the DFD one. In particular, we have chosen four parameter configurations for the ALN distribution (Table 4.9) and four configurations for the EFD (Table 4.11). We have generated 50 samples of size $n = 500$ from each parameter configuration and fitted the models. Tables 4.10 and 4.12 show the mean of the corresponding 50 AICs and BICs.

ID	Crit.	Dir	ALN	FD	EFD	DFD
1	AIC	-742.961	-825.490	-793.643	-790.542	-1978.257
	BIC	-730.317	-804.417	-768.355	-756.825	-1940.325
2	AIC	-740.776	-820.231	-781.488	-779.348	-2021.253
	BIC	-728.132	-799.158	-756.201	-745.632	-1983.322
3	AIC	-727.551	-820.167	-785.455	-783.864	-1998.404
	BIC	-714.907	-799.094	-760.167	-750.147	-1960.472
4	AIC	-734.673	-819.060	-783.541	-780.214	-2002.786
	BIC	-722.029	-797.986	-758.253	-746.497	-1964.854
5	AIC	-745.689	-828.971	-788.784	-787.272	-2071.198
	BIC	-733.045	-807.898	-763.496	-753.555	-2033.267
6	AIC	-734.744	-821.813	-784.915	-781.759	-2052.858
	BIC	-722.100	-800.740	-759.628	-748.043	-2014.927
7	AIC	-735.851	-827.519	-792.128	-791.333	-2059.012
	BIC	-723.207	-806.446	-766.840	-757.616	-2021.081
8	AIC	-741.654	-821.791	-786.221	-782.484	-2031.823
	BIC	-729.010	-800.718	-760.934	-748.767	-1993.892
9	AIC	-728.757	-823.881	-790.891	-789.905	-1956.915
	BIC	-716.113	-802.808	-765.603	-756.188	-1918.983

Tab. 4.8: Mean of the AIC and BIC for the simulation with $n = 500$.

Please note that, despite the DFD does not perform well, the FD and the EFD models allow for a good fit to the data. The DFD is penalized by the high number of parameters it involves and by the data structure (data generated from an ALN do not show a cluster structure, therefore the DFD is an unnecessarily complicated model).

Scenario	μ_1	μ_2	σ_1^2	σ_2^2	$\sigma_{1,2}$
1	0	0	1.3	1.3	0.65
2	0	0	0.21	0.21	0.11
3	1	1.22	7.27	6.83	4.93
4	-1.89	0	2.73	0.79	0.39

Tab. 4.9: Parameter configurations for the ALN simulations.

ID	Crit.	Dir	ALN	FD	EFD	DFD
1	AIC	-892.941	-919.299	-908.689	-908.160	-902.692
	BIC	-880.297	-898.226	-883.401	-874.443	-864.761
2	AIC	-2259.928	-2263.117	-2260.895	-2259.757	-2254.908
	BIC	-2247.284	-2242.044	-2235.607	-2226.040	-2216.977
3	AIC	-1114.436	-1216.120	-1161.433	-1169.229	-1155.437
	BIC	-1101.792	-1195.047	-1136.146	-1135.512	-1117.506
4	AIC	-1445.646	-1538.721	-1479.798	-1502.585	-1473.805
	BIC	-1433.002	-1517.648	-1454.511	-1468.868	-1435.874

Tab. 4.10: Mean of the AIC and the BIC for simulation with $n = 500$. Data generated from ALN distributions.

Scenario	α_1	α_2	α_3	τ	τ	τ	p_1	p_2	p_3
1	10	10	10	30	2	20	0.3	0.2	0.5
2	10	5	30	5	8	32	0.25	0.4	0.35
3	5	13	5	15	15	5	0.25	0.4	0.35
4	20	20	3	10	10	2	1/3	1/3	1/3

Tab. 4.11: Parameter configurations for the EFD simulations.

ID	Crit.	Dir	ALN	FD	EFD	DFD
1	AIC	-884.118	-1208.446	-1317.289	-1584.484	-1450.301
	BIC	-871.474	-1188.489	-1293.340	-1552.553	-1414.378
2	AIC	-1734.925	-1817.169	-1864.653	-1990.095	-1858.658
	BIC	-1722.281	-1797.211	-1840.705	-1958.163	-1822.735
3	AIC	-1077.630	-1180.338	-1271.071	-1400.606	-1265.074
	BIC	-1064.987	-1160.381	-1247.122	-1368.674	-1229.150
4	AIC	-2136.472	-2141.494	-2155.695	-2172.745	-2149.697
	BIC	-2123.828	-2121.536	-2131.746	-2140.814	-2113.774

Tab. 4.12: Mean of the AIC and BIC for the simulation with $n = 500$. Data generated from EFD distributions.

EM algorithm performance evaluation

The last simulation study regards the evaluation of the performance of the ML estimator. For each configuration reported in Table 4.5, 1000 samples of size $n = 150$ have been generated. For each of them, the parameters of the DFD model have been estimated according to the estimation and initialization procedures described in Section 4.4.2. In the following part it is possible to find:

1. The ternary diagram (and its zoomed version) with the true density function (represented as isodensity contour plot).

2. A table reporting the symmetrized Kullback-Leibler divergence measure among clusters.

3. The DFD's correlation matrix, evaluated at the true parameters.

4. A table reporting the results of the simulation for that particular scenario. This table contains:

   a) The true value of the parameters.

   b) The mean of the 1000 estimates for each parameter.

   c) The median of the 1000 estimates for each parameter.

d) The Absolute Relative Bias (Arb), defined as:

$$\text{Arb} = \frac{1}{1000} \sum_{i=1}^{1000} \frac{\left|\hat{\theta}_i - \theta\right|}{\theta}, \tag{4.34}$$

where $\hat{\theta}_i$ is the estimates of θ obtained from the $i-$th sample.

e) The Mean Squared Error (MSE), defined as:

$$\text{MSE} = \frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{\theta}_i - \theta\right)^2. \tag{4.35}$$

f) The standard deviation of the 1000 estimates for each parameter. This quantity can be viewed as the boostrap approximation of the Standard Error of the estimator and, therefore, it is called "Boot. SE".

g) The coverage of the approximated 95% confidence intervals (CI), that is the percentage of times that the approximated 95% CI cointains the true value of the parameter. An approximated $(1-\alpha)$% confidence interval is computed as: $\hat{\theta} \pm z_{1-\alpha/2} \cdot \text{SE}_{Boot.}$.

In general, it is reported that estimating parameters of a finite mixture model through the EM algorithm can encounter several issues, particularly when the sample size is small [38, 60]. In the considered simulations, the relatively small sample size (fixed and equal to $n = 150$) seems to be large enough to produce very good results. In most of scenarios, the covarage level of approximated confidence intervals is very close to the 95% nominal one. Furthermore, the similarity between MLEs mean and MLEs median can be interpreted as an evidence of convergence of the Maximum Likelihood estimator to the Normal distribution, since it is a necessary but not sufficient condition.
It is important to note that scenario 7 represents a configuration of clusters' barycenters that both the FD and the EFD are not capable to recognize (i.e. they form an inverse triangle). Also scenarios 4 and 5 are characterized by a typically DFD configuration of clusters: with two weights equal to zero ($\pi_1 = \pi_5 = 0$), joining the cluster's means produces an oblique and rotate "L". Scenario 9 has 3 clusters very close to one of the edges of the simplex; this means that part of the data have at least one component close to 0. This can be a problem in compositional data analysis, since most density functions are not defined at the boundary of the simplex. The results confirm that scenarios 6 and 8 are quite challenging, since the EM algorithm is not capable of recognize the true number of clusters. Of course, the reason why the EM algorithm fails in providing good estimates of almost all the parameters is entirely due to the overlapping of clusters.

1st Scenario

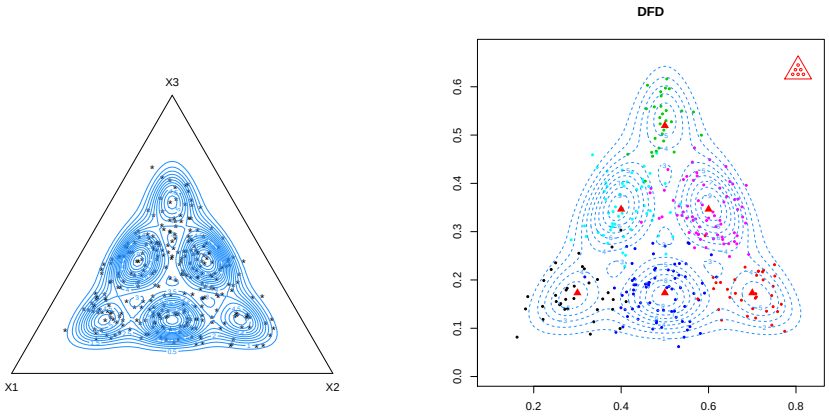

Fig. 4.6: DFD simulation - 1st configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	21.3686
2	$d_{SKL}(f_{1,1}, f_{1,3})$	21.3686
3	$d_{SKL}(f_{1,1}, f_{2,2})$	42.7372
4	$d_{SKL}(f_{1,1}, f_{2,3})$	49.7783
5	$d_{SKL}(f_{1,1}, f_{3,3})$	42.7372
6	$d_{SKL}(f_{1,2}, f_{1,3})$	28.4097
7	$d_{SKL}(f_{1,2}, f_{2,2})$	21.3686
8	$d_{SKL}(f_{1,2}, f_{2,3})$	28.4097
9	$d_{SKL}(f_{1,2}, f_{3,3})$	49.7783
10	$d_{SKL}(f_{1,3}, f_{2,2})$	49.7783
11	$d_{SKL}(f_{1,3}, f_{2,3})$	28.4097
12	$d_{SKL}(f_{1,3}, f_{3,3})$	21.3686
13	$d_{SKL}(f_{2,2}, f_{2,3})$	21.3686
14	$d_{SKL}(f_{2,2}, f_{3,3})$	42.7372
15	$d_{SKL}(f_{2,3}, f_{3,3})$	21.3686

Tab. 4.13: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.50	-0.50
X_2	-0.50	1.00	-0.50
X_3	-0.50	-0.50	1.00

Tab. 4.14: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	10	10	10	15	0.111	0.111	0.111	0.222	0.222	0.222
MLE Mean	10.194	10.202	10.196	15.071	0.116	0.117	0.117	0.219	0.215	0.217
MLE Median	10.128	10.137	10.159	15.010	0.114	0.116	0.115	0.218	0.215	0.217
Arb	0.019	0.020	0.020	0.005	0.046	0.049	0.050	0.016	0.033	0.023
MSE	0.870	0.875	0.897	2.026	0.001	0.001	0.001	0.001	0.001	0.001
Boot. SE	0.913	0.914	0.927	1.422	0.028	0.028	0.028	0.035	0.033	0.036
Coverage	0.937	0.934	0.942	0.948	0.945	0.944	0.940	0.949	0.945	0.942

Tab. 4.15: Simulation results.

2nd Scenario 2

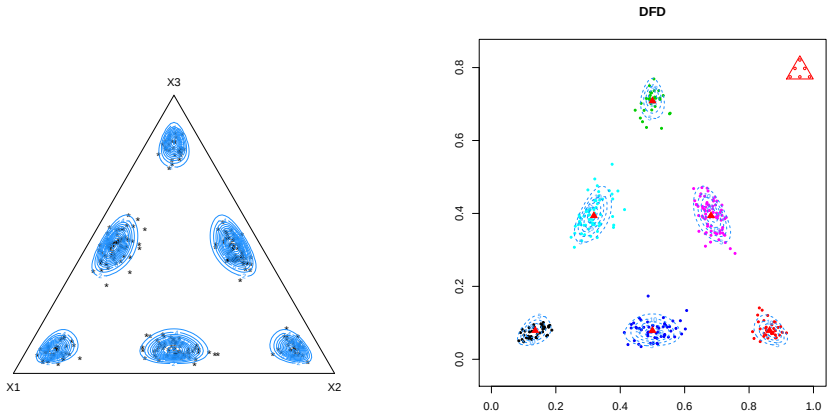

Fig. 4.7: DFD simulation - 2nd configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	89.6996
2	$d_{SKL}(f_{1,1}, f_{1,3})$	89.6996
3	$d_{SKL}(f_{1,1}, f_{2,2})$	179.3993
4	$d_{SKL}(f_{1,1}, f_{2,3})$	221.7186
5	$d_{SKL}(f_{1,1}, f_{3,3})$	179.3993
6	$d_{SKL}(f_{1,2}, f_{1,3})$	132.019
7	$d_{SKL}(f_{1,2}, f_{2,2})$	89.6996
8	$d_{SKL}(f_{1,2}, f_{2,3})$	132.019
9	$d_{SKL}(f_{1,2}, f_{3,3})$	221.7186
10	$d_{SKL}(f_{1,3}, f_{2,2})$	221.7186
11	$d_{SKL}(f_{1,3}, f_{2,3})$	132.019
12	$d_{SKL}(f_{1,3}, f_{3,3})$	89.6996
13	$d_{SKL}(f_{2,2}, f_{2,3})$	89.6996
14	$d_{SKL}(f_{2,2}, f_{3,3})$	179.3993
15	$d_{SKL}(f_{2,3}, f_{3,3})$	89.6996

Tab. 4.16: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.50	-0.50
X_2	-0.50	1.00	-0.50
X_3	-0.50	-0.50	1.00

Tab. 4.17: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	10	10	10	40	0.111	0.111	0.111	0.222	0.222	0.222
MLE Mean	10.153	10.163	10.156	40.575	0.112	0.112	0.112	0.223	0.219	0.222
MLE Median	10.104	10.113	10.116	40.414	0.107	0.113	0.107	0.220	0.220	0.220
Arb	0.015	0.016	0.016	0.014	0.004	0.008	0.006	0.005	0.014	0.001
MSE	0.747	0.749	0.770	11.093	0.001	0.001	0.001	0.001	0.001	0.001
Boot. SE	0.851	0.850	0.864	3.282	0.026	0.026	0.027	0.034	0.032	0.035
Coverage	0.934	0.936	0.949	0.942	0.959	0.960	0.955	0.955	0.951	0.942

Tab. 4.18: Simulation results.

3rd Scenario

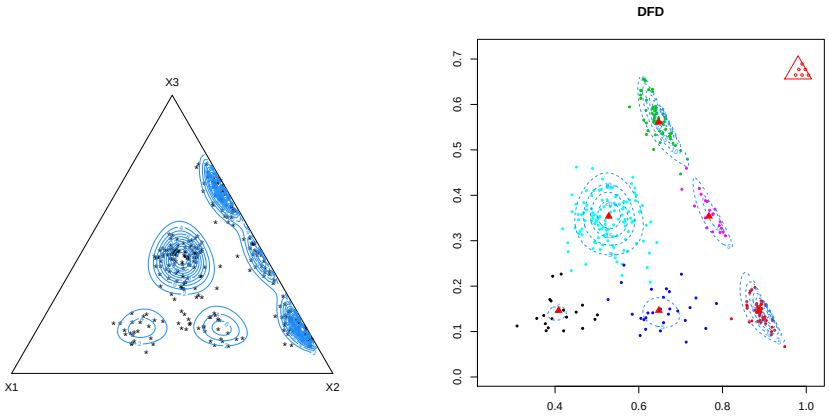

Fig. 4.8: DFD simulation - 3rd configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	20.6449
2	$d_{SKL}(f_{1,1}, f_{1,3})$	26.5024
3	$d_{SKL}(f_{1,1}, f_{2,2})$	69.1464
4	$d_{SKL}(f_{1,1}, f_{2,3})$	78.4857
5	$d_{SKL}(f_{1,1}, f_{3,3})$	76.8715
6	$d_{SKL}(f_{1,2}, f_{1,3})$	24.9904
7	$d_{SKL}(f_{1,2}, f_{2,2})$	48.5016
8	$d_{SKL}(f_{1,2}, f_{2,3})$	57.8408
9	$d_{SKL}(f_{1,2}, f_{3,3})$	75.3594
10	$d_{SKL}(f_{1,3}, f_{2,2})$	73.492
11	$d_{SKL}(f_{1,3}, f_{2,3})$	51.9832
12	$d_{SKL}(f_{1,3}, f_{3,3})$	50.369
13	$d_{SKL}(f_{2,2}, f_{2,3})$	21.5087
14	$d_{SKL}(f_{2,2}, f_{3,3})$	39.0274
15	$d_{SKL}(f_{2,3}, f_{3,3})$	17.5186

Tab. 4.19: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.49	-0.36
X_2	-0.49	1.00	-0.64
X_3	-0.36	-0.64	1.00

Tab. 4.20: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	2	23	12	17	0.080	0.160	0.180	0.100	0.400	0.080
MLE Mean	2.019	23.269	12.151	17.083	0.083	0.163	0.184	0.098	0.397	0.076
MLE Median	2.003	23.142	12.111	17.001	0.083	0.162	0.182	0.098	0.399	0.077
Arb	0.010	0.012	0.013	0.005	0.032	0.016	0.020	0.020	0.008	0.044
MSE	0.043	4.156	1.213	2.461	0.001	0.001	0.001	0.001	0.002	0.0004
Boot. SE	0.206	2.022	1.091	1.567	0.024	0.030	0.033	0.025	0.040	0.022
Coverage	0.946	0.944	0.947	0.946	0.962	0.948	0.941	0.945	0.945	0.948

Tab. 4.21: Simulation results.

4th Scenario

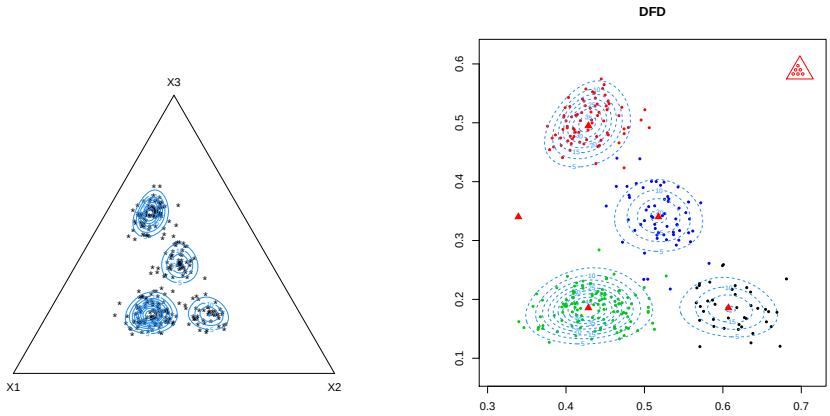

Fig. 4.9: DFD simulation - 4th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	28.8139
2	$d_{SKL}(f_{1,1}, f_{1,3})$	23.5336
3	$d_{SKL}(f_{1,1}, f_{2,2})$	52.2182
4	$d_{SKL}(f_{1,1}, f_{2,3})$	56.417
5	$d_{SKL}(f_{1,1}, f_{3,3})$	45.231
6	$d_{SKL}(f_{1,2}, f_{1,3})$	35.9691
7	$d_{SKL}(f_{1,2}, f_{2,2})$	23.4043
8	$d_{SKL}(f_{1,2}, f_{2,3})$	27.6031
9	$d_{SKL}(f_{1,2}, f_{3,3})$	57.6665
10	$d_{SKL}(f_{1,3}, f_{2,2})$	59.3734
11	$d_{SKL}(f_{1,3}, f_{2,3})$	32.8834
12	$d_{SKL}(f_{1,3}, f_{3,3})$	21.6974
13	$d_{SKL}(f_{2,2}, f_{2,3})$	26.49
14	$d_{SKL}(f_{2,2}, f_{3,3})$	56.5534
15	$d_{SKL}(f_{2,3}, f_{3,3})$	30.0634

Tab. 4.22: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	0.04	-0.64
X_2	0.04	1.00	-0.79
X_3	-0.64	-0.79	1.00

Tab. 4.23: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	40	20	30	25	0	0.160	0.260	0.400	0	0.180
MLE Mean	40.802	20.380	30.587	25.385	0	0.162	0.261	0.397	0	0.180
MLE Median	40.678	20.273	30.513	25.267	0	0.162	0.262	0.397	0	0.178
Arb	0.020	0.019	0.020	0.015	-	0.015	0.004	0.007	-	0.003
MSE	11.163	3.088	6.472	4.612	0	0.001	0.001	0.002	0	0.001
Boot. SE	3.245	1.717	2.477	2.114	0	0.031	0.036	0.041	0.001	0.032
Coverage	0.935	0.930	0.935	0.943	0.996	0.956	0.947	0.953	0.994	0.954

Tab. 4.24: Simulation results.

5th Scenario

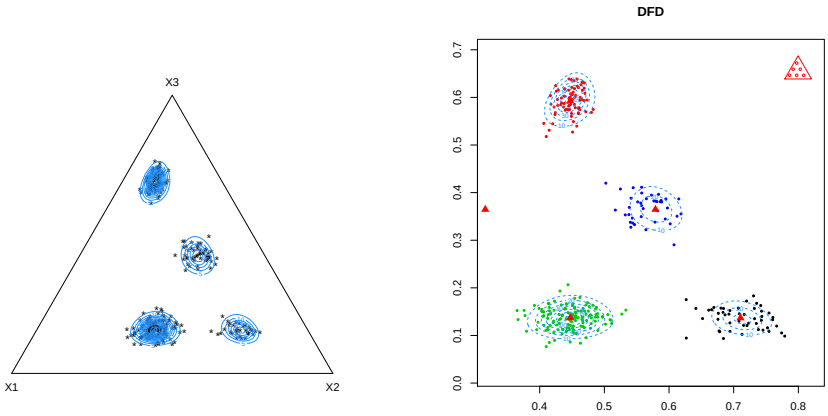

Fig. 4.10: DFD simulation - 5th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	85.7317
2	$d_{SKL}(f_{1,1}, f_{1,3})$	71.7574
3	$d_{SKL}(f_{1,1}, f_{2,2})$	153.7267
4	$d_{SKL}(f_{1,1}, f_{2,3})$	176.1938
5	$d_{SKL}(f_{1,1}, f_{3,3})$	137.0492
6	$d_{SKL}(f_{1,2}, f_{1,3})$	113.1068
7	$d_{SKL}(f_{1,2}, f_{2,2})$	67.995
8	$d_{SKL}(f_{1,2}, f_{2,3})$	90.4621
9	$d_{SKL}(f_{1,2}, f_{3,3})$	178.3987
10	$d_{SKL}(f_{1,3}, f_{2,2})$	181.1019
11	$d_{SKL}(f_{1,3}, f_{2,3})$	104.4364
12	$d_{SKL}(f_{1,3}, f_{3,3})$	65.2918
13	$d_{SKL}(f_{2,2}, f_{2,3})$	76.6655
14	$d_{SKL}(f_{2,2}, f_{3,3})$	164.602
15	$d_{SKL}(f_{2,3}, f_{3,3})$	87.9366

Tab. 4.25: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	0.10	-0.66
X_2	0.10	1.00	-0.81
X_3	-0.66	-0.81	1.00

Tab. 4.26: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	40	20	30	50	0	0.160	0.260	0.400	0	0.180
MLE Mean	40.689	20.387	30.529	50.837	0	0.159	0.260	0.400	0	0.181
MLE Median	40.500	20.293	30.385	50.634	0	0.160	0.260	0.400	0	0.180
Arb	0.017	0.019	0.018	0.017	-	0.008	0	0.001	-	0.008
MSE	10.399	2.998	6.142	16.966	0	0.001	0.001	0.002	0	0.001
Boot. SE	3.152	1.689	2.423	4.035	0	0.030	0.036	0.041	0	0.032
Coverage	0.951	0.945	0.938	0.943	1	0.951	0.945	0.970	1	0.949

Tab. 4.27: Simulation results.

6th Scenario

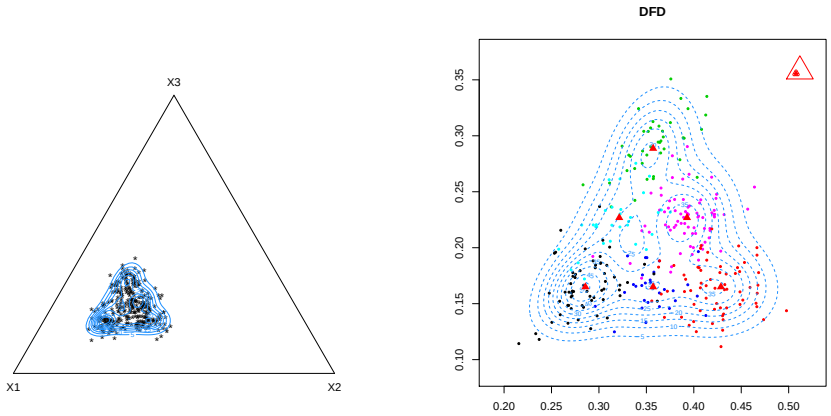

Fig. 4.11: DFD simulation - 6th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	6.6749
2	$d_{SKL}(f_{1,1}, f_{1,3})$	6.6749
3	$d_{SKL}(f_{1,1}, f_{2,2})$	12.4279
4	$d_{SKL}(f_{1,1}, f_{2,3})$	13.6094
5	$d_{SKL}(f_{1,1}, f_{3,3})$	12.4279
6	$d_{SKL}(f_{1,2}, f_{1,3})$	9.6566
7	$d_{SKL}(f_{1,2}, f_{2,2})$	5.7531
8	$d_{SKL}(f_{1,2}, f_{2,3})$	6.9346
9	$d_{SKL}(f_{1,2}, f_{3,3})$	15.4097
10	$d_{SKL}(f_{1,3}, f_{2,2})$	15.4097
11	$d_{SKL}(f_{1,3}, f_{2,3})$	6.9346
12	$d_{SKL}(f_{1,3}, f_{3,3})$	5.7531
13	$d_{SKL}(f_{2,2}, f_{2,3})$	8.4751
14	$d_{SKL}(f_{2,2}, f_{3,3})$	16.9502
15	$d_{SKL}(f_{2,3}, f_{3,3})$	8.4751

Tab. 4.28: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.56	-0.55
X_2	-0.56	1.00	-0.39
X_3	-0.55	-0.39	1.00

Tab. 4.29: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	100	40	40	15	0.220	0.175	0.155	0.150	0.100	0.200
MLE Mean	64.396	26.888	26.664	6.769	0.344	0.353	0.303	0	0	0
MLE Median	64.170	26.854	26.590	6.735	0.344	0.354	0.301	0	0	0
Arb	0.356	0.328	0.333	0.549	0.565	1.017	0.953	1	1	1
MSE	1301.877	177.863	183.370	68.403	0.018	0.035	0.024	0.022	0.010	0.040
Boot. SE	5.854	2.438	2.349	0.805	0.052	0.060	0.052	0	0	0
Coverage	0	0.001	0.001	0	0.327	0.146	0.193	0	0	0

Tab. 4.30: Simulation results.

7th Scenario

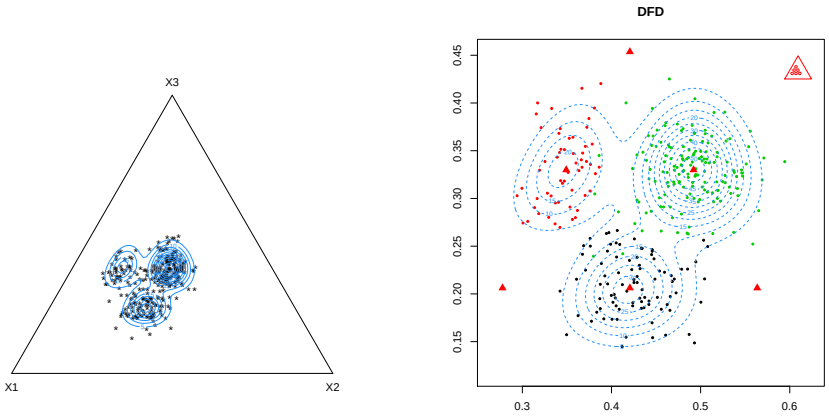

Fig. 4.12: DFD simulation - 7th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	16.6714
2	$d_{SKL}(f_{1,1}, f_{1,3})$	13.4757
3	$d_{SKL}(f_{1,1}, f_{2,2})$	30.4863
4	$d_{SKL}(f_{1,1}, f_{2,3})$	32.0034
5	$d_{SKL}(f_{1,1}, f_{3,3})$	26.0178
6	$d_{SKL}(f_{1,2}, f_{1,3})$	20.3428
7	$d_{SKL}(f_{1,2}, f_{2,2})$	13.8149
8	$d_{SKL}(f_{1,2}, f_{2,3})$	15.332
9	$d_{SKL}(f_{1,2}, f_{3,3})$	32.8849
10	$d_{SKL}(f_{1,3}, f_{2,2})$	34.1577
11	$d_{SKL}(f_{1,3}, f_{2,3})$	18.5277
12	$d_{SKL}(f_{1,3}, f_{3,3})$	12.5421
13	$d_{SKL}(f_{2,2}, f_{2,3})$	15.63
14	$d_{SKL}(f_{2,2}, f_{3,3})$	33.1829
15	$d_{SKL}(f_{2,3}, f_{3,3})$	17.5528

Tab. 4.31: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.49	-0.64
X_2	-0.49	1.00	-0.36
X_3	-0.64	-0.36	1.00

Tab. 4.32: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	40	20	30	18	0	0	0	0.300	0.190	0.510
MLE Mean	38.082	19.645	28.604	16.443	0.043	0.014	0.082	0.255	0.168	0.437
MLE Median	40.455	20.175	30.365	18.198	0	0	0	0.287	0.186	0.497
Arb	0.048	0.018	0.047	0.087	-	-	-	0.149	0.118	0.143
MSE	83.515	10.784	45.855	34.529	0.015	0.011	0.055	0.014	0.006	0.037
Boot. SE	8.940	3.266	6.630	5.669	0.115	0.102	0.221	0.108	0.073	0.179
Coverage	0.867	0.925	0.866	0.863	0.885	0.984	0.881	0.863	0.863	0.863

Tab. 4.33: Simulation results.

8th Scenario

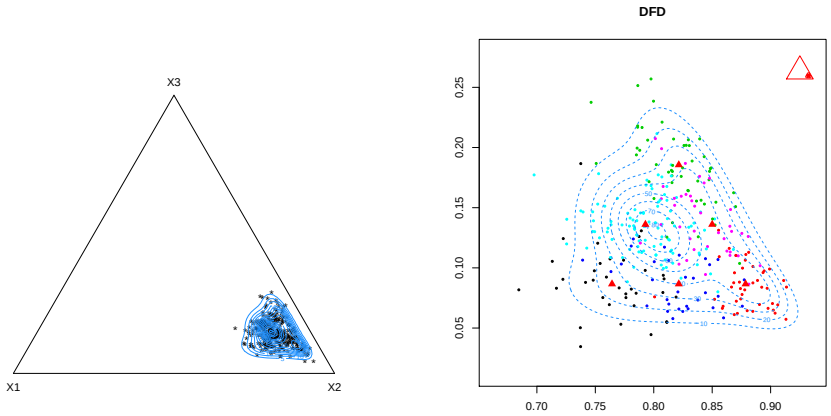

Fig. 4.13: DFD simulation - 8th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	3.6299
2	$d_{SKL}(f_{1,1}, f_{1,3})$	6.733
3	$d_{SKL}(f_{1,1}, f_{2,2})$	9.0888
4	$d_{SKL}(f_{1,1}, f_{2,3})$	12.2364
5	$d_{SKL}(f_{1,1}, f_{3,3})$	14.1481
6	$d_{SKL}(f_{1,2}, f_{1,3})$	4.3405
7	$d_{SKL}(f_{1,2}, f_{2,2})$	5.4589
8	$d_{SKL}(f_{1,2}, f_{2,3})$	8.6065
9	$d_{SKL}(f_{1,2}, f_{3,3})$	11.7555
10	$d_{SKL}(f_{1,3}, f_{2,2})$	9.7994
11	$d_{SKL}(f_{1,3}, f_{2,3})$	5.5033
12	$d_{SKL}(f_{1,3}, f_{3,3})$	7.415
13	$d_{SKL}(f_{2,2}, f_{2,3})$	4.296
14	$d_{SKL}(f_{2,2}, f_{3,3})$	7.4451
15	$d_{SKL}(f_{2,3}, f_{3,3})$	3.149

Tab. 4.34: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.55	-0.29
X_2	-0.55	1.00	-0.64
X_3	-0.29	-0.64	1.00

Tab. 4.35: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	10	100	14	8	0.100	0.150	0.150	0.100	0.400	0.100
MLE Mean	7.334	62.061	10.574	2.784	0.396	0.270	0.334	0	0	0
MLE Median	7.297	61.563	10.484	2.746	0.398	0.267	0.325	0	0	0
Arb	0.267	0.379	0.245	0.652	2.963	0.797	1.227	1	1	1
MSE	7.912	1493.676	13.769	27.477	0.104	0.019	0.048	0.010	0.160	0.010
Boot. SE	0.896	7.373	1.427	0.518	0.126	0.072	0.120	0	0	0
Coverage	0.137	0.001	0.307	0	0.342	0.615	0.700	0	0	0

Tab. 4.36: Simulation results.

9th Scenario

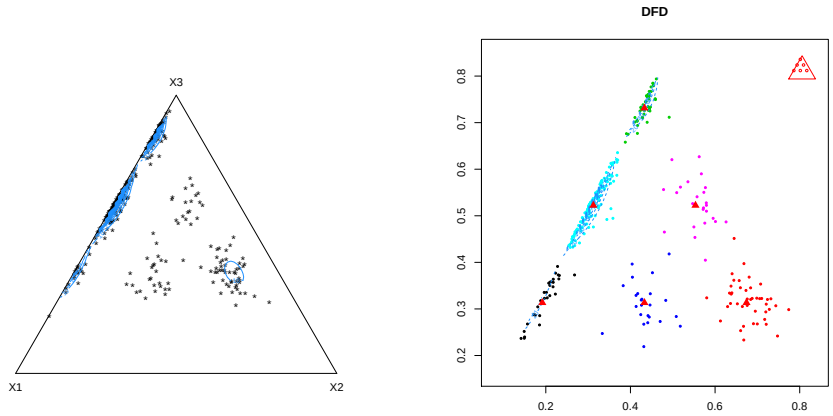

Fig. 4.14: DFD simulation - 9th configuration.

	1	2
1	$d_{SKL}(f_{1,1}, f_{1,2})$	85.2426
2	$d_{SKL}(f_{1,1}, f_{1,3})$	20.1824
3	$d_{SKL}(f_{1,1}, f_{2,2})$	119.0544
4	$d_{SKL}(f_{1,1}, f_{2,3})$	115.741
5	$d_{SKL}(f_{1,1}, f_{3,3})$	47.1167
6	$d_{SKL}(f_{1,2}, f_{1,3})$	85.7623
7	$d_{SKL}(f_{1,2}, f_{2,2})$	33.8118
8	$d_{SKL}(f_{1,2}, f_{2,3})$	30.4984
9	$d_{SKL}(f_{1,2}, f_{3,3})$	112.6965
10	$d_{SKL}(f_{1,3}, f_{2,2})$	119.574
11	$d_{SKL}(f_{1,3}, f_{2,3})$	95.5586
12	$d_{SKL}(f_{1,3}, f_{3,3})$	26.9343
13	$d_{SKL}(f_{2,2}, f_{2,3})$	24.0154
14	$d_{SKL}(f_{2,2}, f_{3,3})$	106.2136
15	$d_{SKL}(f_{2,3}, f_{3,3})$	82.1982

Tab. 4.37: SKL divergence measures.

	X_1	X_2	X_3
X_1	1.00	-0.47	-0.39
X_2	-0.47	1.00	-0.62
X_3	-0.39	-0.62	1.00

Tab. 4.38: Correlation Matrix.

	α_1	α_2	α_3	τ	π_1	π_2	π_3	π_4	π_5	π_6
True	12	0.900	30	20	0.080	0.160	0.180	0.100	0.400	0.080
MLE Mean	12.264	0.915	30.663	20.317	0.083	0.162	0.182	0.099	0.393	0.080
MLE Median	12.167	0.910	30.393	20.211	0.082	0.161	0.181	0.100	0.395	0.080
Arb	0.022	0.016	0.022	0.016	0.036	0.014	0.014	0.010	0.016	0.001
MSE	1.205	0.008	7.609	3.399	0.001	0.001	0.001	0.001	0.002	0.0004
Boot. SE	1.066	0.088	2.679	1.817	0.024	0.030	0.033	0.024	0.040	0.021
Coverage	0.941	0.955	0.940	0.945	0.955	0.950	0.941	0.946	0.944	0.958

Tab. 4.39: Simulation results.

4-part compositions scenario

In this subsection, results of a simulation study similar to the previous one are results. Data are generated from a DFD distribution with $D = 4$ and three configigurations of parameter are considered. For each scenario we generated 100 samples of size $n = 150$ and evaluated the indices previously described. Results suggest that the EM algorithm works quite well also with 4-part compositions, since the Arbs are small and the coverage levels are close to the nominal one $1 - \alpha = 0.95$.

First Scenario:

	X_1	X_2	X_3	X_4
X_1	1.00	-0.33	-0.33	-0.33
X_2	-0.33	1.00	-0.33	-0.33
X_3	-0.33	-0.33	1.00	-0.33
X_4	-0.33	-0.33	-0.33	1.00

Tab. 4.40: Correlation Matrix.

	True	MLE Mean	MLE Median	Arb	MSE	Boot. SE	Coverage
α_1	10.000	9.938	9.878	0.006	0.675	0.823	0.980
α_2	10.000	10.021	10.083	0.002	0.618	0.790	0.990
α_3	10.000	9.975	9.982	0.003	0.655	0.813	0.970
α_4	10.000	9.952	9.979	0.005	0.580	0.764	0.990
τ	15.000	14.874	14.731	0.008	1.491	1.221	0.970
π_1	0.100	0.106	0.104	0.063	0.001	0.026	0.960
π_2	0.100	0.101	0.101	0.008	0.001	0.026	0.970
π_3	0.100	0.105	0.103	0.049	0.001	0.026	0.940
π_4	0.100	0.104	0.104	0.037	0.001	0.025	0.950
π_5	0.100	0.101	0.099	0.013	0.001	0.027	0.950
π_6	0.100	0.096	0.095	0.039	0.001	0.022	0.950
π_7	0.100	0.097	0.099	0.028	0.001	0.027	0.970
π_8	0.100	0.098	0.097	0.021	0.001	0.023	0.960
π_9	0.100	0.099	0.098	0.011	0.001	0.027	0.960
π_{10}	0.100	0.093	0.094	0.070	0.001	0.027	0.940

Tab. 4.41: Simulation results.

Second Scenario:

	X_1	X_2	X_3	X_4
X_1	1.00	-0.08	-0.42	-0.33
X_2	-0.08	1.00	-0.30	-0.36
X_3	-0.42	-0.35	1.00	-0.46
X_4	-0.28	-0.36	-0.46	1.00

Tab. 4.42: Correlation Matrix.

	True	MLE Mean	MLE Median	Arb	MSE	Boot. SE	Coverage
α_1	10.000	10.019	9.872	0.002	0.507	0.715	0.960
α_2	17.000	17.095	16.991	0.006	1.570	1.256	0.940
α_3	22.000	22.132	22.016	0.006	2.817	1.682	0.960
α_4	13.000	12.949	12.877	0.004	0.907	0.956	0.970
τ	20.000	19.904	19.781	0.005	2.206	1.490	0.950
π_1	0.051	0.053	0.052	0.042	0.000	0.020	0.950
π_2	0.021	0.022	0.021	0.058	0.000	0.012	0.970
π_3	0.158	0.157	0.154	0.007	0.001	0.026	0.930
π_4	0.151	0.155	0.153	0.022	0.001	0.026	0.940
π_5	0.151	0.154	0.155	0.021	0.001	0.028	0.970
π_6	0.078	0.078	0.078	0.001	0.001	0.022	0.950
π_7	0.141	0.143	0.141	0.012	0.001	0.029	0.940
π_8	0.087	0.084	0.085	0.037	0.001	0.024	0.950
π_9	0.077	0.074	0.070	0.044	0.001	0.023	0.960
π_{10}	0.084	0.081	0.081	0.043	0.000	0.022	0.950

Tab. 4.43: Simulation results.

Third Scenario:

	X_1	X_2	X_3	X_4
X_1	1.00	-0.30	-0.17	-0.28
X_2	-0.30	1.00	-0.37	-0.40
X_3	-0.17	-0.23	1.00	-0.44
X_4	-0.45	-0.40	-0.44	1.00

Tab. 4.44: Correlation Matrix.

	True	MLE Mean	MLE Median	Arb	MSE	Boot. SE	Coverage
α_1	20.000	20.003	19.940	0.000	1.984	1.416	0.950
α_2	0.900	0.902	0.897	0.002	0.006	0.079	0.990
α_3	13.000	12.932	12.859	0.005	1.112	1.057	0.950
α_4	15.000	15.001	14.956	0.000	1.394	1.187	0.970
τ	22.000	21.919	21.756	0.004	2.704	1.651	0.970
π_1	0.019	0.019	0.020	0.043	0.000	0.012	0.960
π_2	0.143	0.145	0.140	0.013	0.001	0.027	0.970
π_3	0.110	0.111	0.113	0.012	0.001	0.023	0.940
π_4	0.172	0.172	0.173	0.001	0.001	0.033	0.960
π_5	0.042	0.040	0.034	0.038	0.000	0.017	0.980
π_6	0.138	0.139	0.133	0.003	0.001	0.030	0.930
π_7	0.158	0.160	0.159	0.007	0.001	0.030	0.940
π_8	0.024	0.024	0.022	0.019	0.000	0.013	0.950
π_9	0.069	0.069	0.067	0.001	0.000	0.020	0.980
π_{10}	0.126	0.121	0.121	0.036	0.001	0.027	0.930

Tab. 4.45: Simulation results.

5 Applications

5.1 Italian election results

Compositional data can arise in different field, a good example is in politics. The number of votes collected by each party in a particular constituency can form an interesting basis to study. On 4th March 2018, italian population voted for the two Chambers of Parliament and data regarding the results can be downloaded from the web (from now on, this dataset will be referred to as "election data"). Election data consist of 231 single-member constituencies (out of 232, because data for the Aosta constituency are not available) and of $D = 7$ components (parties): Movimento 5 Stelle (**M5S**), Partito Democratico (**PD**), Forza Italia (**FI**), Lega (**L**), Fratelli D'Italia (**FDI**), Liberi e Uguali (**LeU**) and Other parties. Figure 5.1 shows which party won in each constituency.

This dataset is of interest because votes gained by some parties seem to be positively correlated, as it can be evinced by Table 5.1.

	M5S	PD	FI	L	FDI	LeU	Other
M5S	1.000	-0.609	0.601	-0.783	-0.173	-0.185	-0.309
PD	-0.609	1.000	-0.596	0.295	0.137	0.553	-0.065
FI	0.601	-0.596	1.000	-0.505	-0.084	-0.353	-0.312
L	-0.783	0.295	-0.505	1.000	0.024	-0.255	-0.098
FDI	-0.173	0.137	-0.084	0.024	1.000	0.066	-0.065
LeU	-0.185	0.553	-0.353	-0.255	0.066	1.000	0.268
Other	-0.309	-0.065	-0.312	-0.098	-0.065	0.268	1.000

Tab. 5.1: Election data: correlation matrix.

In the following subsections, results of some interesting compositions are shown. The focus is on 3-part compositions, obtained considering two elements and amalgamating the remaining ones. Points in the ternary diagrams are colored by the geographical area of the constituency: north-west, north-east, south, center and islands.

In most cases, the preferred model is the EFD (14 cases out of $\frac{7 \cdot 6}{2} = 21$). This is due to the very general flexibility allowed by this distribution in cluster modelling, with respectively few additional parameters (with respect to the FD). In all the other

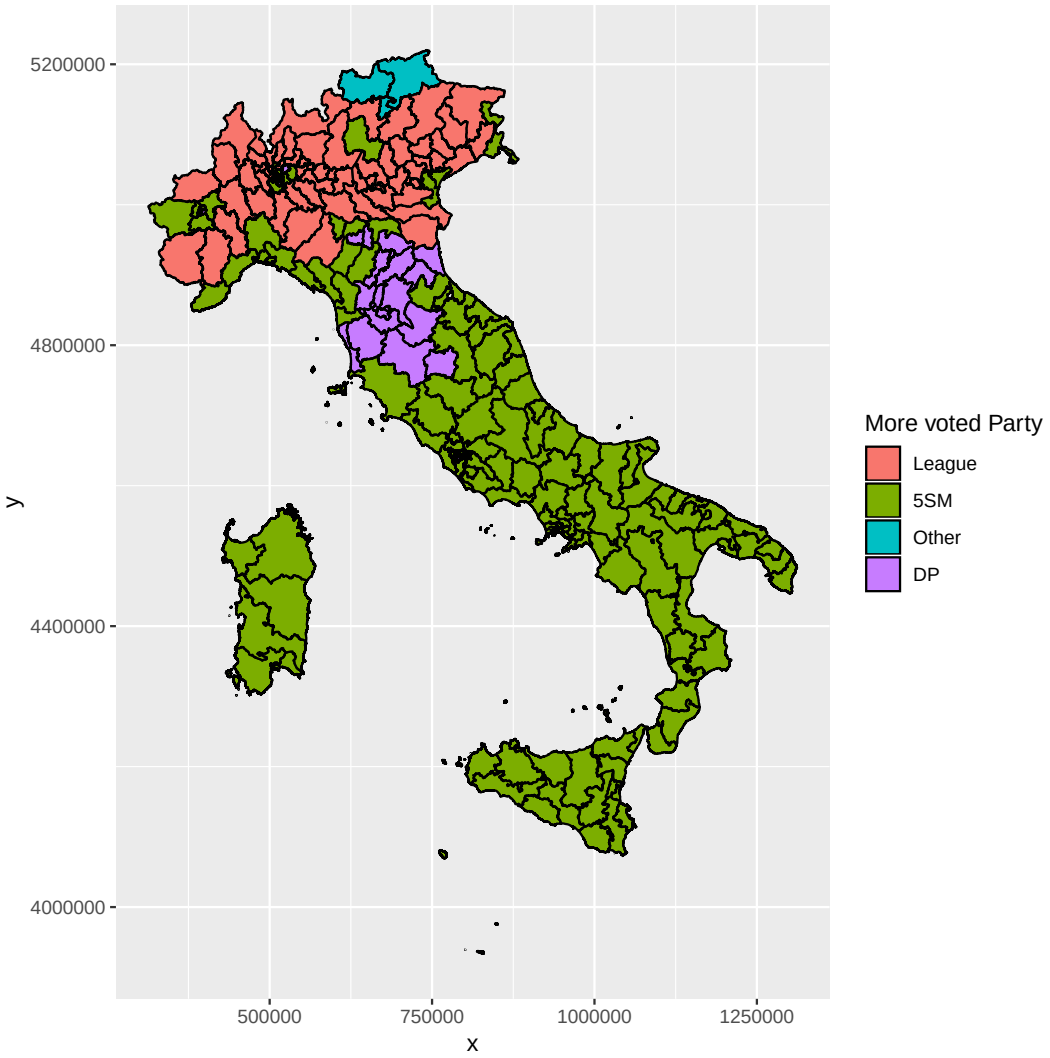

Fig. 5.1: Election results maps for the Chamber of Deputies.

seven cases, the preferred one is the ALN, because data do not show clusters and then the particular structure of the ALN is more suitable. Comparing only the Flexible models, the EFD usually has a better fit than the FD, suggesting some superiority of the Extended model. Since election data do not show a particular cluster structure, the DFD can not show its advantages over the FD. Nonetheless, in four cases the DFD outperforms the FD (look at subsections 5.1.1, 5.1.2,5.1.5 and 5.1.6) and in two more cases it has also a better fit than the Extended Flexible Dirichlet. This is due to the particular configuration of this two scenarios: inspecting plots in subsections 5.1.3 and 5.1.4 it is possible to note that three clusters are located on a straight line, that is a configuration not considered by other Flexible models.

This application confirms that the Flexible family of distribution on the simplex provides some useful distributions to describe compositional data. The best model among this family seems to be the EFD, but in some situations it fails in identifying some clusters.

5.1.1 PD Vs Lega

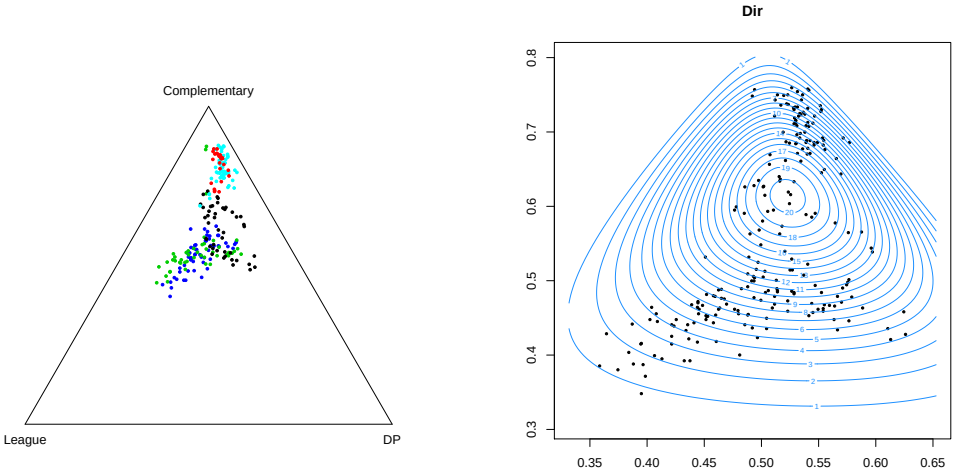

Fig. 5.2: Ternary plot and Dirichlet isodensity contour plot: PD Vs Lega. Each color refers to different geographical areas.

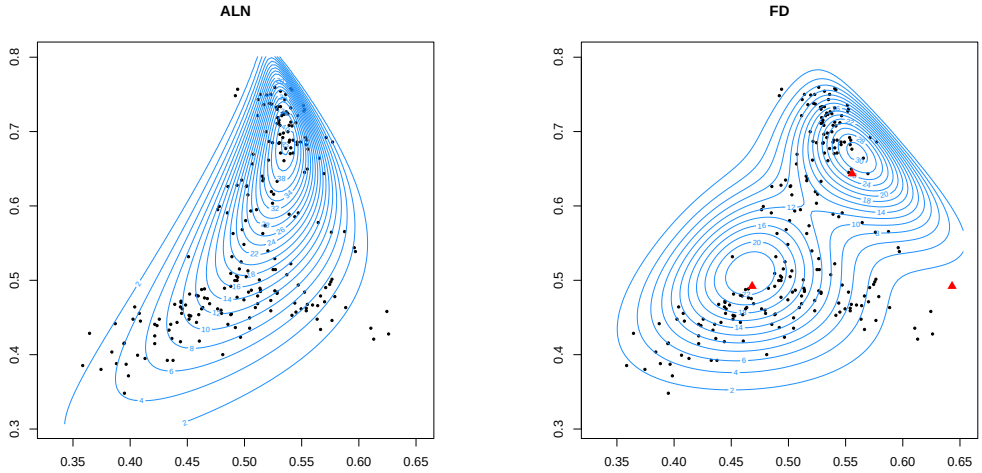

Fig. 5.3: ALN and FD isodensity contour plots: PD Vs Lega. Red triangles represent cluster means.

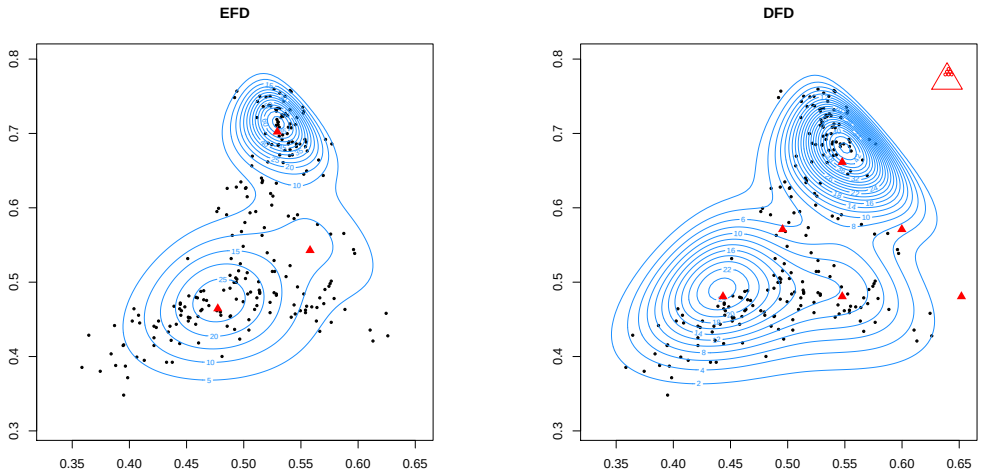

Fig. 5.4: EFD and DFD isodensity contour plots: PD Vs Lega. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-976.76	-1178.80	-1116.96	-1251.54	-1179.72
BIC	-979.46	-1161.59	-1096.31	-1224.00	-1148.74

Tab. 5.2: AIC and BIC for several models.

5.1.2 PD Vs other parties

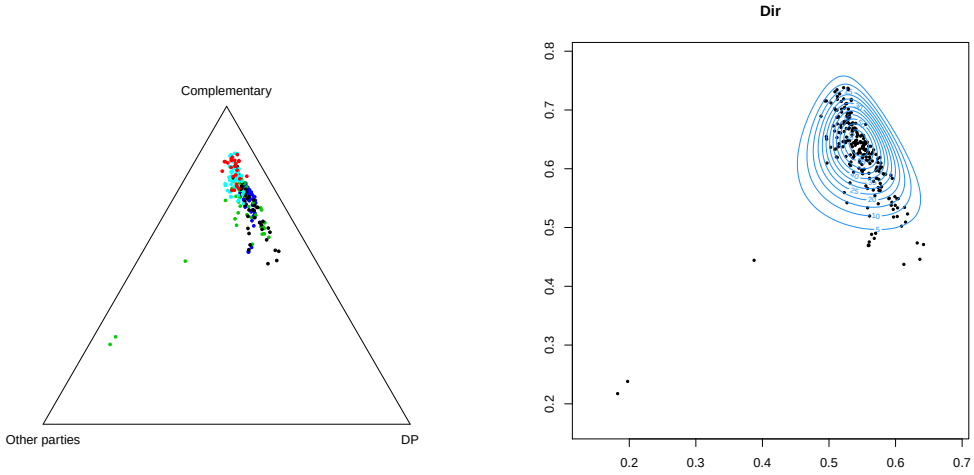

Fig. 5.5: Ternary plot and Dirichlet isodensity contour plot: PD Vs Other parties. Each color refers to different geographical areas.

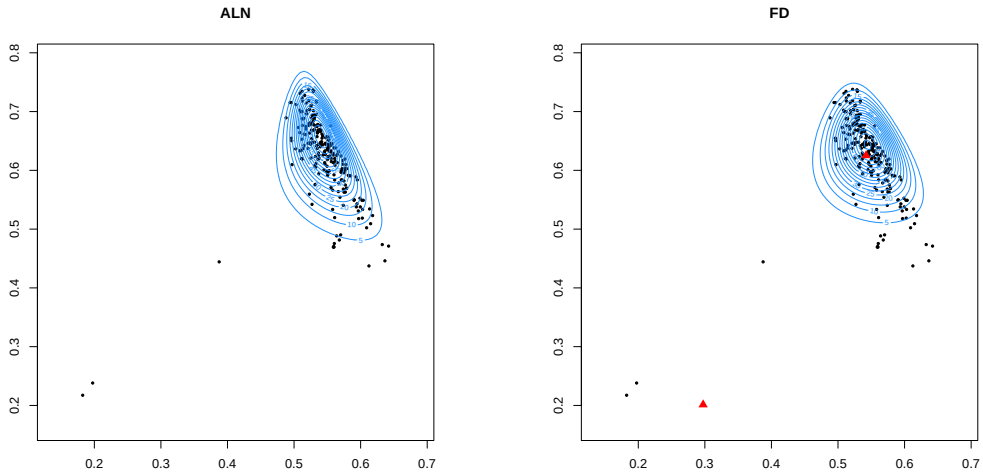

Fig. 5.6: ALN and FD isodensity contour plots: PD Vs Other parties. Red triangles represent cluster means.

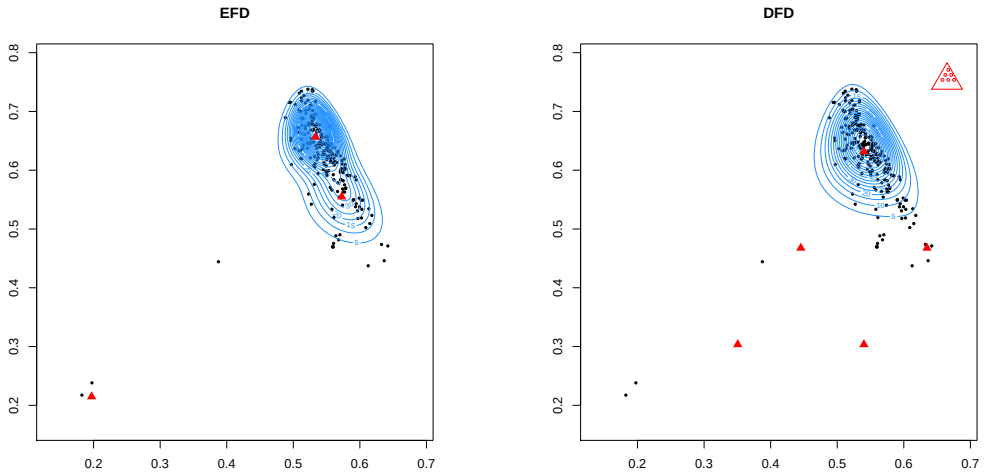

Fig. 5.7: EFD and DFD isodensity contour plots: PD Vs Other parties. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-1474.21	-1593.67	-1577.04	-1693.00	-1591.49
BIC	-1476.92	-1576.46	-1556.38	-1665.46	-1560.51

Tab. 5.3: AIC and BIC for several models.

5.1.3 Lega Vs FDI

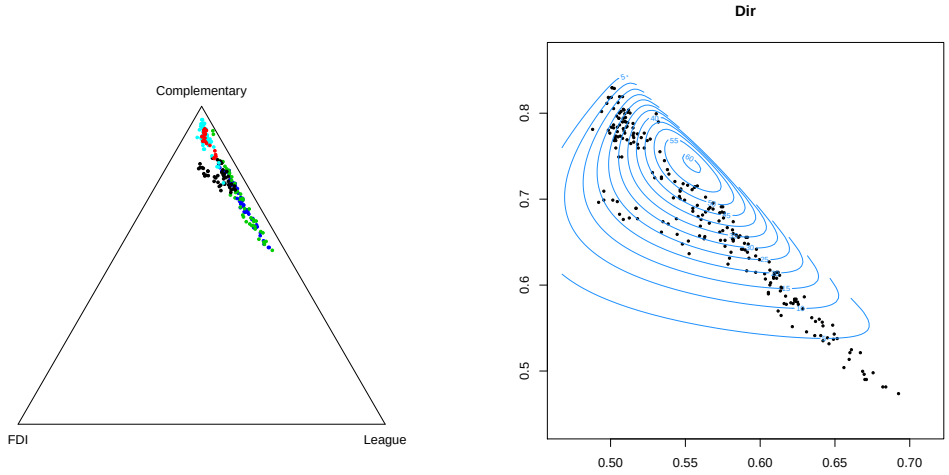

Fig. 5.8: Ternary plot and Dirichlet isodensity contour plot: Lega Vs FDI. Each color refers to different geographical areas.

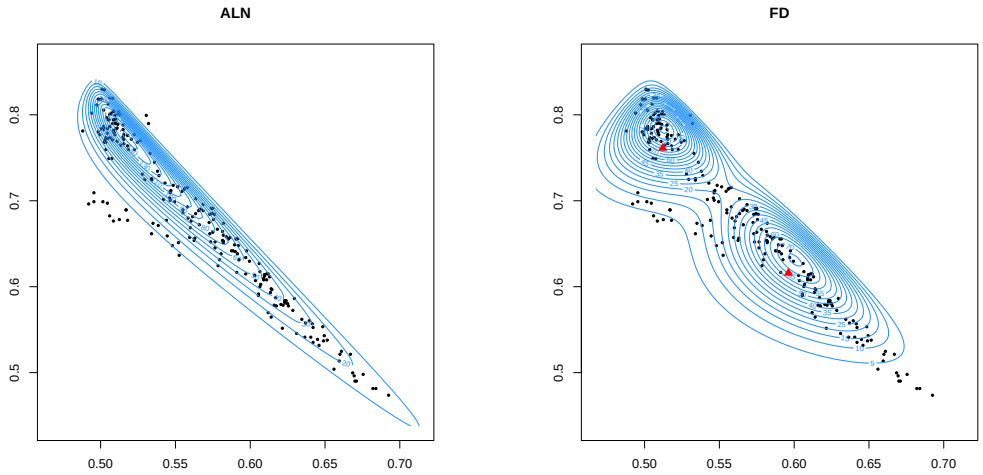

Fig. 5.9: ALN and FD isodensity contour plots: Lega Vs FDI. Red triangles represent cluster means.

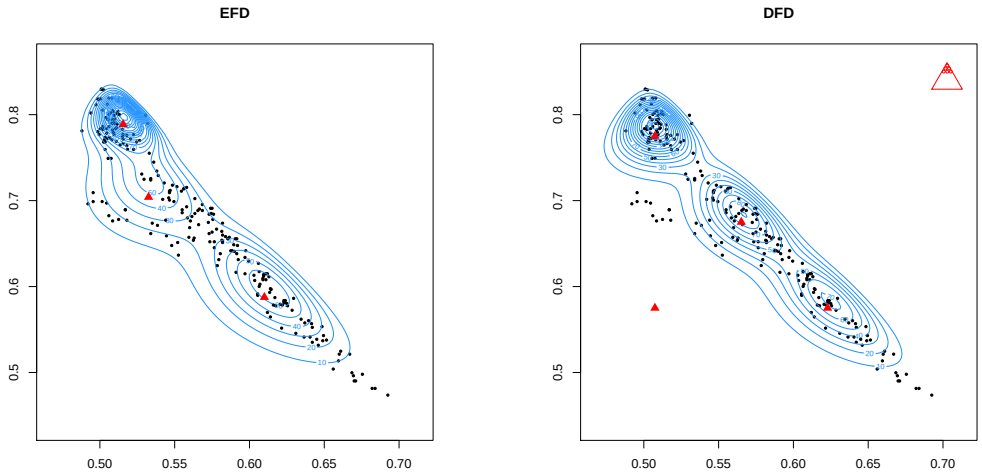

Fig. 5.10: EFD and DFD isodensity contour plots: Lega Vs FDI. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-1397.38	-1758.81	-1630.63	-1675.22	-1739.61
BIC	-1400.09	-1741.59	-1609.97	-1647.68	-1708.63

Tab. 5.4: AIC and BIC for several models.

5.1.4 Lega Vs LEU

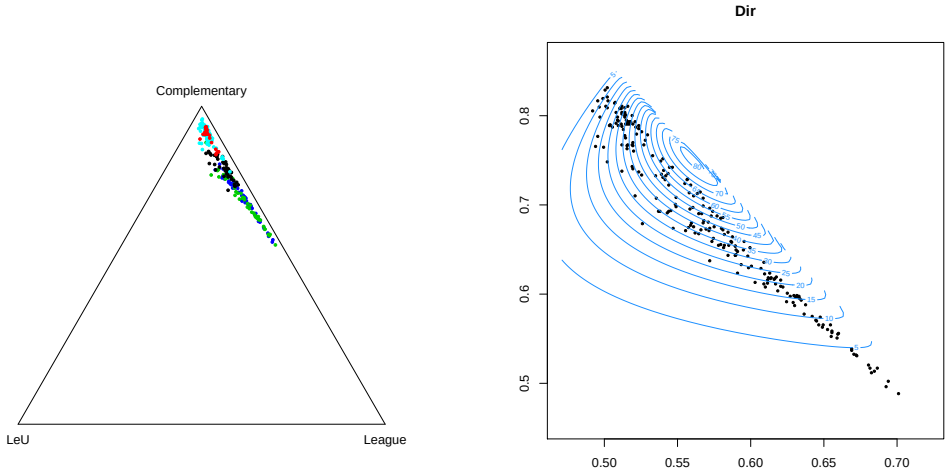

Fig. 5.11: Ternary plot and Dirichlet isodensity contour plot: Lega Vs LeU. Each color refers to different geographical areas.

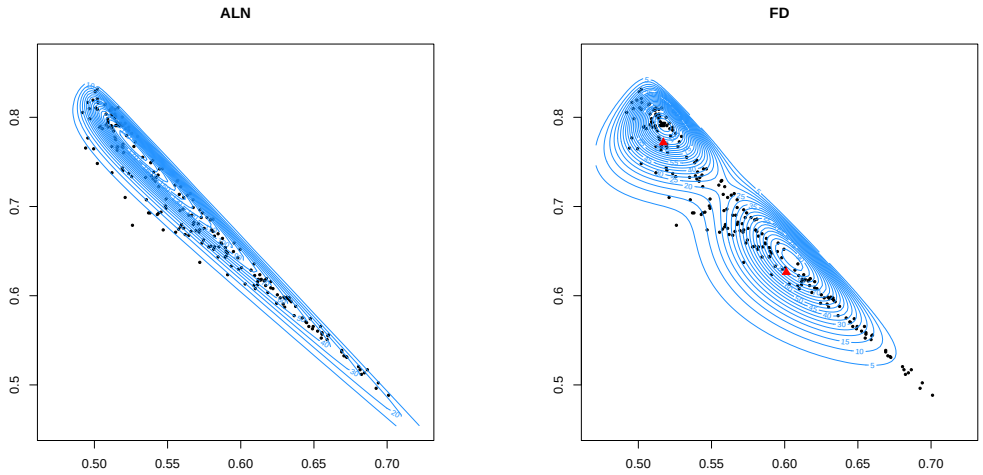

Fig. 5.12: ALN and FD isodensity contour plots: Lega Vs LeU. Red triangles represent cluster means.

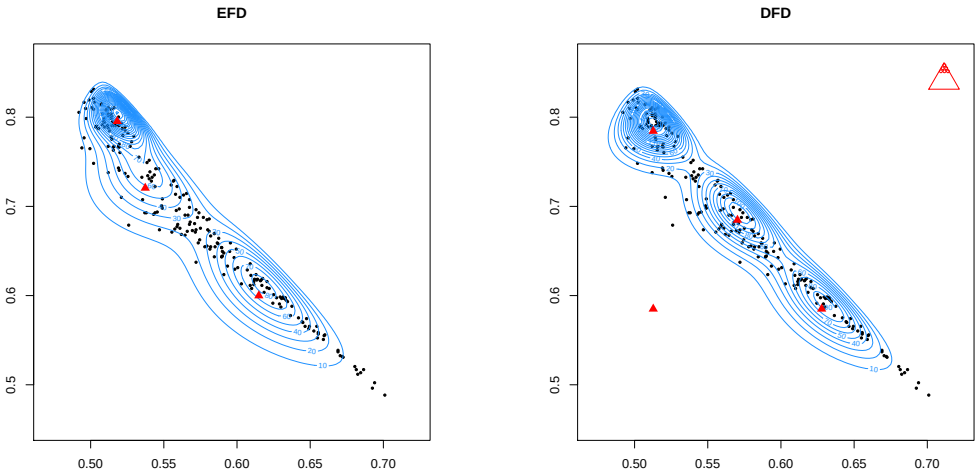

Fig. 5.13: EFD and DFD isodensity contour plots: Lega Vs LeU. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-1479.38	-1919.20	-1734.91	-1762.14	-1860.91
BIC	-1482.09	-1901.99	-1714.26	-1734.60	-1829.92

Tab. 5.5: AIC and BIC for several models.

5.1.5 Lega Vs other parties

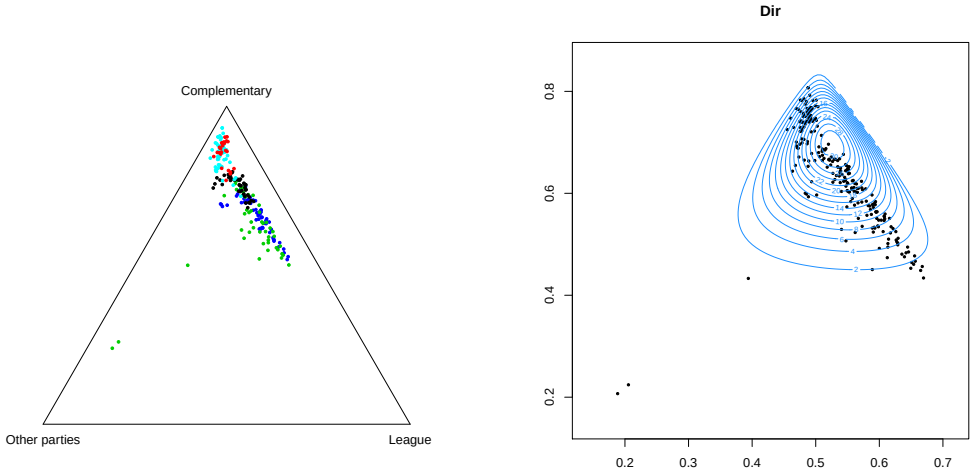

Fig. 5.14: Ternary plot and Dirichlet isodensity contour plot: Lega Vs Other parties. Each color refers to different geographical areas.

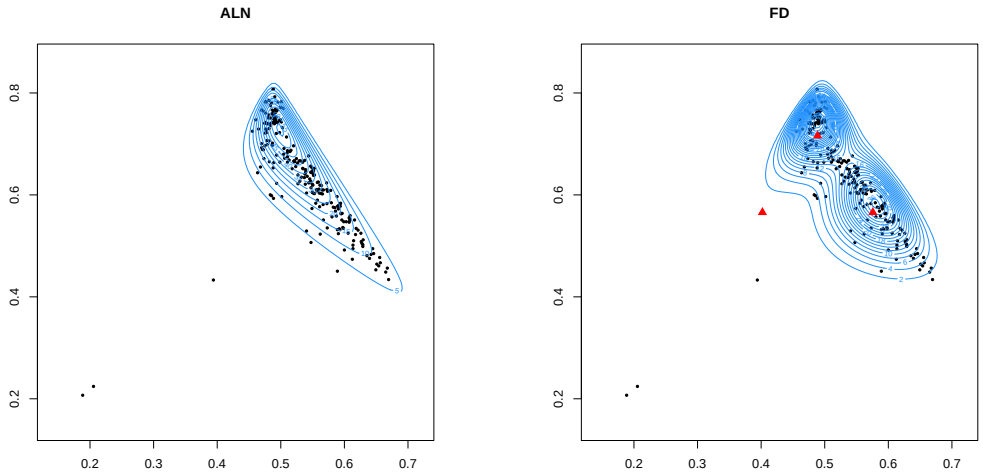

Fig. 5.15: ALN and FD isodensity contour plots: Lega Vs Other parties. Red triangles represent cluster means.

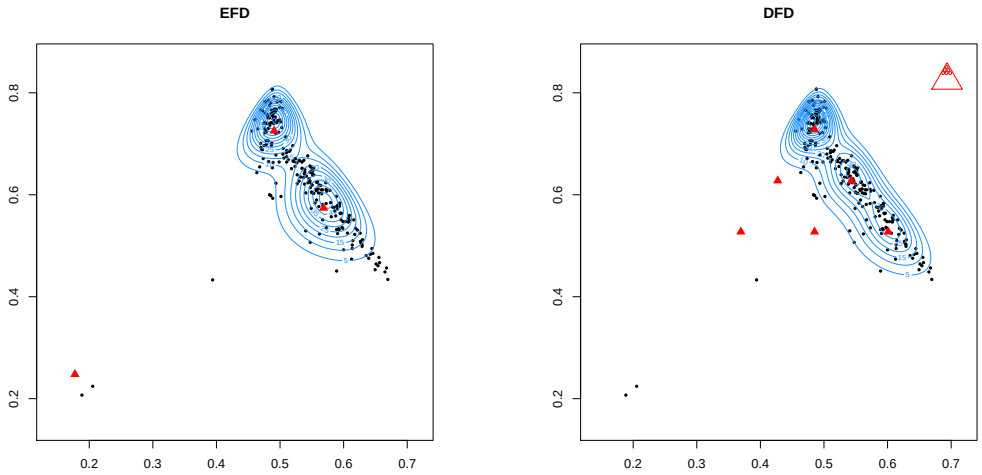

Fig. 5.16: EFD and DFD isodensity contour plots: Lega Vs Other parties. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-1156.02	-1364.50	-1329.95	-1405.60	-1399.53
BIC	-1158.72	-1347.28	-1309.30	-1378.06	-1368.55

Tab. 5.6: AIC and BIC for several models.

5.1.6 FDI Vs other parties

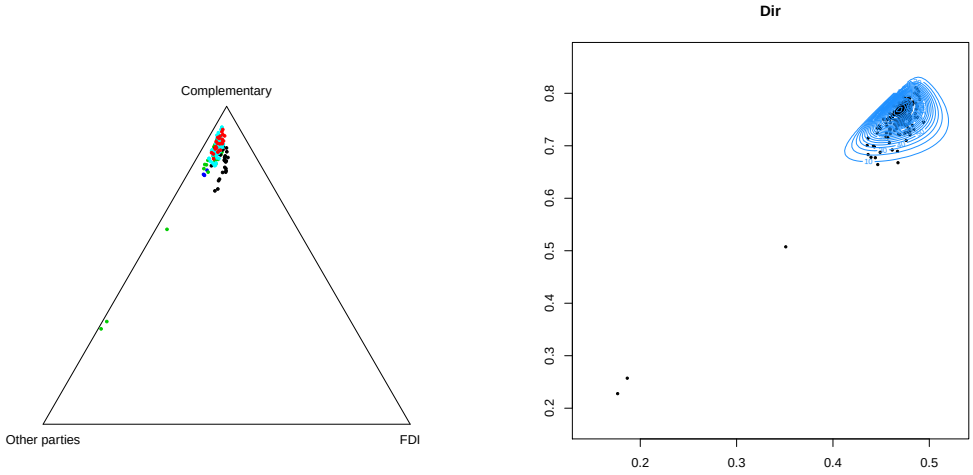

Fig. 5.17: Ternary plot and Dirichlet isodensity contour plot: FDI Vs Other parties. Each color refers to different geographical areas.

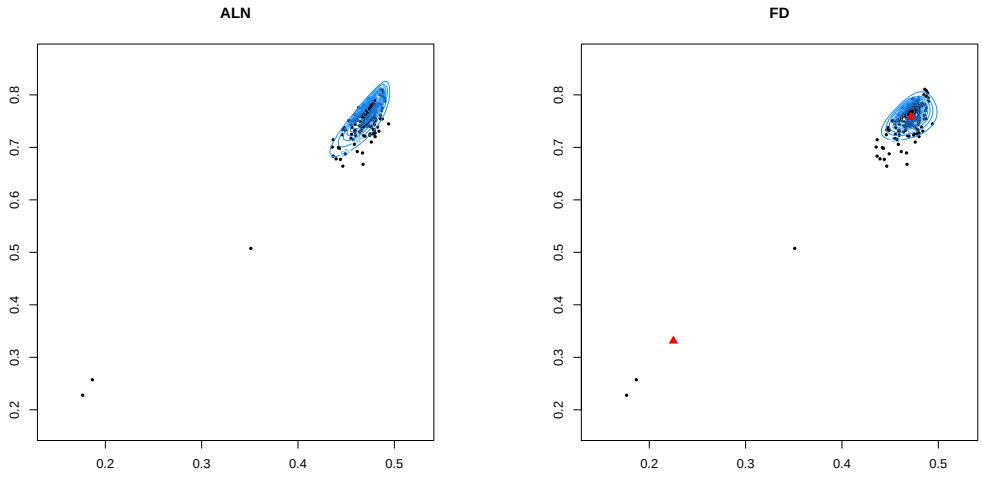

Fig. 5.18: ALN and FD isodensity contour plots: FDI Vs Other parties. Red triangles represent cluster means.

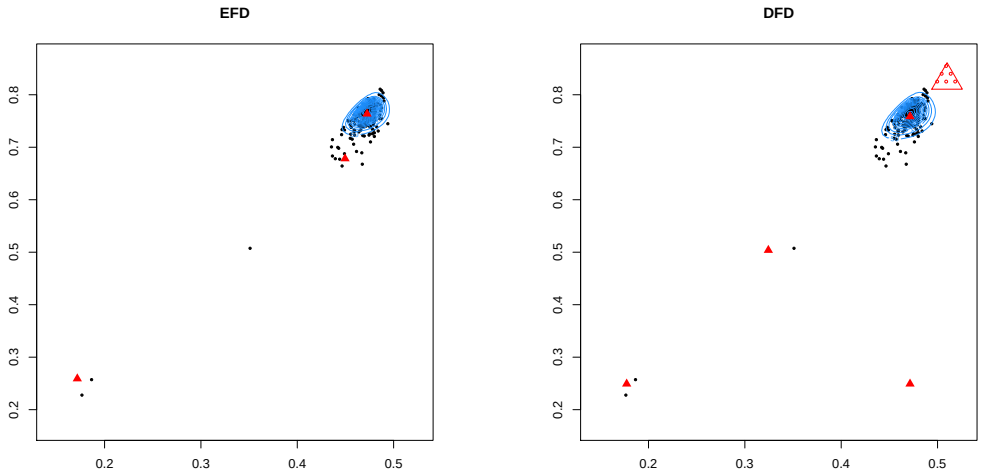

Fig. 5.19: EFD and DFD isodensity contour plots: FDI Vs Other parties. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-2101.86	-2427.62	-2447.84	-2529.58	-2495.19
BIC	-2104.57	-2410.41	-2427.19	-2502.04	-2464.20

Tab. 5.7: AIC and BIC for several models.

5.2 Olive oils

The Olive oil data have been discussed, for the first time, in a work by Forina et al. [36] and it has been made available in the R package **pdfCluster** [11]. These data have been largely analyzed in the literature with clustering aims [10, 87] as well to compare models [63]. This dataset is composed by 572 rows representing a different specimen of olive oil produced in Italy, whereas the 10 columns represent some characteristics of a specific oil. In particular, the first two variables correspond to the macro-area and the region of origin of each oil; the other 8 columns represent measurements regarding the fatty acid composition: Palmitic, Palmitoleic, Stearic, Oleic, Linoleic, Linolenic, Arachidic and Eicosenoic.

Linolenic and Arachidic variables presented some zero values: since the proposed distributions are not defined on the border of the simplex, in order to keep all the elements of the basis, the oils connected to the zero values have been removed by the sample. The final sample is then composed by 535 oils, and the relative composition has been obtained, closing each row of the dataset.

Acid	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Linolenic	Arachidic	Eicosenoic
Mean	12.467	1.281	2.273	72.973	9.873	0.340	0.620	0.173

Tab. 5.8: Mean of the components (in percentages).

These data show some positive (and high) correlation. For example, the correlation among Palmitic and Palmitoleic is 0.85, as it can be seen in Table 5.9.

	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Linolenic	Arachidic	Eicosenoic
Palmitic	1.00	0.85	-0.12	-0.82	0.47	0.20	0.05	0.47
Palmitoleic	0.85	1.00	-0.21	-0.86	0.62	0.02	0.00	0.40
Stearic	-0.12	-0.21	1.00	0.09	-0.20	0.19	0.14	0.22
Oleic	-0.82	-0.86	0.09	1.00	-0.88	-0.08	-0.22	-0.37
Linoleic	0.47	0.62	-0.20	-0.88	1.00	-0.14	0.21	0.07
Linolenic	0.20	0.02	0.19	-0.08	-0.14	1.00	0.38	0.55
Arachidic	0.05	0.00	0.14	-0.22	0.21	0.38	1.00	0.22
Eicosenoic	0.47	0.40	0.22	-0.37	0.07	0.55	0.22	1.00

Tab. 5.9: Olive Oil data: correlation coefficients.

5.2.1 2-part compositions

The first part of this application regards the eight 2-part compositions (i.e. the one-dimensional marginals). These marginals allow us to compare distributions in a variety of scenarios regarding symmetry (asymmetric and symmetric cases) and cluster structure (perfect unimodality and bimodality). Table 5.10 shows AIC and

BIC values for the each marginal and eache estimated model. The FD and the EFD distributions are the one with the best fit; this means that univariate marginals of the Olive data exhibit a cluster and/or a covariance structure that neither the Dirichlet nor the ALN can cover. Moreover, these structures do not require a more complex model (the DFD) than the FD and the EFD that, with their flexibility can fit data in a good way (look at Table 5.10 and Figures 5.20 and 5.21). Please not that the DFD generally performs better than the Dirichlet and the ALN.

Component	Crit.	Dir	FD	ALN	EFD	DFD
X_1 = Palmitic	AIC	-2884.73	-2937.41	-2888.62	-2942.85	-2927.33
	BIC	-2876.17	-2920.28	-2880.05	-2921.44	-2888.79
X_2 = Palmitoleic	AIC	-4101.42	-4147.60	-4072.50	-4145.86	-4135.82
	BIC	-4092.86	-4130.47	-4063.94	-4124.45	-4097.28
X_3 = Stearic	AIC	-4555.94	-4581.79	-4568.91	-4579.67	-4570.91
	BIC	-4547.38	-4564.66	-4560.34	-4558.26	-4532.37
X_4 = Oleic	AIC	-1904.75	-1970.07	-1900.14	-1976.37	-1960.06
	BIC	-1896.19	-1952.94	-1891.58	-1954.95	-1921.52
X_5 = Linoleic	AIC	-2415.18	-2607.50	-2396.43	-2607.56	-2597.46
	BIC	-2406.62	-2590.37	-2387.87	-2586.15	-2558.92
X_6 = Linolenic	AIC	-5746.04	-5828.12	-5676.07	-5848.01	-5766.32
	BIC	-5737.48	-5810.99	-5667.51	-5826.60	-5727.78
X_7 = Arachidic	AIC	-5127.11	-5355.77	-4989.70	-5353.76	-5328.67
	BIC	-5118.55	-5338.64	-4981.14	-5332.35	-5290.13
X_8 = Eicosenoic	AIC	-5735.89	-6182.78	-5652.57	-6384.73	-6212.04
	BIC	-5727.32	-6165.65	-5644.00	-6363.32	-6173.50

Tab. 5.10: AIC and BIC for the 2-part compositions - Olive data. Values in red are the maxima of each row.

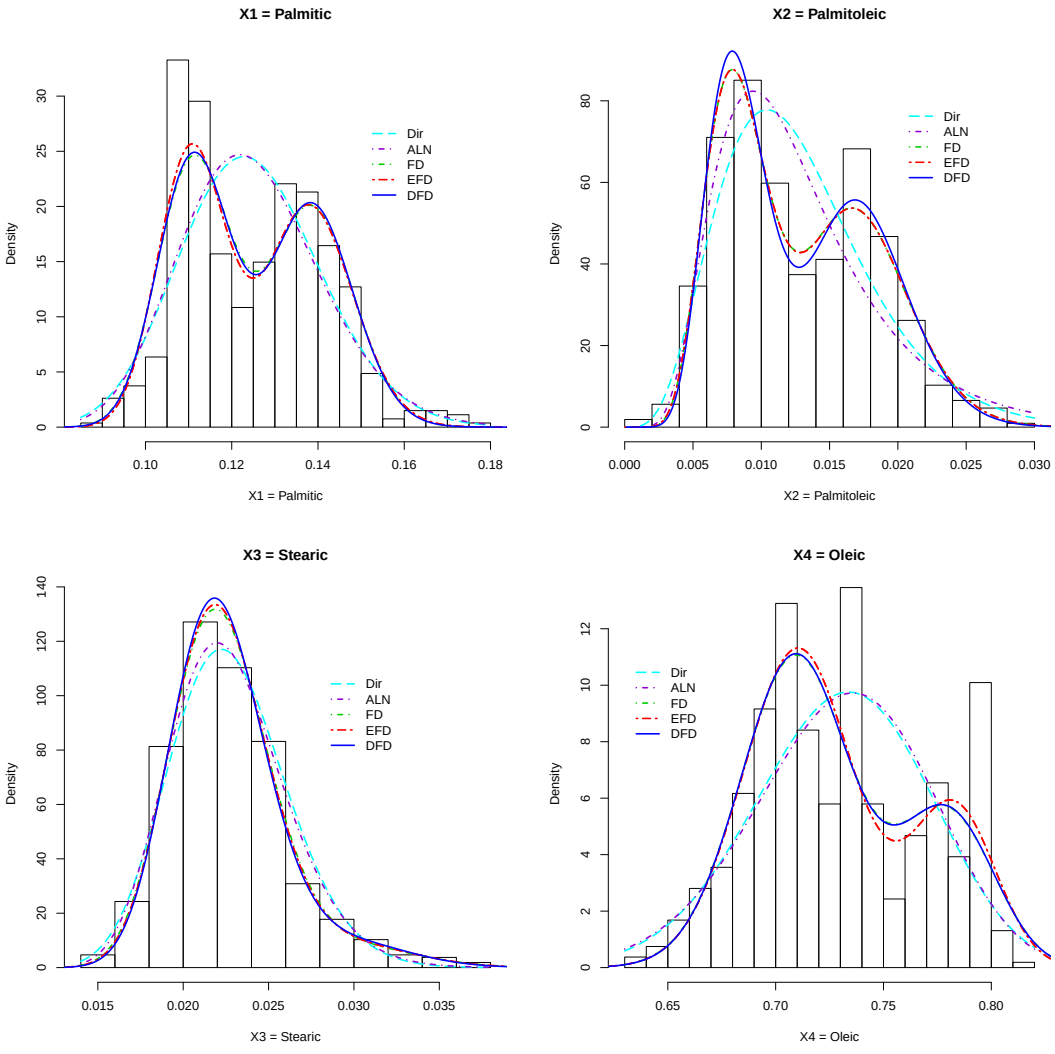

Fig. 5.20: Histograms and estimated densities of 2-part compositions.

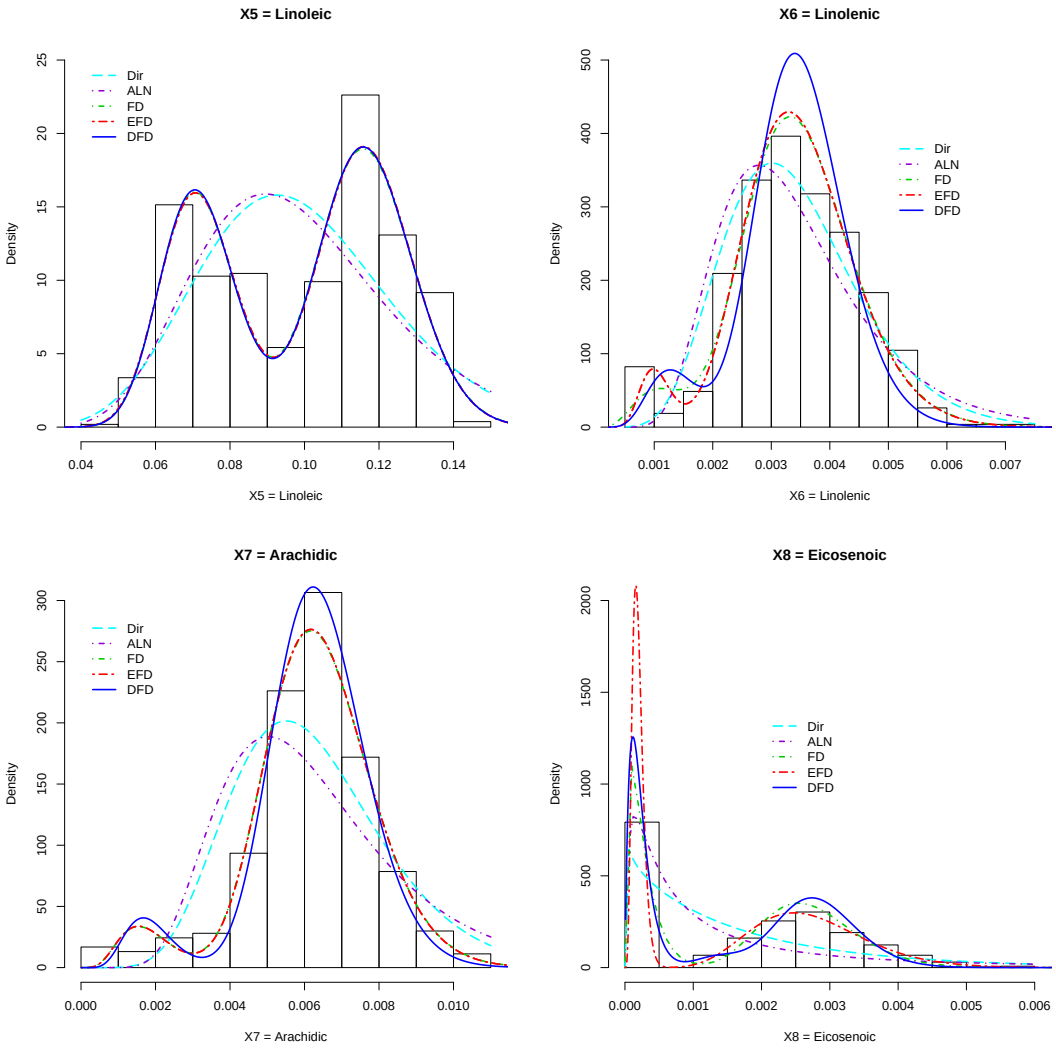

Fig. 5.21: Histograms and estimated densities of 2-part compositions.

5.2.2 3-part compositions

Due to the high number of 3-components compositions available (two components versus the amalgamation of the not considered components), only the results of some interesting and representative scenarios are reported.
The results of this application are not so different by the ones of the Election Data:

- **Palmitic Vs Palmitoleic** (Section 5.2.2): despite the correlation coefficient among two components is very elevated (0.85), data do not show a particular cluster structure. Then, the cluster structure assumed by the DFD is not supported by data.

- **Stearic Vs Oleic** (Section 5.2.2): three clusters are located on a straight line. The Flexible model that better recognize this configuration is the DFD, indicating that allowing for a more complex model (with respect to the EFD) helps in fitting data. Despite this advantage, the best model is the ALN: this means that the cluster structure can be ignored.

- **Stearic Vs Linoleic** (Section 5.2.2): both the EFD and the DFD recognize three cluster along the boundary of the simplex. These three clusters are not perfectly located on a straight line and for this reason the EFD has a better fit than the DFD's one. Nonetheless, the FD does not provide good AIC and BIC as the DFD's ones.

- **Oleic vs Linoleic** (Section 5.2.2): despite inspecting the contour plots, the DFD seems to be the best model (i.e. it catches a third clusters in the bottom area), the criterions point at the EFD as the best model. This is a very clear way to see that the great number of parameters considered heavily penalize the DFD model.

- **Linolenic Vs Eicosenoic** (Section 5.2.2): in this scenario, looking at contour plots, the DFD seems to be the best model, since it dedicates a cluster to some isolated observations. Nonetheless, both the AIC and the BIC point at the EFD as the best model. Comparing the criterions of the DFD and of the FD models and looking at the correspondent contour plots, one can conclude that the DFD is preferred: the criterions are very similar whereas the cluster structure is better recognized by the DFD model.

Palmitic Vs Palmitoleic

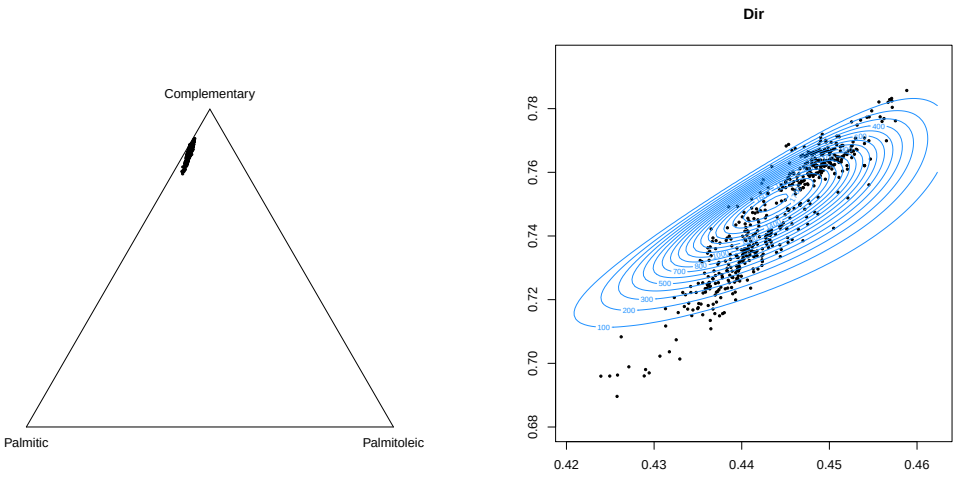

Fig. 5.22: Ternary plot and Dirichlet isodensity contour plot: Palmitic Vs Palmitoleic.

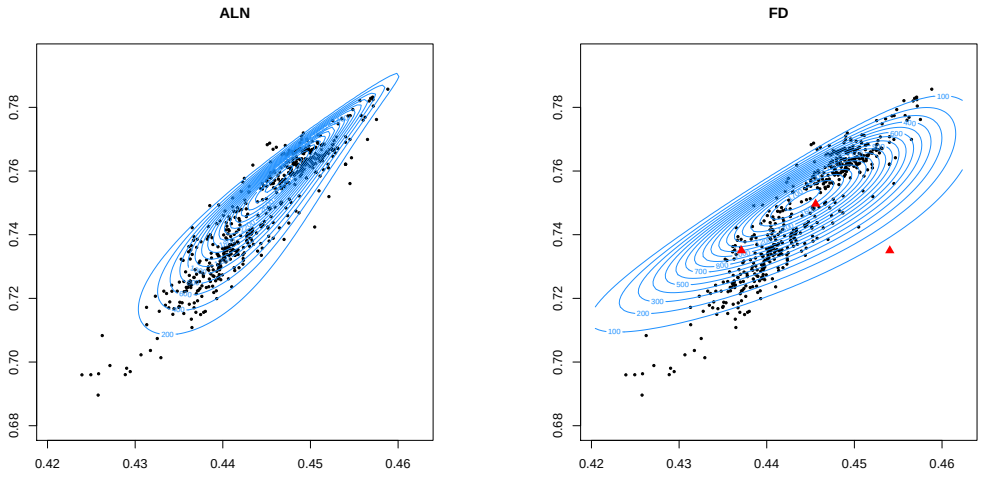

Fig. 5.23: ALN and FD isodensity contour plots: Palmitic Vs Palmitoleic. Red triangles represent cluster means.

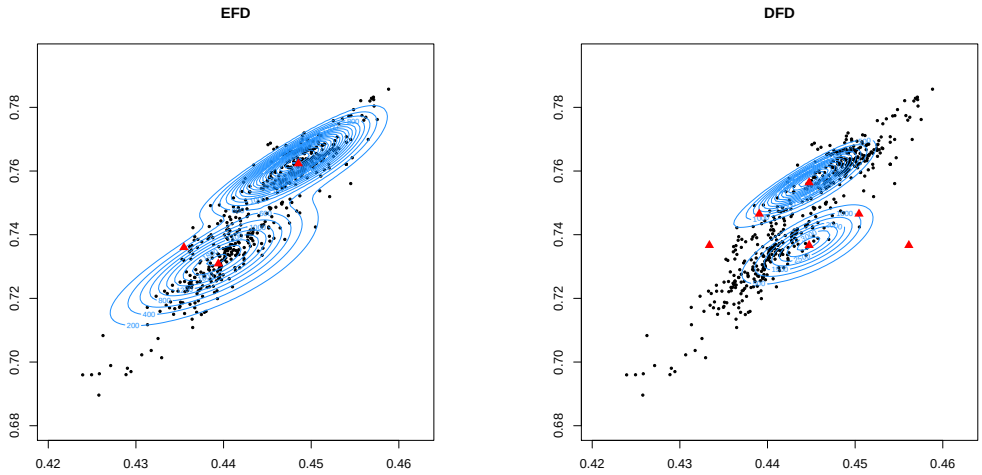

Fig. 5.24: EFD and DFD isodensity contour plots: Palmitic Vs Palmitoleic. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-6949.5	-7505.3	-6946.1	-7534.6	-6450.01
BIC	-6936.7	-7483.69	-6920.4	-7500.3	-6411.47

Tab. 5.11: AIC and BIC for several models.

Stearic Vs Oleic

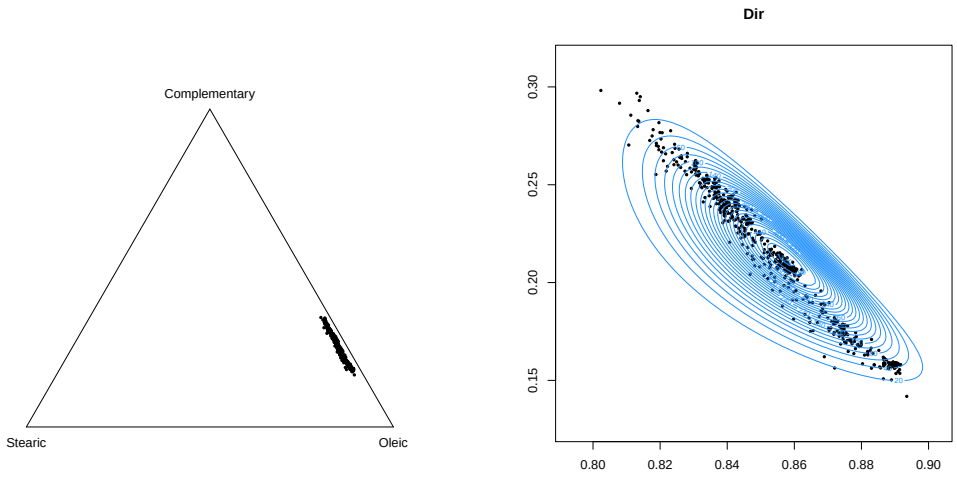

Fig. 5.25: Ternary plot and Dirichlet isodensity contour plot: Stearic Vs Oleic.

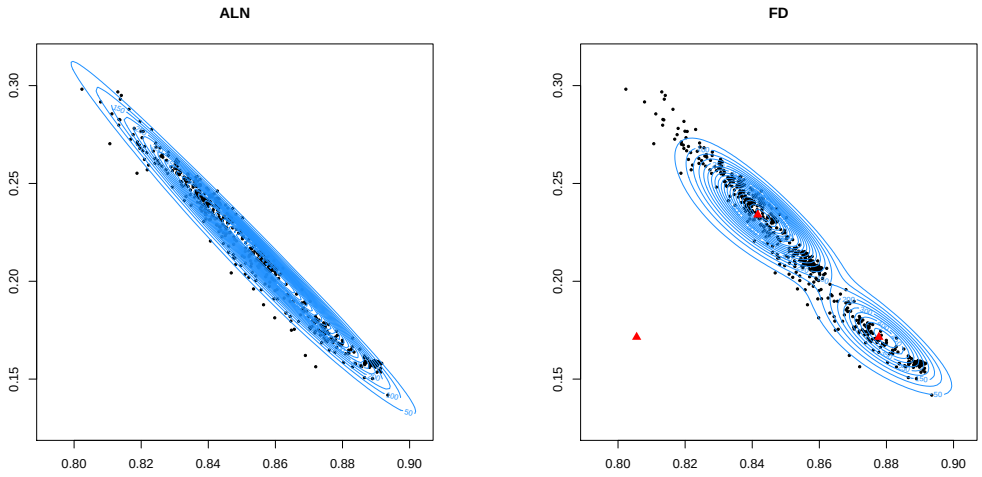

Fig. 5.26: ALN and FD isodensity contour plots: Stearic Vs Oleic. Red triangles represent cluster means.

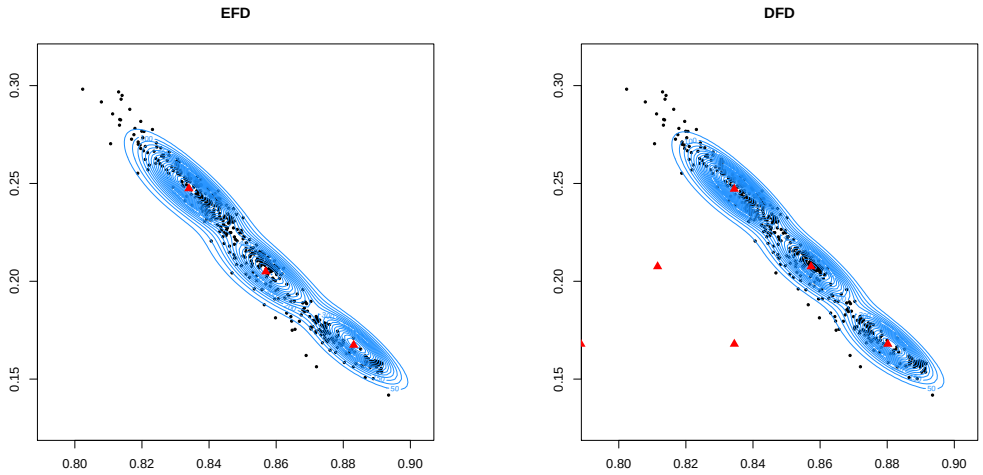

Fig. 5.27: EFD and DFD isodensity contour plots: Stearic Vs Oleic. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-5618	-6478.3	-6190.4	-6388.8	-6458.2
BIC	-5605.4	-6456.8	-6164.7	-6354.5	-6419.7

Tab. 5.12: AIC and BIC for several models.

Stearic Vs Linoleic

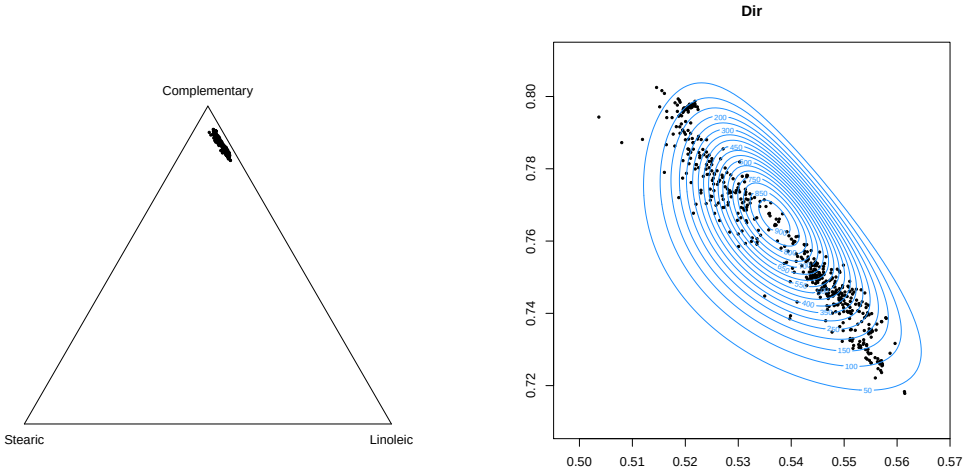

Fig. 5.28: Ternary plot and Dirichlet isodensity contour plot: Stearic Vs Linoleic.

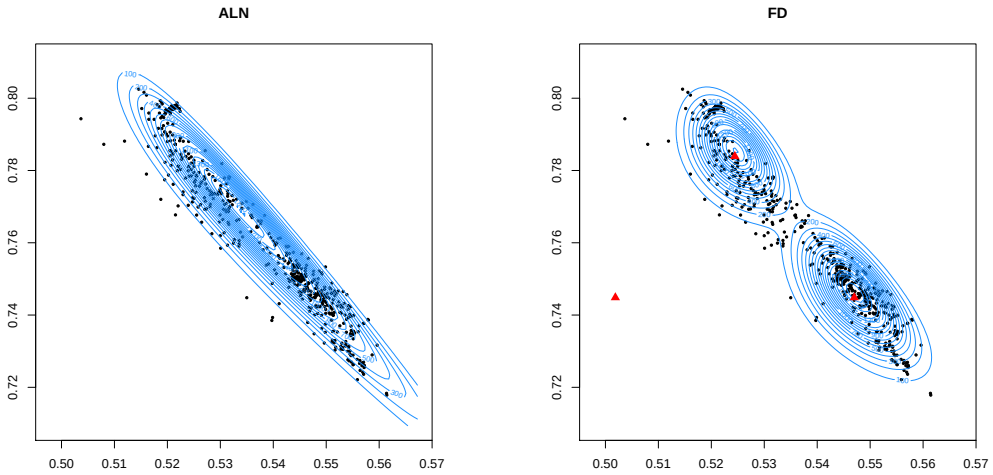

Fig. 5.29: ALN and FD isodensity contour plots: Stearic Vs Linoleic. Red triangles represent cluster means.

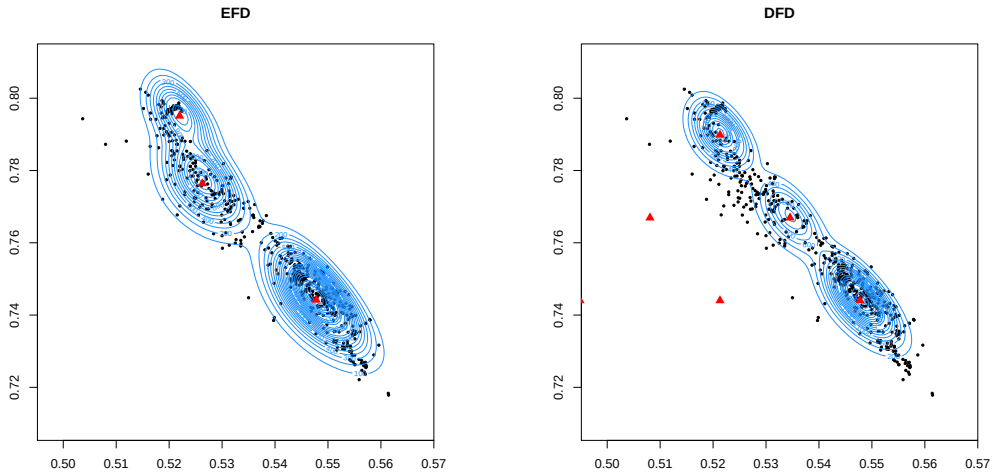

Fig. 5.30: EFD and DFD isodensity contour plots: Stearic Vs Linoleic. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-6237.0	-6982.7	-7022.0	-7153.4	-7022.7
BIC	-6224.2	-6961.3	-6996.3	-7119.1	-6984.2

Tab. 5.13: AIC and BIC for several models.

Oleic Vs Linoleic

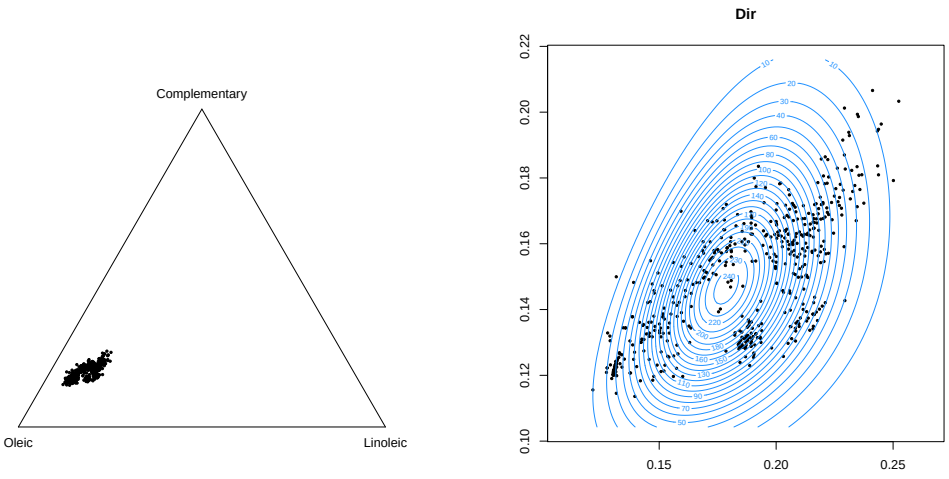

Fig. 5.31: Ternary plot and Dirichlet isodensity contour plot: Oleic Vs Linoleic.

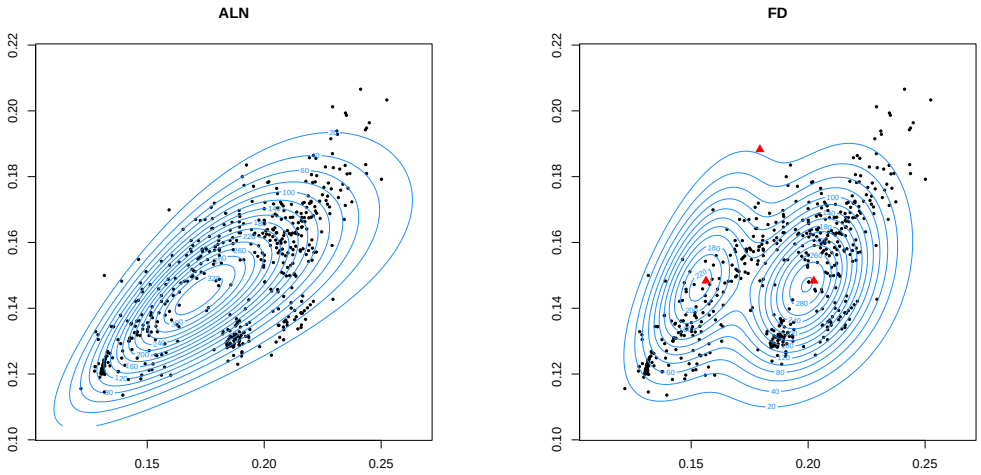

Fig. 5.32: ALN and FD isodensity contour plots: Oleic Vs Linoleic. Red triangles represent cluster means.

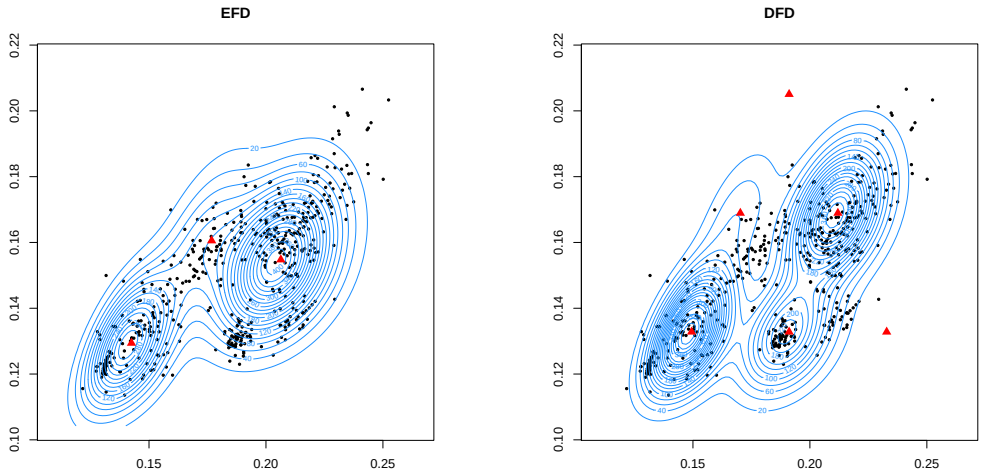

Fig. 5.33: EFD and DFD isodensity contour plots: Oleic Vs Linoleic. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-4822.5	-5120.0	-5060.1	-5280.6	-5224.1
BIC	-4809.6	-5098.6	-5034.4	-5246.3	-5185.5

Tab. 5.14: AIC and BIC for several models.

Linolenic Vs Eicosenoic

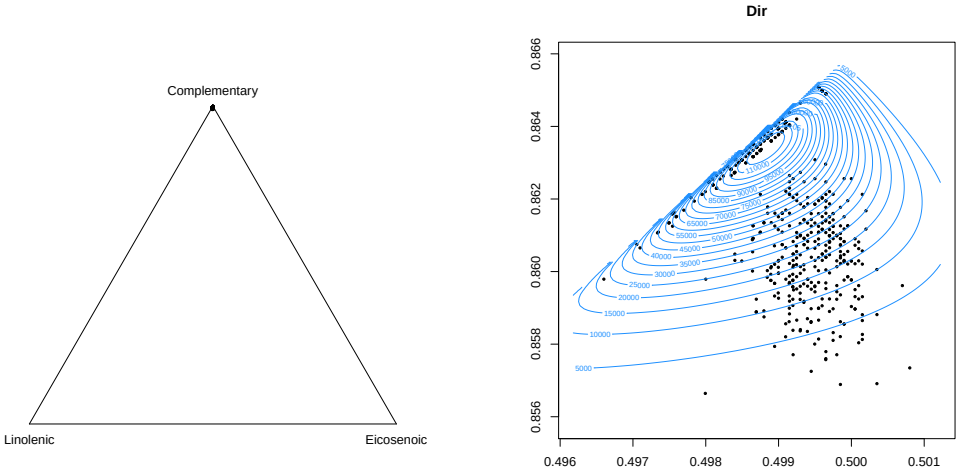

Fig. 5.34: Ternary plot and Dirichlet isodensity contour plot: Linolenic Vs Eicosenoic.

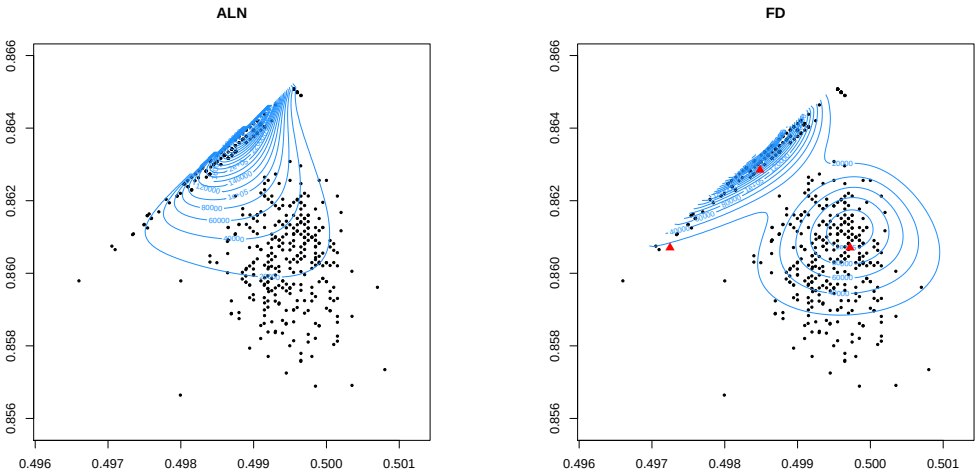

Fig. 5.35: ALN and FD isodensity contour plots: Linolenic Vs Eicosenoic. Red triangles represent cluster means.

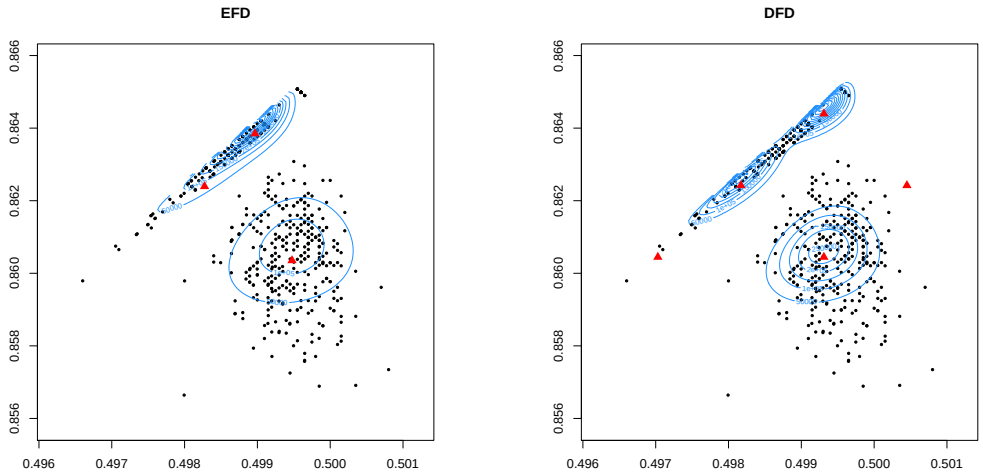

Fig. 5.36: EFD and DFD isodensity contour plots: Linolenic Vs Eicosenoic. Red triangles represent cluster means.

	Dir	ALN	FD	EFD	DFD
AIC	-11208.2	-11499.2	-11847.5	-12091.7	- 11835.4
BIC	-11195.4	-11477.8	-11821.8	-12057.4	-11797.4

Tab. 5.15: AIC and BIC for several models.

6

A Flexible distribution for count data

An important kind of data is represented by counts (i.e. non-negative integers). According to their support and to the nature of the phenomenon, univariate count data are usually treated as Poisson, Binomial or Negative Binomial random variables. The Binomial is the distribution of the number of successes in n independent experiments. These experiments can lead to a dichotomous outcome (success or failure) and the parameter π represents the probability of a single success. If the outcome of the experiment can assume more than two values, then the Binomial distribution is generalized by the Multinomial distribution:

$$\begin{cases} \mathbf{X} \sim \text{Multinomial}(n, \boldsymbol{\pi}), \quad n \in \mathbb{N},\ \boldsymbol{\pi} \in \mathcal{S}^D, \\ \mathbb{S}_{\mathbf{X}} = \mathcal{S}_n^D, \\ \mathbb{E}[X_r] = n\pi_r, \qquad r = 1, \dots, D, \\ \text{Var}(X_r) = n\pi_r(1-\pi_r), \qquad r = 1, \dots, D, \\ \text{Cov}(X_r, X_h) = -n\pi_r\pi_h, \quad r, h = 1, \dots, D, r \neq h. \end{cases} \tag{6.1}$$

where $\mathbb{S}_{\mathbf{X}}$ denotes the support of the random vector X. The support for the Multinomial distribution is the $\{n, D\}$-simplex [81]:

$$\mathcal{S}_n^D = \left\{ \mathbf{x} = (x_1, \dots, x_D)^\intercal : x_k \in \mathbb{N} \cup \{0\},\ k = 1, \dots, D,\ \sum_{k=1}^{D} x_k = n \right\}. \tag{6.2}$$

This means that the support $\mathcal{S}_n^D$ shares the same problems as the unitary simplex in Definition 1, since each element in $\mathcal{S}_n^D$ must have components summing to n. From (6.1) it is possible to see that also the covariance matrix of the Multinomial distribution suffers of the same problems of the Dirichlet's covariance matrix listed in section 3.1. In order to obtain a more flexible distribution for multivariate count data with the same structure of Multinomial data (i.e. subject to a sum constraint), compound distributions are often used [9, 17, 19, 49, 55, 64, 84, 88, 89].

6.1 Compound distributions for count data

Compound distributions are probability distributions obtained by a two-step approach. The first step consists in assuming that the parameter $\boldsymbol{\theta}$ of the distribution of a random vector $\mathbf{X}$ is not constant but follow a specific distribution. Then, the resulting joint distribution is marginalized, integrating over $\boldsymbol{\theta}$. More formally, let $\mathbf{X}|\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ be two random vectors with probability density function $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$, respectively. Then, the marginal distribution of $\mathbf{X}$ is identified by the probability density function $f(\mathbf{x}) = \int_{\Theta} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$, where Θ is the support of $\boldsymbol{\theta}$.

This approach leads to more flexible distributions and has been used in several fields, such as epidemiology [51], text analysis [90], marketing [80], biometrics [49, 88], information retrieval [89], psychology [9] and species composition [17]. Particular interest has been regarded in compound distributions for modelling multivariate count data; In this area, a popular choice is to assume that $\mathbf{X}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ and then impose a distribution $\mathcal{F}$ on $\boldsymbol{\Pi}$ ($\boldsymbol{\Pi}$ is the random vector that assumes value $\boldsymbol{\pi}$). The support of $\boldsymbol{\Pi}$ is the D-part simplex $\mathcal{S}^D$. Recalling now that, if $\mathbf{X}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$, then:

- $\mathbb{E}[X_r|\boldsymbol{\Pi} = \boldsymbol{\pi}] = n\pi_r$
- $\text{Var}(X_r|\boldsymbol{\Pi} = \boldsymbol{\pi}) = n\pi_r(1 - \pi_r)$
- $\text{Cov}(X_r, X_h|\boldsymbol{\Pi} = \boldsymbol{\pi}) = -n\pi_r\pi_h$

Thanks to the well-known laws of total expectation, total variance and total covariance, it is easy to define the first two orders moments of $\mathbf{X}$ whatever $\mathcal{F}$ is:

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbb{E}[\mathbf{X}|\boldsymbol{\Pi} = \boldsymbol{\pi}]] = \mathbb{E}[n\boldsymbol{\Pi}] = n\mathbb{E}[\boldsymbol{\Pi}] \tag{6.3}$$

$$\begin{aligned} \text{Var}(X_r) &= \mathbb{E}[\text{Var}(X_r|\boldsymbol{\Pi} = \boldsymbol{\pi})] + \text{Var}(\mathbb{E}[X_r|\boldsymbol{\Pi} = \boldsymbol{\pi}]) \\ &= \mathbb{E}[n\Pi_r(1 - \Pi_r)] + \text{Var}(n\Pi_r) \\ &= (n^2 - n)\text{Var}(\Pi_r) + n\mathbb{E}[\Pi_r](1 - \mathbb{E}[\Pi_r]) \\ &= n\left(n\text{Var}(\Pi_r) + \mathbb{E}[\Pi_r] - \mathbb{E}\left[\Pi_r^2\right]\right) \end{aligned} \tag{6.4}$$

$$\begin{aligned}
\text{Cov}\left(X_r, X_h\right) &= \mathbb{E}\left[\text{Cov}\left(X_r, X_h|\boldsymbol{\Pi}=\boldsymbol{\pi}\right)\right] + \text{Cov}\left(\mathbb{E}\left[X_r|\boldsymbol{\Pi}=\boldsymbol{\pi}\right], \mathbb{E}\left[X_h|\boldsymbol{\Pi}=\boldsymbol{\pi}\right]\right) \\
&= \mathbb{E}\left[-n\Pi_r\Pi_h\right] + \text{Cov}\left(n\Pi_r, n\Pi_h\right) \\
&= (n^2-n)\text{Cov}\left(\Pi_r, \Pi_h\right) - n\mathbb{E}\left[\Pi_r\right]\mathbb{E}\left[\Pi_h\right] \\
&= n\left(n\text{Cov}\left(\Pi_r, \Pi_h\right) - \mathbb{E}\left[\Pi_r \cdot \Pi_h\right]\right)
\end{aligned} \tag{6.5}$$

It is also possible to define the correlation coefficient among two distinct components of $\mathbf{X}$:

$$\rho_{X_r,X_h} = \frac{(n-1)\text{Cov}\left(\Pi_r, \Pi_h\right) - \mathbb{E}\left[\Pi_r\right]\mathbb{E}\left[\Pi_h\right]}{\sqrt{\left[(n-1)\text{Var}\left(\Pi_r\right) + \mathbb{E}\left[\Pi_r\right]\left(1-\mathbb{E}\left[\Pi_r\right]\right)\right] \cdot \left[(n-1)\text{Var}\left(\Pi_h\right) + \mathbb{E}\left[\Pi_h\right]\left(1-\mathbb{E}\left[\Pi_h\right]\right)\right]}} \tag{6.6}$$

These results are extremely general and does not depend on the specific distribution imposed on $\boldsymbol{\Pi}$. Evaluating (6.6) for $n = 1$, the resulting correlation in always negative:

$$\rho_{X_r,X_h}|_{n=1} = -\sqrt{\frac{\mathbb{E}\left[\Pi_r\right]\mathbb{E}\left[\Pi_h\right]}{\left(1-\mathbb{E}\left[\Pi_r\right]\right)\left(1-\mathbb{E}\left[\Pi_h\right]\right)}} < 0. \tag{6.7}$$

This fact is not surprising: if $n = 1$ then only one element of $\mathbf{X}$ can be different from 0 and, therefore, negative dependence should exist. Moreover, if $D = 2$, (6.7) is equal to -1, since that is a situation of perfect (negative) linear dependence among the two distinct elements of $\mathbf{X} = (X_1, X_2)^\intercal$. It is also easy to show that:

$$\lim_{n\to+\infty} \rho_{X_r,X_h} = \frac{n\text{Cov}\left(\Pi_r, \Pi_h\right)}{\sqrt{n\text{Var}\left(\Pi_r\right) n\text{Var}\left(\Pi_h\right)}} = \frac{\text{Cov}\left(\Pi_r, \Pi_h\right)}{\sqrt{\text{Var}\left(\Pi_r\right)\text{Var}\left(\Pi_h\right)}} = \rho_{\Pi_r,\Pi_h}. \tag{6.8}$$

With $n = 1$, correlation among counts in distinct categories has negative value. Increasing the value of n, this correlation tends to the correlation among the underling probabilities. This means that the choice of $\mathcal{F}$ is decisive: if $\mathcal{F}$ allows for $\text{Cov}\left(\Pi_r, \Pi_h\right) > 0$, then $\rho_{X_r,X_h} > 0$ for large enough values of n, otherwise correlation are always negative.

6.2 The Dirichlet-Multinomial distribution

A very popular choice for $\mathcal{F}$ is the Dirichlet distribution. As recalled in Section 3.1, the random vector $\mathbf{\Pi}$ is Dirichlet distributed if and only if its probability density function can be written as:

$$g(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha^+)}{\prod_{r=1}^{D} \Gamma(\alpha_r)} \prod_{r=1}^{D} \pi_r^{\alpha_r - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^\intercal$ is a vector of positive real numbers and $\alpha^+ = \sum_{r=1}^{D} \alpha_r$. Then its moments are: $\mathbb{E}[\Pi_r] = \frac{\alpha_r}{\alpha^+}$, $\text{Var}(\Pi_r) = \frac{\alpha_r}{\alpha^+}\left(1 - \frac{\alpha_r}{\alpha^+}\right)\frac{1}{(\alpha^+ + 1)}$ and $\text{Cov}(\Pi_r, \Pi_h) = -\frac{\alpha_r}{\alpha^+}\frac{\alpha_h}{\alpha^+}\frac{1}{(\alpha^+ + 1)}$ for every $r, h = 1, \ldots, D$ $(r \neq h)$. Covariances (and correlations) among any two distinct components of a Dirichlet distribution are negative.

Let $\mathbf{X}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ and $\mathbf{\Pi} \sim \text{Dir}(\boldsymbol{\alpha})$. Then the probability mass function of $\mathbf{X}$ can be obtained in the following way:

$$\begin{aligned} f_{DM}(\mathbf{x}; \boldsymbol{\alpha}) &= P(\mathbf{X} = \mathbf{x}) \\ &= \int_{\mathcal{S}^D} \frac{n!}{x_1! \ldots x_D!} \prod_{r=1}^{D} \pi_r^{x_r} \cdot \frac{\Gamma(\alpha^+)}{\prod_{r=1}^{D} \Gamma(\alpha_r)} \prod_{r=1}^{D} \pi_r^{\alpha_r - 1} d\boldsymbol{\pi} \\ &= \frac{n!\,\Gamma(\alpha^+)}{\Gamma(n + \alpha^+)} \prod_{r=1}^{D} \frac{\Gamma(x_r + \alpha_r)}{x_r!\,\Gamma(\alpha_r)} \end{aligned} \tag{6.9}$$

This probability function defines the so-called Dirichlet-Multinomial (DM) distribution (also known as Polya-Eggenberger distribution) with parameter n and $\boldsymbol{\alpha}$. It is more flexible than the Multinomial: they have the same number of parameters but the vector $\boldsymbol{\alpha}$ is not constrained to belong to the simplex as $\boldsymbol{\pi}$ does. Figure 6.1 shows the heat map of the DM probability function in some parametric configurations with $n = 50$.

The principal moments of a Dirichlet-Multinomial distribution are:

$$\mathbb{E}[X_r] = n \cdot \frac{\alpha_r}{\alpha^+} \tag{6.10}$$

$$\text{Var}(X_r) = \mathbb{E}[X_r]\,(n - \mathbb{E}[X_r]) \left[\frac{n + \alpha^+}{n(\alpha^+ + 1)}\right] \tag{6.11}$$

$$\text{Cov}(X_r, X_h) = -n \cdot \frac{\alpha_r}{\alpha^+}\frac{\alpha_h}{\alpha^+}\left(\frac{n + \alpha^+}{\alpha^+ + 1}\right) \tag{6.12}$$

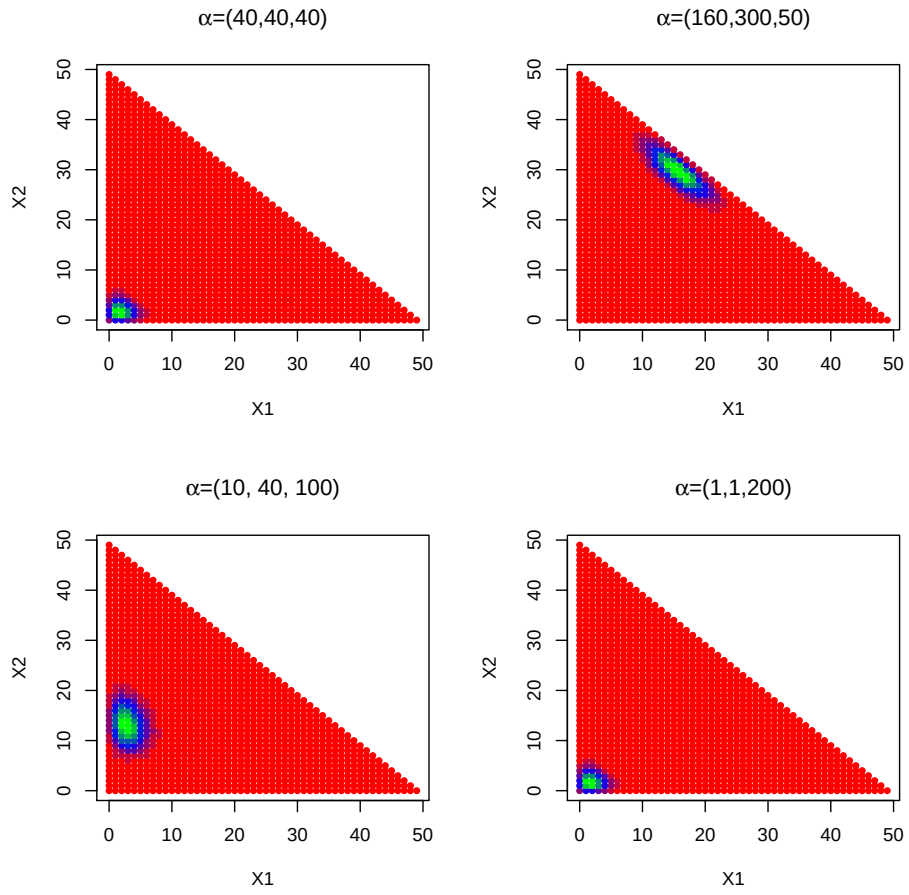

Fig. 6.1: Heat maps of DM probability function with $n = 50$.

It is interesting to note that, despite the DM provides several advantages with respect to the Multinomial distribution (i.e. more flexible covariance structure and ability of capture burstiness in text document classification, that is the phenomenon that most words never appear in any document, but in the moment they appear once, they are likely to appear several times [65]), it allows only for negative correlations. Indeed, covariances are negative and, furthermore, the correlation coefficient ρ_{X_r,X_h} is not function of n and it is equal to ρ_{Π_r,Π_h}:

$$\begin{aligned}
\rho_{X_r,X_h} &= \frac{-n \cdot \frac{\alpha_r}{\alpha^+}\frac{\alpha_h}{\alpha^+}\left(\frac{n+\alpha^+}{1+\alpha^+}\right)}{\sqrt{n \cdot \frac{\alpha_r}{\alpha^+}\left(1-\frac{\alpha_r}{\alpha^+}\right)\left(\frac{n+\alpha^+}{1+\alpha^+}\right) n \cdot \frac{\alpha_h}{\alpha^+}\left(1-\frac{\alpha_h}{\alpha^+}\right)\left(\frac{n+\alpha^+}{1+\alpha^+}\right)}} \\
&= -\sqrt{\frac{\alpha_r \alpha_h}{(\alpha^+ - \alpha_r)(\alpha^+ - \alpha_h)}} \\
&= \rho_{\Pi_r,\Pi_h}.
\end{aligned} \tag{6.13}$$

The R package **MGLM** [91] is based on the work of Zhang et al. [92] and allows to fit several kind of distributions to count data. One of these model is the Dirichlet-Multinomial model; hence it is possible to obtain an estimate for the parameter vector $\boldsymbol{\alpha}$.

6.3 Changing the distribution of $\boldsymbol{\Pi}$: the EFD-Multinomial distribution

The Multinomial and the Dirichlet-Multinomial distributions share the assumption that counts can be only negatively correlated. This is a strong restriction to data analysis and the necessity for a more flexible proposal arose [17, 92]. In this Section, a new proposal is provided, relying on a compound distribution we obtain by assuming that $\boldsymbol{\Pi} \sim \text{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$. The EFD has been chosen because the possibility of having positive covariances among Π_r and Π_h will make possible a positive association between categories of counts.

Definition 23. *Let $\boldsymbol{X}|\boldsymbol{\Pi} = \boldsymbol{\pi} \sim$ Multinomial$(n, \boldsymbol{\pi})$ and $\boldsymbol{\Pi} \sim$ EFD$(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$. Then the marginal distribution of $\boldsymbol{X}$ is called **Extended Flexible Dirichlet-Multinomial** with parameters $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$ and $\boldsymbol{p}$ (i.e. $\boldsymbol{X} \sim$ EFDM$(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$).*

The probability function of the EFDM distribution is complicated; it is much more interesting to note that it can me expressed as finite mixture of particular Dirichlet-Multinomial components:

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathcal{S}^D} f_{\mathbf{X},\boldsymbol{\Pi}}(\mathbf{x}, \boldsymbol{\pi}) d\boldsymbol{\pi} \\
&= \frac{n!}{x_1! \dots x_D!} \sum_{i=1}^{D} p_i \frac{\Gamma(\alpha^+ + \tau_i)\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)\prod_{r=1}^{D}\Gamma(\alpha_r)} \int_{\mathcal{S}^D} \pi_i^{\tau_i} \prod_{r=1}^{D} \pi_r^{\alpha_r - 1} d\boldsymbol{\pi} \\
&= \frac{n!}{x_1! \dots x_D!} \sum_{i=1}^{D} p_i \frac{\Gamma(\alpha^+ + \tau_i)\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)\prod_{r=1}^{D}\Gamma(\alpha_r)} \frac{\Gamma(\alpha_i + \tau_i + x_i)\prod_{r=1}^{D}\Gamma(\alpha_r + x_r)}{\Gamma(\alpha^+ + \tau_i + n)\Gamma(\alpha_i + x_i)} \\
&= \sum_{i=1}^{D} p_i \frac{n!\,\Gamma(\alpha^+ + \tau_i)}{\Gamma(n + \alpha^+ + \tau_i)} \left(\prod_{r=1}^{D} \frac{\Gamma(\alpha_r + x_r)}{x_r!\,\Gamma(\alpha_r)} \right) \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)} \frac{\Gamma(\alpha_i + \tau_i + x_i)}{\Gamma(\alpha_i + x_i)} \\
&= \sum_{i=1}^{D} p_i f_{DM}(\mathbf{x}; \boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i),
\end{aligned}
\tag{6.14}
$$

where $\mathbf{e}_i$ is the canonical vector with all entries equal to zero and the i-th equal to 1. Figure 6.2 shows the heat map of the EFDM probability function in some parametric configurations.

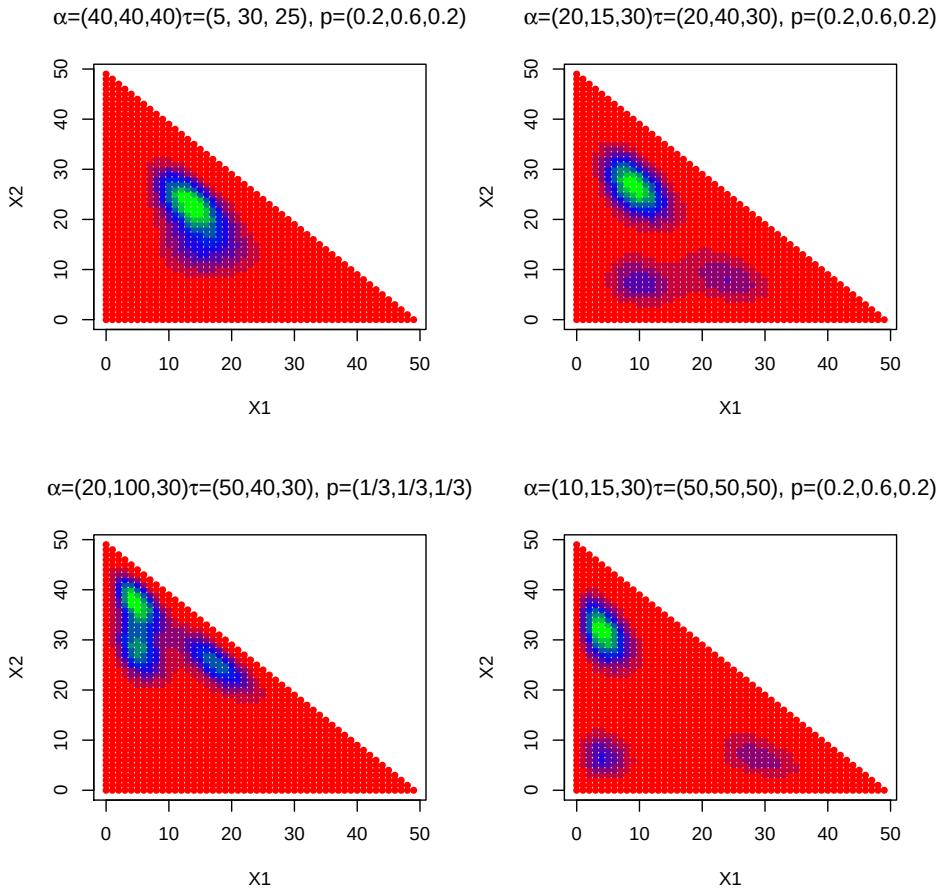

Fig. 6.2: Heat maps of EFDM probability function with $n = 50$.

Recall from (3.44)-(3.46) that, if $\mathbf{\Pi} \sim \text{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$, then:

$$\mathbb{E}\left[\Pi_r\right] = \alpha_r \, k_1 + \frac{p_r \tau_r}{\alpha^+ + \tau_r}$$

$$\text{Var}\left(\Pi_r\right) = \alpha_r^2 (k_2 - k_1^2) + \alpha_r k_2 + p_r \frac{\tau_r (2\alpha_r + \tau_r + 1)}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)} - \frac{p_r^2 \tau_r^2}{(\alpha^+ + \tau_r)^2} - k_1 \frac{2\alpha_r p_r \tau_r}{\alpha^+ + \tau_r}$$

$$\text{Cov}\left(\Pi_r, \Pi_h\right) = \alpha_r \alpha_h (k_2 - k_1^2) + \frac{\alpha_h p_r \tau_r}{\alpha^+ + \tau_r}\left(\frac{1}{\alpha^+ + \tau_r + 1} - k_1\right) + \\ + \frac{\alpha_r p_h \tau_h}{\alpha^+ + \tau_h}\left(\frac{1}{\alpha^+ + \tau_h + 1} - k_1\right) - \frac{p_r p_h \tau_r \tau_h}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_h)}$$

where $k_1 = \sum_{r=1}^{D} \frac{p_r}{\alpha^+ + \tau_r}$ and $k_2 = \sum_{r=1}^{D} \frac{p_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)}$.

From these moments and thanks to the rules (6.3)-(6.5), it is possible to derive the first two orders moments of the EFD-Multinomial distribution:

$$\mathbb{E}\left[X_r\right] = n \cdot \left[\alpha_r \, k_1 + \frac{p_r \tau_r}{\alpha^+ + \tau_r}\right] \tag{6.15}$$

$$\text{Var}\left(X_r\right) = \mathbb{E}\left[X_r\right]\left(1 - \mathbb{E}\left[X_r\right]\right) + (n^2 - n)\left\{\alpha_r k_2 \left(\alpha_r + 1\right) + \frac{p_r \tau_r (2\alpha_r + \tau_r + 1)}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)}\right\} \tag{6.16}$$

$$\text{Cov}\left(X_r, X_h\right) = \alpha_r \alpha_h \left[n^2 (k_2 - k_1^2) - n k_2\right] + \\ - n^2 k_1 \left(\frac{\alpha_r p_h \tau_h}{\alpha^+ + \tau_h} + \frac{\alpha_h p_r \tau_r}{\alpha^+ + \tau_r}\right) - \frac{n^2 p_r p_h \tau_r \tau_h}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_h)} + \\ + (n^2 - n)\left[\frac{\alpha_r p_h \tau_h}{(\alpha^+ + \tau_h)(\alpha^+ + \tau_h + 1)} + \frac{\alpha_h p_r \tau_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)}\right] \tag{6.17}$$

Covariance (6.3) is complicated to study, but it is possible to see that some parametric configurations allow for positive covariance, as in Example 13

Example 13. *Let* $\boldsymbol{\alpha} = (4, 6, 19)^\intercal$, $\boldsymbol{\tau} = (5, 1, 42)^\intercal$ *and* $\boldsymbol{p} = (0.23, 0.12, 0.65)^\intercal$. *Then the correlation matrices of* $\boldsymbol{X} \sim EFDM(n, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{p})$ *for* $n = 50$, $n = 500$ *and* $n \to +\infty$ *are:*

$n = 50$	X_1	X_2	X_3
X_1	1.000	0.296	-0.849
X_2	0.296	1.000	-0.756
X_3	-0.849	-0.756	1.000

$\longrightarrow$

$n = 500$	X_1	X_2	X_3
X_1	1.000	0.410	-0.885
X_2	0.410	1.000	-0.788
X_3	-0.885	-0.788	1.000

$\longrightarrow$

$n \to +\infty$	X_1	X_2	X_3
X_1	1.000	0.426	-0.889
X_2	0.426	1.000	-0.793
X_3	-0.889	-0.793	1.000

Note that, because of (6.8), the Table associated to $n \to +\infty$ *coincides with the correlation matrix of an EFD distribution with the same values for* $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$ *and* $\boldsymbol{p}$.

In order to get Maximum Likelihood estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$ and $\mathbf{p}$, one can, once again, rely on the EM algorithm, as showed in the previous sections. Suppose that a sample of size N has been collected: $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal$. It is extremely important to note that the notation has changed with respect to previous sections: now the sample size is denoted with N, whereas n represents the sum of each unit. Let $z_{s,i}$ be an indicator such that:

$$z_{s,i} = \begin{cases} 1 \text{ if the } s\text{-th subject comes from the } i\text{-th component of the mixture} \\ 0 \text{ otherwise} \end{cases}$$

$i = 1, \ldots, D$ and $s = 1, \ldots, N$. This indicator represents the unobserved component label; then it is possible to define the complete-data log likelihood function:

$$\ln L_C(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{s=1}^{N} \sum_{i=1}^{D} z_{s,i} \left\{ \ln p_i + \ln f_{DM}(\mathbf{x}_s; \boldsymbol{\alpha} + \tau_i \, \mathbf{e}_i) \right\}. \tag{6.18}$$

In order to evaluate the performance of the EM algorithm, a simulation study has been conducted.

6.4 A simulation study

In this Sections, we present results from a simulation study aimed at investigating the behaviour of the EM algorithm. Four configurations for the vector $(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})^\intercal$ (Table 6.1) and four different values for the parameter n $(50, 100, 250$ and $500)$ have been considered.

Config.	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
1	15	15	15	20	20	20	0.3333	0.3333	0.3333
2	40	40	40	5	30	25	0.2	0.6	0.2
3	5	30	70	10	25	15	0.1	0.75	0.15
4	4	6	19	5	1	42	0.23	0.12	0.65

Tab. 6.1: Configurations of the vector $(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})^\intercal$ for the EFDM simulation study.

For each combination of $(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})^\intercal$ and n, 300 samples of size $N = 150$ have been generated and the EM algorithm has been applied to each. Higher sample size has been considered $(N = 300)$ just for $n = 50$ and 250. The EM uses as initial values the real ones, in order to look at the performance in the best case scenario. In the following subsections it is possible to find:

1. A ternary plot (the triangle represents now the $\{n, D\}$-simplex instead of the unitary one) and the scatterplots of component pairs (data are simulated considering $n = 250$ and $N = 150$).

2. The correlation matrix for $n = 50$, $n = 500$ and $n \to +\infty$.

3. A table reporting the results of the simulation for that particular scenario. This table contains:

   a) The true value of the parameters.

   b) The mean of the 300 estimates for each parameter.

   c) The median of the 300 estimates for each parameter.

   d) The Absolute Relative Bias (Arb), defined as in (4.34)

   e) The square root of the Mean Squared Error (MSE), defined as in (4.35). The square root has been considered because of the magnitude of the MSEs.

   f) The standard deviation of the 300 estimates for each parameter ("Boot. SE").

   g) The coverage of the approximated 95% confidence intervals.

For each configuration, the results are quite unsatisfactory. The Arbs are too high and the coverage levels are too far from nominal value $1 - \alpha = 0.95$. Furthermore, not always increasing the sample size leads to a better results. Inspecting the results of the simulations more carefully, it is possible to note that, despite the EM uses the true parameters as starting point, the final estimates are quite far from the true values.

As a practical example, let us focus on a dataset generated according to the first configuration with $n = 50$. Table 6.2 compares the true parameters and their estimates:

	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
Estimate	29.02	28.63	29.26	43.28	40.75	43.20	0.376	0.312	0.312

Tab. 6.2: Comparison between true parameters and final estimates. Simulated data from the first configuration and $n = 50$.

It is well-known that the EM algorithm maximizes the complete-data log-likelihood function. Figure 6.3 illustrates that the complete-data log-likelihood function increases at each iteration of the algorithm, as expected. Nonetheless, the log-likelihood that one should maximize is the mixture one, defined as:

$$\ln L_M(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \ln \left(\prod_{s=1}^{N} f_{EFDM}(\mathbf{x}_s; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) \right). \tag{6.19}$$

The mixture log-likelihood increases in the early steps of the algorithm but then it decreases, converging to a point that is not a maxima.

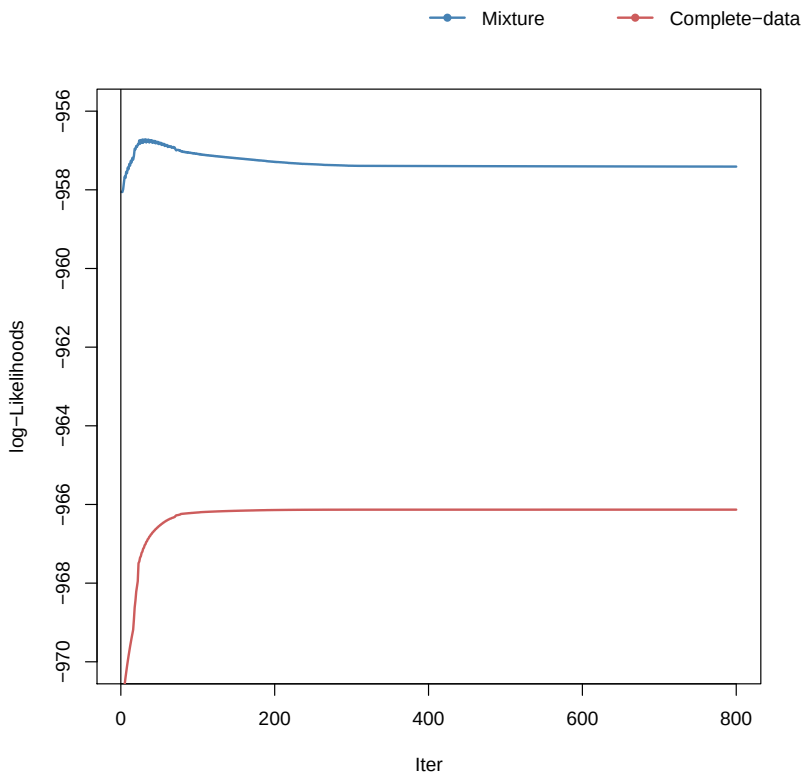

Fig. 6.3: Mixture and Complete-data log-likelihoods as function of the EM's iteration.

Evaluating both the complete-data log-likelihood and the mixture log-likelihood at the true and estimated parameters, one obtains the values reported in Table 6.3:

Param. \ log-lik.	Mixture	Complete-data
True	-959.5013	-974.3026
Estimates	-957.4180	-966.1317

Tab. 6.3: Mixture and complete-data log-likelihoods evaluated at the original and estimated parameters.

The final estimates lead to an higher value of the complete-data log-likelihood as well as to a lower value of the mixture likelihood. Since the EFDM can be expressed

as a finite mixture of particular DM components and these have at most a finite mode, it is clear that the issue is in the estimation procedure and not in the distribution (i.e. if the components have more than one mode, the log-likelihood could have several maximizers).
Two possible ways to deal with this issue are:

- Define a different estimation approach not involving the complete-data log-likelihood.

- Define a Bayesian procedure.

The latter approach is going to be considered in next section.

6.4.1 First configuration

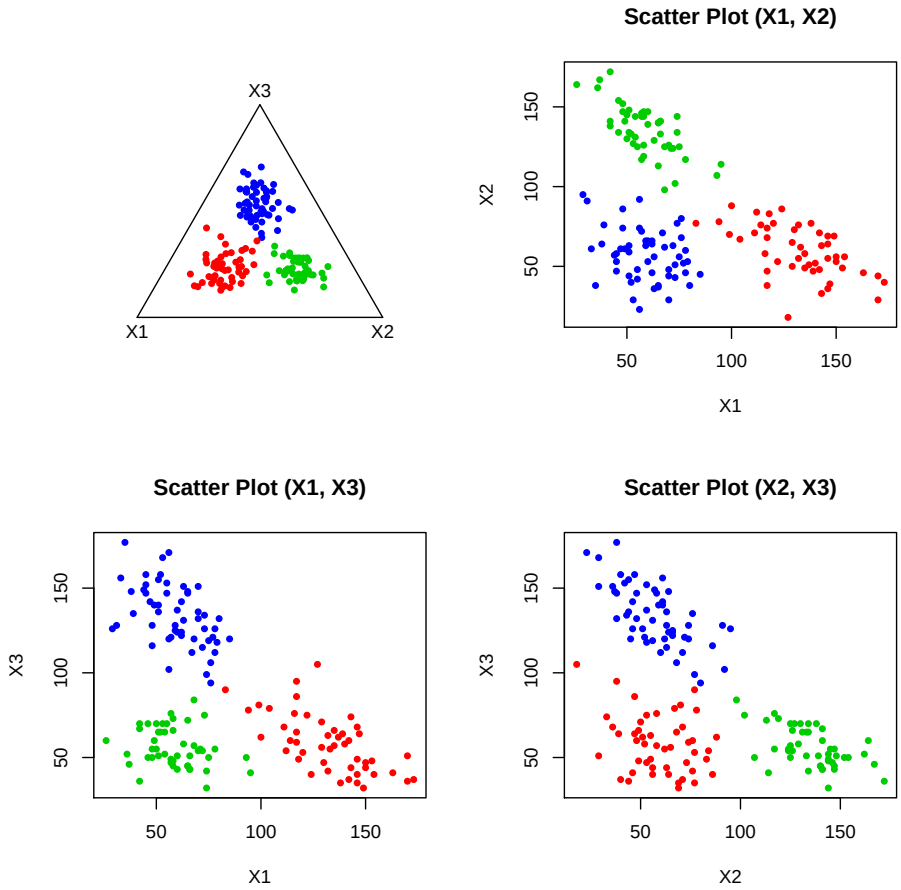

Fig. 6.4: First Configuration plots - $n = 250$.

$n = 50$	X_1	X_2	X_3
X_1	1.000	-0.500	-0.500
X_2	-0.500	1.000	-0.500
X_3	-0.500	-0.500	1.000

$\longrightarrow$

$n = 500$	X_1	X_2	X_3
X_1	1.000	-0.500	-0.500
X_2	-0.500	1.000	-0.500
X_3	-0.500	-0.500	1.000

$\longrightarrow$

$n \to +\infty$	X_1	X_2	X_3
X_1	1.000	-0.500	-0.500
X_2	-0.500	1.000	-0.500
X_3	-0.500	-0.500	1.000

Tab. 6.4: Correlation matrix for $n = 50$, $n = 500$ and $n \to +\infty$ (EFD).

Sample size 150

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	23.767	23.761	23.783	33.316	33.214	33.535	0.333	0.335	0.332
MLE median	23.255	22.990	23.139	32.349	32.111	32.353	0.332	0.333	0.332
Arb	0.584	0.584	0.586	0.666	0.661	0.677	0.002	0.005	0.004
$\sqrt{\text{MSE}}$	10.243	10.339	10.382	15.593	15.369	15.559	0.039	0.041	0.041
Boot. SE	5.288	5.481	5.525	8.099	7.835	7.660	0.039	0.041	0.041
Coverage	0.653	0.693	0.683	0.69	0.647	0.640	0.95	0.94	0.95
$n = 100$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	17.866	17.821	17.818	24.321	24.361	24.384	0.333	0.335	0.332
MLE median	17.642	17.592	17.593	24.331	23.944	23.878	0.334	0.332	0.332
Arb	0.191	0.188	0.188	0.216	0.218	0.219	0.0001	0.004	0.003
$\sqrt{\text{MSE}}$	3.757	3.734	3.688	5.650	5.667	5.711	0.040	0.041	0.038
Boot. SE	2.425	2.443	2.376	3.635	3.613	3.654	0.040	0.041	0.038
Coverage	0.810	0.823	0.800	0.823	0.797	0.827	0.957	0.957	0.953
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	16.025	16.058	16.050	21.677	21.571	21.495	0.334	0.335	0.331
MLE median	16.039	15.968	16.043	21.474	21.485	21.477	0.333	0.333	0.334
Arb	0.068	0.071	0.070	0.084	0.079	0.075	0.003	0.004	0.006
$\sqrt{\text{MSE}}$	1.796	1.885	1.895	3.045	2.925	2.826	0.040	0.038	0.038
Boot. SE	1.472	1.558	1.574	2.538	2.463	2.395	0.040	0.037	0.038
Coverage	0.903	0.907	0.907	0.877	0.897	0.893	0.953	0.943	0.950
$n = 500$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	15.814	15.785	15.832	21.253	21.109	21.077	0.331	0.333	0.336
MLE median	15.663	15.670	15.738	21.080	21.131	20.911	0.332	0.329	0.336
Arb	0.054	0.052	0.055	0.063	0.055	0.054	0.006	0.002	0.007
$\sqrt{\text{MSE}}$	1.628	1.724	1.675	2.729	2.480	2.612	0.040	0.035	0.038
Boot. SE	1.408	1.532	1.452	2.420	2.215	2.376	0.039	0.035	0.038
Coverage	0.897	0.913	0.907	0.903	0.923	0.910	0.937	0.973	0.950

Tab. 6.5: Simulation results for the first configuration - $N = 150$.

Sample size 300

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	22.921	22.916	22.955	32.065	31.944	31.947	0.332	0.334	0.334
MLE median	22.787	22.987	22.940	31.858	31.739	31.702	0.331	0.334	0.335
Arb	0.528	0.528	0.530	0.603	0.597	0.597	0.004	0.002	0.002
$\sqrt{\text{MSE}}$	8.397	8.384	8.416	12.696	12.563	12.619	0.025	0.023	0.022
Boot. SE	2.782	2.758	2.744	3.948	3.889	4.055	0.025	0.023	0.022
Coverage	0.203	0.200	0.183	0.133	0.127	0.157	0.943	0.947	0.957
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	15	15	15	20	20	20	0.333	0.333	0.333
MLE mean	15.986	16.014	15.990	21.470	21.382	21.401	0.333	0.334	0.334
MLE median	15.945	15.987	15.970	21.336	21.217	21.471	0.332	0.333	0.333
Arb	0.066	0.068	0.066	0.073	0.069	0.070	0.002	0.001	0.001
$\sqrt{\text{MSE}}$	1.431	1.455	1.459	2.346	2.216	2.100	0.028	0.027	0.029
Boot. SE	1.036	1.041	1.070	1.825	1.729	1.561	0.028	0.027	0.028
Coverage	0.850	0.853	0.850	0.857	0.880	0.857	0.953	0.963	0.953

Tab. 6.6: Simulation results for the first configuration - $N = 300$.

6.4.2 Second configuration

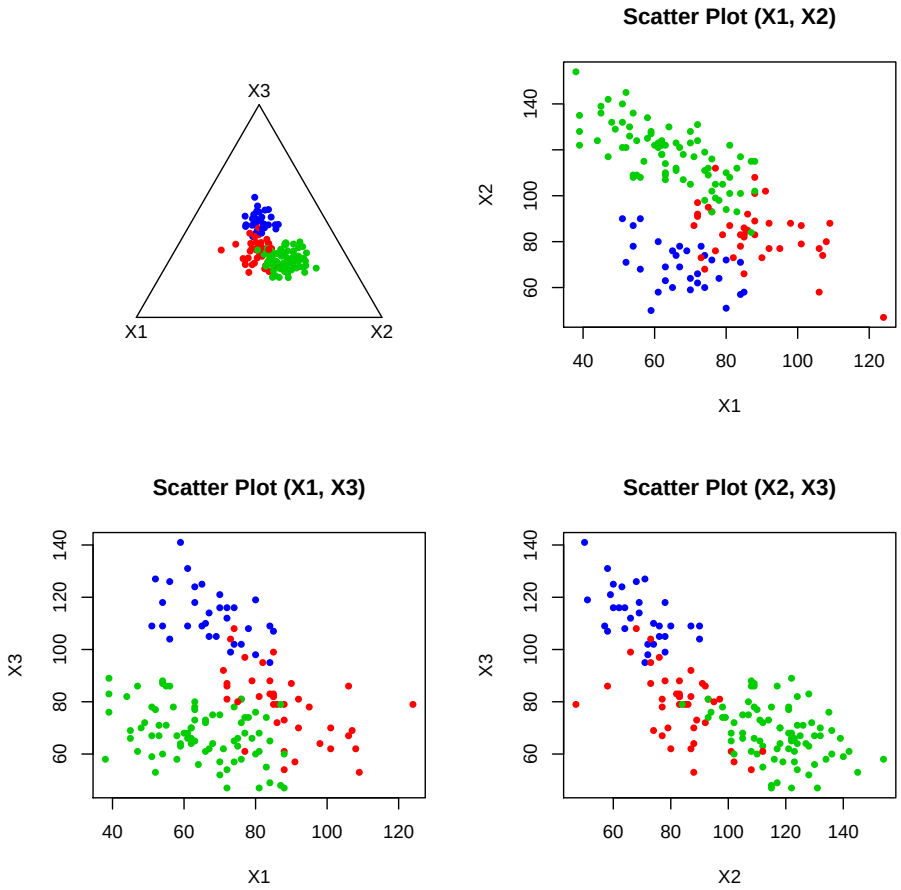

Fig. 6.5: Second Configuration plots - $n = 250$.

$n = 50$	X_1	X_2	X_3
X_1	1.000	-0.498	-0.243
X_2	-0.498	1.000	-0.720
X_3	-0.243	-0.720	1.000

$\longrightarrow$

$n = 500$	X_1	X_2	X_3
X_1	1.000	-0.506	-0.096
X_2	-0.506	1.000	-0.810
X_3	-0.096	-0.810	1.000

$\longrightarrow$

$n \to +\infty$	X_1	X_2	X_3
X_1	1.000	-0.510	-0.064
X_2	-0.510	1.000	-0.826
X_3	-0.064	-0.826	1.000

Tab. 6.7: Correlation matrix for $n = 50$, $n = 500$ and $n \to +\infty$ (EFD).

Sample size 150

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	30.166	22.458	30.295	46.919	24.821	38.371	0.009	0.923	0.068
MLE median	13.906	9.250	14.916	37.129	14.197	32.627	0	0.993	0.000
Arb	0.246	0.439	0.243	8.384	0.174	0.535	0.956	0.538	0.658
$\sqrt{\text{MSE}}$	49.162	49.371	49.206	59.151	41.322	42.242	0.196	0.343	0.169
Boot. SE	48.088	46.073	48.159	41.664	40.928	40.003	0.041	0.114	0.106
Coverage	0.977	0.980	0.977	0.913	0.970	0.980	0.013	0.213	0.987
$n = 100$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	64.151	60.684	63.620	45.635	49.334	40.292	0.057	0.694	0.249
MLE median	51.507	46.597	49.524	43.573	40.121	28.167	0.013	0.687	0.246
Arb	0.604	0.517	0.590	8.127	0.644	0.612	0.714	0.156	0.246
$\sqrt{\text{MSE}}$	54.894	55.341	55.474	44.342	44.934	41.366	0.169	0.135	0.103
Boot. SE	49.214	51.244	50.111	17.717	40.495	38.371	0.091	0.098	0.090
Coverage	0.950	0.950	0.950	0.37	0.950	0.940	0.433	0.830	0.913
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	55.684	55.177	55.860	17.483	42.280	36.426	0.158	0.620	0.222
MLE median	54.582	53.813	55.026	16.533	41.926	36.290	0.161	0.618	0.220
Arb	0.392	0.379	0.396	2.497	0.409	0.457	0.211	0.033	0.111
$\sqrt{\text{MSE}}$	19.709	19.782	20.010	15.144	16.079	16.797	0.092	0.059	0.067
Boot. SE	11.916	12.666	12.181	8.559	10.363	12.291	0.081	0.056	0.063
Coverage	0.787	0.820	0.783	0.727	0.803	0.870	0.887	0.940	0.930
$n = 500$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	50.135	50.093	50.229	12.103	37.660	32.817	0.180	0.612	0.208
MLE median	49.854	50.867	50.517	11.953	37.585	32.840	0.178	0.613	0.205
Arb	0.253	0.252	0.256	1.421	0.255	0.313	0.100	0.021	0.038
$\sqrt{\text{MSE}}$	11.857	12.075	11.970	8.596	9.413	10.336	0.066	0.050	0.049
Boot. SE	6.145	6.618	6.206	4.833	5.462	6.751	0.062	0.048	0.049
Coverage	0.613	0.667	0.607	0.760	0.693	0.783	0.933	0.940	0.937

Tab. 6.8: Simulation results for the second configuration - $N = 150$.

Sample size 300

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	20.594	13.046	21.023	51.129	18	31.671	0.001	0.955	0.043
MLE median	11.685	5.282	12.525	43.956	11.942	30.646	0	1	0.000
Arb	0.485	0.674	0.474	9.226	0.400	0.267	0.994	0.592	0.783
$\sqrt{\text{MSE}}$	31.718	35.279	30.611	53.948	23.689	13.175	0.199	0.368	0.182
Boot. SE	25.047	22.724	23.979	27.926	20.391	11.343	0.007	0.096	0.093
Coverage	0.977	0.983	0.977	0.867	0.960	0.970	0	0.113	0.207
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	40	40	40	5	30	25	0.200	0.600	0.200
MLE mean	56.567	56.132	56.758	15.675	42.588	35.952	0.166	0.616	0.217
MLE median	55.922	55.391	56.040	15.683	42.019	36.101	0.163	0.617	0.215
Arb	0.414	0.403	0.419	2.135	0.420	0.438	0.169	0.027	0.087
$\sqrt{\text{MSE}}$	18.345	18.177	18.577	11.811	14.320	13.320	0.065	0.041	0.045
Boot. SE	7.865	8.363	8.004	5.046	6.815	7.569	0.055	0.037	0.041
Coverage	0.47	0.557	0.480	0.417	0.573	0.723	0.910	0.937	0.927

Tab. 6.9: Simulation results for the second configuration - $N = 300$.

6.4.3 Third configuration

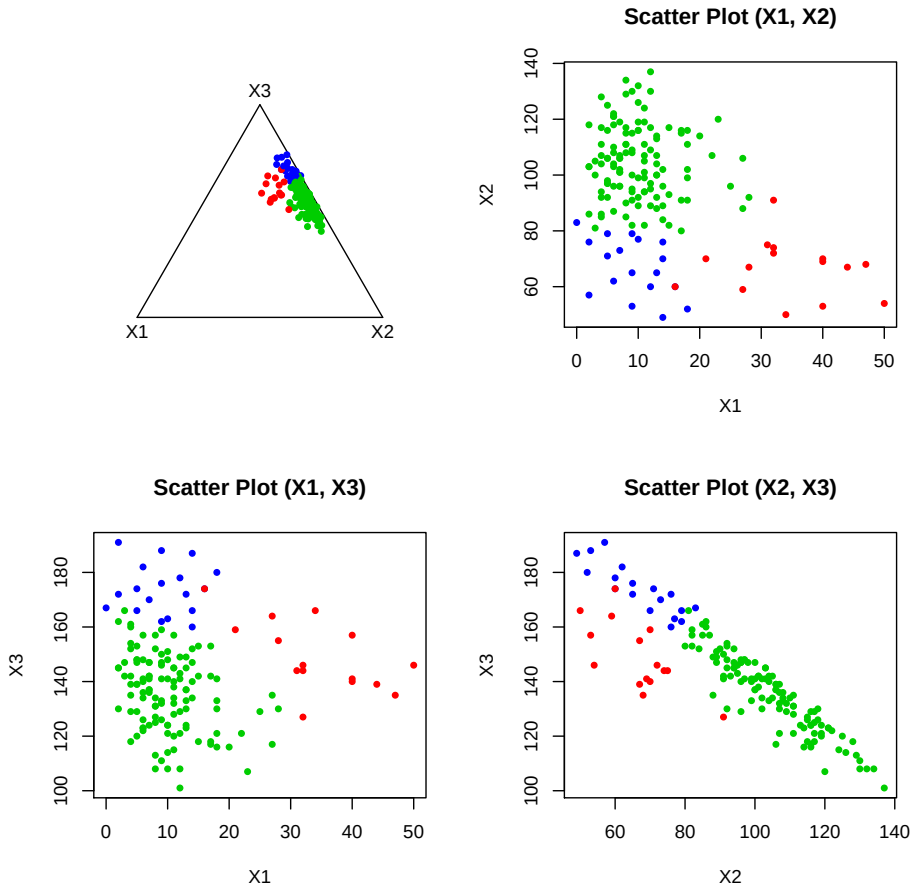

Fig. 6.6: Third Configuration plots - $n = 250$.

$n = 50$	X_1	X_2	X_3
X_1	1.000	-0.339	-0.080
X_2	-0.339	1.000	-0.911
X_3	-0.080	-0.911	1.000

$\longrightarrow$

$n = 500$	X_1	X_2	X_3
X_1	1.000	-0.442	0.049
X_2	-0.442	1.000	-0.918
X_3	0.049	-0.918	1.000

$\longrightarrow$

$n \to +\infty$	X_1	X_2	X_3
X_1	1.000	-0.462	0.076
X_2	-0.462	1.000	-0.919
X_3	0.076	-0.919	1.000

Tab. 6.10: Correlation matrix for $n = 50$, $n = 500$ and $n \to +\infty$ (EFD).

Sample size 150

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	3.612	13.781	47.126	16.405	20.250	52.698	0.024	0.961	0.015
MLE median	2.319	7.223	27.261	13.941	13.378	42.633	0.007	0.989	0.000
Arb	0.278	0.541	0.327	0.640	0.190	2.513	0.761	0.281	0.898
$\sqrt{\text{MSE}}$	4.798	30.852	71.494	16.581	29.237	54.627	0.084	0.219	0.140
Boot. SE	4.585	26.201	67.624	15.268	28.801	39.469	0.036	0.059	0.038
Coverage	0.970	0.977	0.977	0.960	0.967	0.897	0.283	0.100	0.053
$n = 100$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	6.844	38.989	95.268	18.651	34.805	46.171	0.081	0.818	0.101
MLE median	6.160	34.718	86.695	17.118	31.472	45.624	0.081	0.807	0.106
Arb	0.369	0.300	0.361	0.865	0.392	2.078	0.191	0.090	0.324
$\sqrt{\text{MSE}}$	4.015	25.973	58.066	12.073	22.137	37.522	0.048	0.106	0.080
Boot. SE	3.561	24.327	52.193	8.407	19.814	20.853	0.044	0.082	0.063
Coverage	0.923	0.937	0.937	0.843	0.940	0.770	0.923	0.857	0.830
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	6.123	36.933	86.361	14.081	30.658	27.453	0.100	0.757	0.142
MLE median	6.029	36.601	85.676	13.795	30.231	27.398	0.099	0.755	0.142
Arb	0.225	0.231	0.234	0.408	0.226	0.830	0.004	0.010	0.050
$\sqrt{\text{MSE}}$	1.436	9.104	20.524	4.834	7.979	15.814	0.031	0.044	0.041
Boot. SE	0.894	5.890	12.371	2.586	5.617	9.730	0.031	0.043	0.041
Coverage	0.770	0.797	0.740	0.670	0.850	0.773	0.963	0.963	0.953
$n = 500$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	5.705	34.859	80.314	12.404	28.114	24.502	0.100	0.757	0.143
MLE median	5.715	34.730	80.079	12.355	28.356	24.334	0.100	0.756	0.141
Arb	0.141	0.162	0.147	0.240	0.125	0.633	0.003	0.009	0.049
$\sqrt{\text{MSE}}$	0.950	6.496	13.546	3.094	5.109	11.928	0.027	0.042	0.039
Boot. SE	0.635	4.305	8.767	1.945	4.044	7.198	0.027	0.041	0.038
Coverage	0.803	0.783	0.783	0.787	0.880	0.767	0.940	0.927	0.937

Tab. 6.11: Simulation results for the third configuration - $N = 150$.

Sample size 300

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	4.998	21.402	66.956	20.756	28.674	62.178	0.024	0.952	0.025
MLE median	2.142	5.584	25.531	15.133	13.373	48.203	0.003	0.996	0.000
Arb	0.0001	0.287	0.043	1.076	0.147	3.145	0.764	0.269	0.835
$\sqrt{\text{MSE}}$	7.969	47.441	115.110	23.678	46.338	71.272	0.085	0.216	0.134
Boot. SE	7.955	46.577	114.878	21.059	46.115	53.333	0.037	0.076	0.048
Coverage	0.957	0.960	0.960	0.943	0.953	0.927	0.283	0.223	0.180
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	5	30	70	10	25	15	0.100	0.750	0.150
MLE mean	6.124	36.675	85.735	13.644	30.436	27.744	0.098	0.760	0.141
MLE median	6.100	36.597	84.983	13.532	30.114	27.426	0.099	0.760	0.143
Arb	0.225	0.222	0.225	0.364	0.217	0.850	0.015	0.014	0.060
$\sqrt{\text{MSE}}$	1.300	7.946	18.207	4.108	6.750	14.374	0.022	0.031	0.031
Boot. SE	0.652	4.303	9.145	1.893	3.996	6.636	0.022	0.029	0.029
Coverage	0.607	0.677	0.613	0.530	0.750	0.540	0.963	0.927	0.943

Tab. 6.12: Simulation results for the third configuration - $N = 300$.

6.4.4 Forth configuration

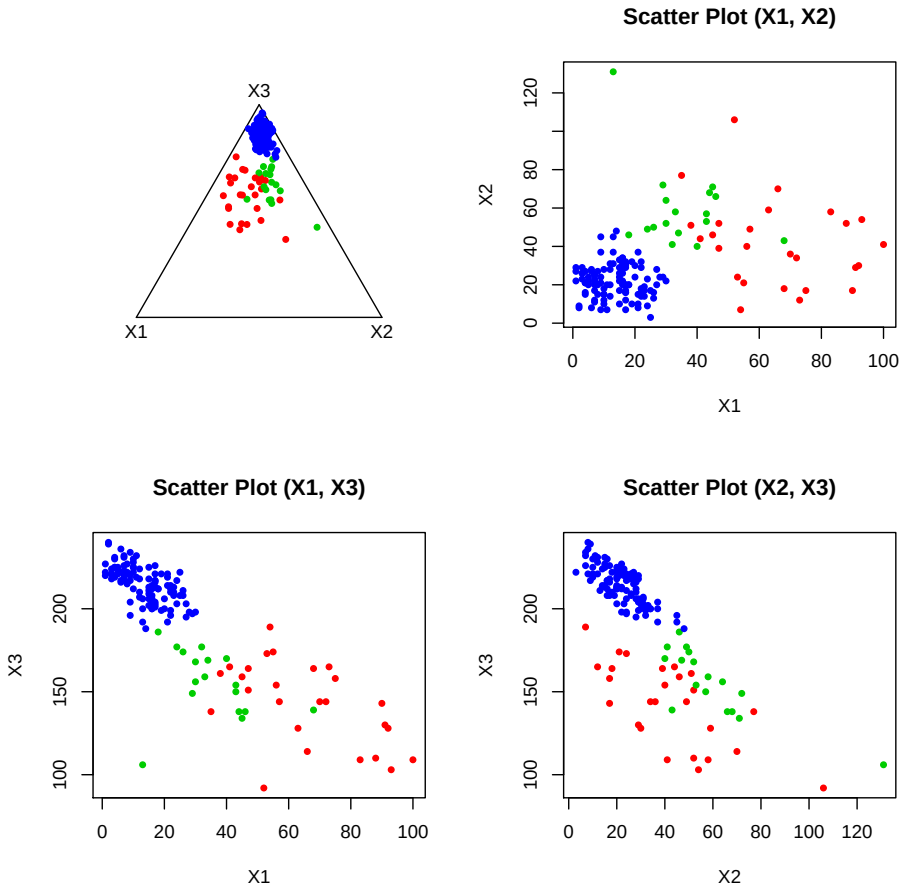

Fig. 6.7: Forth Configuration plots - $n = 250$.

$n = 50$	X_1	X_2	X_3
X_1	1.000	0.296	-0.849
X_2	0.296	1.000	-0.756
X_3	-0.849	-0.756	1.000

$\longrightarrow$

$n = 500$	X_1	X_2	X_3
X_1	1.000	0.410	-0.885
X_2	0.410	1.000	-0.788
X_3	-0.885	-0.788	1.000

$\longrightarrow$

$n \to +\infty$	X_1	X_2	X_3
X_1	1.000	0.426	-0.889
X_2	0.426	1.000	-0.793
X_3	-0.889	-0.793	1.000

Tab. 6.13: Correlation matrix for $n = 50$, $n = 500$ and $n \to +\infty$ (EFD).

Sample size 150

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	5.991	9.187	28.908	8.504	15.614	61.184	0.278	0.038	0.684
MLE median	5.301	8.204	24.764	6.290	14.195	57.154	0.279	0.010	0.676
Arb	0.498	0.531	0.521	0.701	14.614	0.457	0.208	0.682	0.052
$\sqrt{\text{MSE}}$	3.370	4.860	17.374	7.358	16.004	32.519	0.086	0.100	0.064
Boot. SE	2.714	3.663	14.248	6.460	6.512	26.213	0.072	0.057	0.054
Coverage	0.920	0.897	0.917	0.927	0.477	0.930	0.913	0.497	0.880
$n = 100$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	4.823	7.403	23.431	6.489	10.769	49.961	0.284	0.047	0.669
MLE median	4.617	7.149	22.178	5.586	10.253	47.965	0.293	0.018	0.668
Arb	0.369	0.300	0.361	0.865	0.392	2.078	0.191	0.090	0.324
$\sqrt{\text{MSE}}$	1.360	2.041	7.677	4.163	10.857	14.061	0.091	0.098	0.048
Boot. SE	1.081	1.480	6.259	3.881	4.728	11.570	0.073	0.066	0.044
Coverage	0.873	0.857	0.867	0.920	0.503	0.883	0.897	0.983	0.933
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	4.436	6.760	21.947	6.205	7.732	45.436	0.268	0.067	0.665
MLE median	4.364	6.706	21.842	5.834	7.432	44.562	0.276	0.048	0.669
Arb	0.109	0.127	0.155	0.241	6.732	0.082	0.163	0.441	0.024
$\sqrt{\text{MSE}}$	0.779	1.176	5.109	3.218	7.805	8.203	0.085	0.091	0.045
Boot. SE	0.645	0.895	4.167	2.979	3.943	7.437	0.076	0.073	0.042
Coverage	0.890	0.877	0.880	0.943	0.653	0.927	0.947	0.963	0.963
$n = 500$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	4.332	6.626	21.625	5.915	7.092	44.861	0.268	0.074	0.657
MLE median	4.281	6.578	21.720	5.729	6.413	44.345	0.268	0.070	0.658
Arb	0.083	0.104	0.138	0.183	6.092	0.068	0.167	0.380	0.011
$\sqrt{\text{MSE}}$	0.650	0.971	4.583	2.865	7.185	6.750	0.080	0.084	0.042
Boot. SE	0.557	0.741	3.751	2.710	3.804	6.104	0.070	0.071	0.041
Coverage	0.897	0.853	0.883	0.973	0.700	0.910	0.907	0.977	0.950

Tab. 6.14: Simulation results for the forth configuration - $N = 150$.

Sample size 300

$n = 50$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	5.371	8.397	25.114	6.366	17.761	56.455	0.297	0.020	0.682
MLE median	5.133	8.017	22.868	5.271	16.713	54.967	0.304	0.004	0.685
Arb	0.343	0.400	0.322	0.273	16.761	0.344	0.293	0.832	0.050
$\sqrt{\text{MSE}}$	1.873	3.006	9.936	3.677	17.847	18.646	0.085	0.105	0.049
Boot. SE	1.274	1.810	7.819	3.408	6.121	11.758	0.051	0.033	0.036
Coverage	0.863	0.800	0.880	0.923	0.253	0.817	0.730	0.147	0.860
$n = 250$	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3
True	4	6	19	5	1	42	0.230	0.120	0.650
MLE mean	4.305	6.611	21.018	5.194	9.007	44.379	0.287	0.047	0.665
MLE median	4.225	6.522	20.985	5.079	8.339	43.677	0.290	0.034	0.667
Arb	0.076	0.102	0.106	0.039	8.007	0.057	0.249	0.606	0.024
$\sqrt{\text{MSE}}$	0.600	0.960	4.283	2.251	8.931	5.685	0.079	0.089	0.034
Boot. SE	0.515	0.739	3.771	2.239	3.948	5.155	0.054	0.051	0.030
Coverage	0.900	0.847	0.917	0.970	0.543	0.917	0.803	0.580	0.923

Tab. 6.15: Simulation results for the forth configuration - $N = 300$.

6.5 A Bayesian approach

A Bayesian approach has been proposed to produce estimates of the EFDM parameters. This procedure is based on Hamiltonian Monte Carlo (HMC) [14, 31, 66], that is a generalization of the Metropolis algorithm which incorporates both deterministic and MCMC simulation methods. The HMC has been implemented through the Stan modeling language [86]. To sample from the posterior distribution, Stan requires the probability density function (as in 6.14) and the prior distibutions for $\mathbf{p}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$. An interesting feature of Stan is that it makes possible to define prior also for transformed version of the parameters. We took advantage of this aspect and defined the following priors:

- $\mathbf{p} \sim \mathcal{D}(d_0, \ldots, d_0)$
- $\alpha^+ = \sum_{r=1}^{D} \alpha_r \sim \text{Gamma}(g_1, g_2)$
- $\tau^+ = \sum_{r=1}^{D} \tau_r \sim \text{Gamma}(h_1, h_2)$
- $\bar{\boldsymbol{\alpha}} = \dfrac{\boldsymbol{\alpha}}{\alpha^+} \sim \mathcal{D}(e_0, \ldots, e_0)$

- $\bar{\boldsymbol{\tau}} = \dfrac{\boldsymbol{\tau}}{\tau^+} \sim \mathcal{D}(f_0, \ldots, f_0)$

We set $d_0 = e_0 = f_0 = 1$ in order to have uniform priors on the simplex.

A small simulation study has been conducted. It refers to two hyperparameters configurations for the priors of α^+ and τ^+:

- Gamma priors with mean the true parameter and variance equal to 1 (informative priors)
- Gamma priors with mean the true parameter and variance equal to 500 (weakly informative priors).

In this simulation study, 200 samples of size 150 have been generated from the EFDM for four different parameter configurations:

1. $n = 50$, $\boldsymbol{\alpha} = (15, 15, 15)^\intercal$, $\boldsymbol{\tau} = (20, 20, 20)^\intercal$ and $\mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^\intercal$
2. $n = 250$, $\boldsymbol{\alpha} = (15, 15, 15)^\intercal$, $\boldsymbol{\tau} = (20, 20, 20)^\intercal$ and $\mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^\intercal$
3. $n = 50$, $\boldsymbol{\alpha} = (4, 6, 19)^\intercal$, $\boldsymbol{\tau} = (5, 1, 42)^\intercal$ and $\mathbf{p} = (0.23, 0.12, 0.65)^\intercal$
4. $n = 250$, $\boldsymbol{\alpha} = (4, 6, 19)^\intercal$, $\boldsymbol{\tau} = (5, 1, 42)^\intercal$ and $\mathbf{p} = (0.23, 0.12, 0.65)^\intercal$

Scenarios 1 and 2 refer to configuration 1 in Table 6.1 with $n = 50$ and $n = 250$ respectively, whereas scenarios 3 and 4 refer to configuration 4. We have chosen these configurations because we wanted to test our Bayesian procedure both in a scenario characterized by well-separated components and in a scenario characterized by positive correlations among counts.

The results encourage to prefer the Bayesian approach instead of the classical one. Indeed, the Arbs are smaller than the ones obtained with the EM algorithm (Tables 6.5 and 6.14). Estimates are close to the real values also using the weakly informative priors and this suggests that the Bayesian procedure is robust with respect to the choice of the hyperparameters of the Gamma priors of α^+ and τ^+. More intensive simulation studies need to be conducted, inspecting different parameter configurations and changing the priors for the unknown parameters.

6.5.1 Bayesian - Informative Priors

Scenarios 1 and 2:

$n = 50$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	0.333	15	15	15	20	20	20
Mean Post. Means	0.333	0.338	0.329	14.946	14.986	15.027	20.196	20.013	19.816
Mean Post. Medians	0.332	0.337	0.328	14.939	14.979	15.021	20.181	19.997	19.796
Arb	0.108	0.099	0.093	0.031	0.030	0.029	0.090	0.082	0.089
$\sqrt{MSE}$	0.043	0.041	0.039	0.560	0.557	0.537	2.254	2.065	2.186
Coverage	0.940	0.960	0.965	0.985	0.980	0.985	0.935	0.965	0.965
$n = 250$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	0.333	15	15	15	20	20	20
Mean Post. Means	0.334	0.337	0.329	14.967	15.002	14.966	19.967	20.143	19.921
Mean Post. Medians	0.333	0.336	0.328	14.963	14.998	14.962	19.960	20.136	19.913
Arb	0.100	0.094	0.091	0.023	0.023	0.022	0.061	0.060	0.060
$\sqrt{MSE}$	0.040	0.039	0.038	0.428	0.430	0.405	1.509	1.487	1.495
Coverage	0.950	0.945	0.950	0.985	0.975	0.995	0.945	0.945	0.960

Tab. 6.16: Simulation results for scenarios 1 and 2 with informative priors.

Scenarios 3 and 4:

$n = 50$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.23	0.12	0.65	4	6	19	5	1	42
Mean Post. Means	0.245	0.112	0.643	3.978	5.877	19.135	5.125	3.059	39.824
Mean Post. Medians	0.244	0.102	0.645	3.972	5.875	19.128	5.023	2.470	40.113
Arb	0.185	0.329	0.058	0.067	0.048	0.022	0.192	2.063	0.054
$\sqrt{MSE}$	0.054	0.049	0.045	0.328	0.363	0.543	1.236	2.500	2.760
Coverage	0.955	0.985	0.965	0.965	0.945	0.995	0.970	0.995	0.965
$n = 250$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.23	0.12	0.65	4	6	19	5	1	42
Mean Post. Means	0.246	0.111	0.643	3.987	5.922	19.047	4.964	2.330	40.733
Mean Post. Medians	0.245	0.106	0.644	3.985	5.921	19.039	4.921	1.990	40.864
Arb	0.187	0.305	0.052	0.046	0.033	0.020	0.181	1.346	0.036
$\sqrt{MSE}$	0.054	0.045	0.041	0.231	0.256	0.470	1.154	1.896	1.993
Coverage	0.930	0.960	0.950	0.965	0.955	1.000	0.940	0.990	0.975

Tab. 6.17: Simulation results for scenarios 3 and 4 with informative priors.

6.5.2 Bayesian - Weakly Informative Priors

Scenarios 1 and 2:

$n = 50$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	0.333	15	15	15	20	20	20
Mean Post. Means	0.333	0.338	0.329	15.180	15.235	15.280	20.570	20.380	20.215
Mean Post. Medians	0.332	0.337	0.328	14.921	14.971	15.016	20.185	19.992	19.823
Arb	0.108	0.099	0.093	0.105	0.110	0.110	0.143	0.136	0.139
$\sqrt{MSE}$	0.043	0.041	0.040	1.950	2.057	2.074	3.624	3.402	3.418
Coverage	0.940	0.960	0.960	0.990	0.990	0.995	0.990	0.990	0.990
$n = 250$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	0.333	15	15	15	20	20	20
Mean Post. Means	0.334	0.337	0.329	14.936	14.975	14.941	19.972	20.163	19.951
Mean Post. Medians	0.333	0.336	0.328	14.868	14.904	14.872	19.867	20.054	19.845
Arb	0.100	0.094	0.090	0.069	0.073	0.072	0.098	0.091	0.091
$\sqrt{MSE}$	0.040	0.039	0.038	1.307	1.361	1.360	2.393	2.295	2.320
Coverage	0.950	0.940	0.950	0.970	0.975	0.975	0.970	0.980	0.965

Tab. 6.18: Simulation results for scenarios 1 and 2 with weakly informative priors.

Scenarios 3 and 4:

$n = 50$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.23	0.12	0.65	4	6	19	5	1	42
Mean Post. Means	0.248	0.108	0.644	4.124	6.113	20.054	5.461	3.987	41.357
Mean Post. Medians	0.248	0.098	0.645	4.030	5.975	19.557	5.177	2.644	40.389
Arb	0.189	0.323	0.057	0.138	0.144	0.163	0.284	2.987	0.149
$\sqrt{MSE}$	0.055	0.047	0.045	0.706	1.071	3.887	1.847	3.504	7.776
Coverage	0.975	0.985	0.965	0.985	0.995	0.990	0.990	0.990	0.995
$n = 250$	p_1	p_2	p_3	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.23	0.12	0.65	4	6	19	5	1	42
Mean Post. Means	0.249	0.108	0.643	4.024	5.975	19.189	5.025	2.863	41.364
Mean Post. Medians	0.248	0.103	0.644	4.000	5.941	19.064	4.943	2.039	41.094
Arb	0.196	0.323	0.052	0.084	0.081	0.102	0.237	1.873	0.096
$\sqrt{MSE}$	0.056	0.047	0.041	0.418	0.626	2.474	1.561	2.447	5.168
Coverage	0.910	0.960	0.950	0.975	0.965	0.960	0.915	0.990	0.960

Tab. 6.19: Simulation results for scenarios 3 and 4 with weakly informative priors.

7 Conclusion

This thesis was aimed at presenting several distributions for compositional data, in order to overcome the main drawbacks of the Dirichlet distribution: rigid dependence structure, severe scheme of simplicial independences and unimodality. The first proposal to deal with these issues is the Additive Logistic-Normal proposed by Aitchison and based on a log-ratio transformation to compositional data. In such a way it is possible to map the D-dimensional simplex into $\mathbb{R}^{D-1}$. This approach allows to use all the standard statistical methods defined for multivariate Normal data. Despite this advantage, the ALN is an unimodal distribution and does not allow for positive covariances among components of the compositions.
A group of finite mixture models (each one including the Dirichlet as an inner point) have been then introduced: the Flexible Dirichlet, the Extended Flexible Dirichlet and the Double Flexible Dirichlet. These distributions are obtained closing different basis, obtained combining Gamma and Multinomial random variables. Both the FD and the EFD allow for at most D modes, whereas the DFD allows for at most $\frac{D(D+1)}{2}$ clusters. The EFD generalizes the FD, moving the cluster means along the lines joining each vertex of the simplex to a common barycenter (Section 3.4), whereas the DFD allows for an higher number of mixture components, increasing the number of clusters and varying their location on the simplex (Figures 4.3 and 4.4). If a subset of weights of the mixture assume value zero, then interesting cluster configurations are considered.

In this thesis, two works involving the FD and the EFD models, presented at two conferences, have been included: a new Bayesian estimation procedure for the parameters of a FD distribution and an intensive simulation study aimed at assessing the reliability of the EM algorithm implemented for the EFD. The core of this thesis regards the introduction of the DFD model. Several statistical properties have been derived for this new model. The DFD, as well as the EFD, allows for covariances that can assume also positive values. Since real compositions can present positive linear dependence (look at correlation matrices 5.1 and 5.9), this is a reasonable demand to a distribution defined on the simplex.

The EFD and the DFD resulted as the more interesting models for real data features, because of the flexibility they allow in the modelization of the covariance matrix and

clusters' position on the simplex. Simulation studies and applications to real data confirm the existence of cases where such cluster configurations make sense.

This family of "Flexible" distribution can be used also to generate new models. In Section 6 the Extended Flexible Dirichlet-Multinomial has been introduced compounding the Multinomial distribution with the EFD one. In this work, study of the EFDM distribution has begun, showing that it overcomes the rigidity of the Multinomial and Dirichlet-multinomial distributions. Since the EM algorithm does not provide satisfactory estimate of parameters, a better estimation procedure needs to be implemented. In future works, a Bayesian procedure will be defined and tested.

7.1 Future Works

7.1.1 Reducing the number of parameters in the DFD model

It has already been said that the number of parameters that the Double Flexible Dirichlet needs to estimate is very high. This fact penalizes the DFD fit to real data (as in Table 4.7). Indeed, as shown by Figure 7.1, the number of parameters grows quadratically with D, whereas most of the other models (expecpt for the ALN) is a linear function of D.

It is very easy to note that the number of parameters is strongly influenced by the matrix $\mathbf{P}$: even if it is assumed to be symmetric, it has $\frac{D(D+1)}{2}$ distinct elements to be estimated. In order to reduce this number of parameters, one can make some assumption on the matrix $\mathbf{P}$. The simpler assumption possible is to assume that the vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent. In this way one can obtain the following relationship:

$$p_{i,j} = P\left(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j\right) = P\left(\mathbf{Z}_1 = \mathbf{e}_i\right) \cdot P\left(\mathbf{Z}_2 = \mathbf{e}_j\right) = p_{i\cdot} \cdot p_{\cdot j}. \tag{7.1}$$

This assumption does not affect the number of clusters (components of the mixture), neither their position in the simplex. It alters only the "size" of each subpopulation. Moreover, with this assumption it is sufficient to estimate the vector $\mathbf{p}$ (recall from 4 that it is the probability vector of $\mathbf{Z}_1$ and $\mathbf{Z}_2$) in order to obtain an estimate of the entire matrix $\mathbf{P}$. It follows that the number of parameters of a DFD model coincides with the FD's one. This assumption brings three problems:

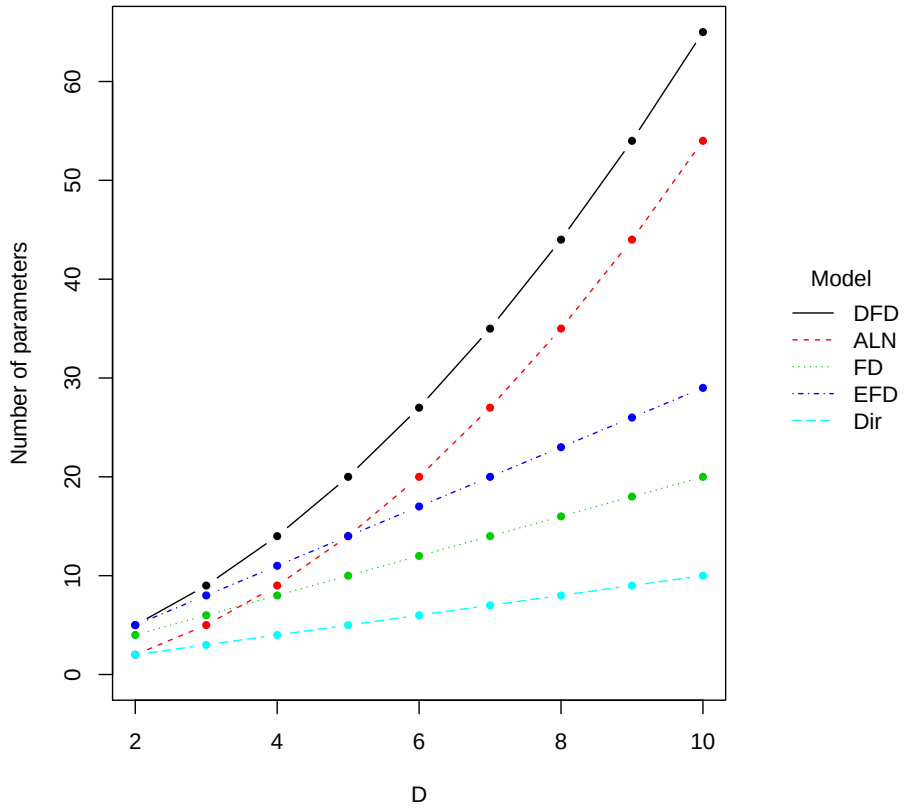

Fig. 7.1: Number of parameters of several models for compositional data.

- The EM algorithm needs to be modified. In general, the EM algorithm estimate the complete weights vector of a finite mixture model, so a different approach must to be implemented.

- The DFD model does not allow for positive covariances anymore. Indeed, from Equation 4.21 it is possible to see that $\text{Cov}\left(X_r, X_h\right) > 0 \Longrightarrow \left(p_{r,h} - 2p_{r\cdot} \cdot p_{\cdot h}\right) > 0$. But if (7.1) holds, then $p_{r,h} - 2p_{r\cdot} \cdot p_{\cdot h} = p_{r\cdot} \cdot p_{\cdot h} - 2p_{r\cdot} \cdot p_{\cdot h} = -p_{r\cdot} \cdot p_{\cdot h} < 0$.

- It is well-known that, if $\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2$, then $p_{i,j} = P\left(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j\right)$ can not take the value zero. This means that this assumption prevent the possibility of having empty clusters.

An alternative could be adding to this assumption some parameters that regulate the dependence among $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The number of these new parameters should not be too high, otherwise the simplification imposed to the parametric space would be pointless.

In conclusion, the parametrization of the DFD can be relaxed. This brings some drawbacks with respect to the properties showed in Section 4. Nonetheless, since the simplified version of the DFD model has the same number of parameters of the Flexible Dirichlet, it remain an interesting alternative to use when data show more than D clusters or when one looks for parametric forms of dependence.

7.1.2 Moving the clusters: the Extended Double Flexible Dirichlet

Looking at Figure 4.3, it is clear that the cluster structure imposed by the Double Flexible Dirichlet distribution is very rigid. For example, if all the $\frac{D(D+1)}{2}$ are present, the two following hold:

- Joining the cluster means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ (with respect to (4.30)), one obtains an equilateral triangle (as happens in the FD case).

- Clusters 4, 5 and 6 are forced to have mean vector at the midpoint of two out of the three vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$.

In order to relax this structure, one can combine pecularities of the DFD and EFD models. In particular, the basis (4.2) can be extended in the following way. Let $\mathbf{W} = (W_1, \dots, W_D)^\intercal$, $\mathbf{U}_1 = (U_{1,1}, \dots, U_{1,D})^\intercal$, $\mathbf{U}_2 = (U_{2,1}, \dots, U_{2,D})^\intercal$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)^\intercal$ be jointly independent and $W_r \sim \text{Gamma}(\alpha_r, 1)$, $U_{1,r}, U_{2,r} \sim \text{Gamma}(\tau_r, 1)$ $(r = 1, \dots, D)$ and $\mathbf{Z}_1, \mathbf{Z}_2 \sim \text{Multinomial}(1, \mathbf{p})$. Assuming also that:

- the W_r's are independent on each other,

- $\mathbf{U}_1$ and $\mathbf{U}_2$ have independent elements

then it is possible to construct the r-th element of the new basis $\mathbf{Y}$:

$$Y_r = W_r + U_{1,r} Z_{1,r} + U_{2,r} Z_{2,r}, \qquad r = 1, \dots, D. \tag{7.2}$$

Note that $\mathbf{Z}_1 \not\perp\!\!\!\perp \mathbf{Z}_2$, in general. Indeed, the only assumption made on the matrix $\mathbf{P}$ (i.e. the matrix whose generic element is $p_{i,j} = P(\mathbf{Z}_1 = \mathbf{e}_i, \mathbf{Z}_2 = \mathbf{e}_j)$) is that it must be symmetric. Closing this basis, a new distribution called Extended Double Flexible

Dirichlet (EDFD) is obtained. It is not surprising that the EDFD allows for a finite mixture structure:

$$EDFD(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{P}) = \sum_{i=1}^{D}\sum_{j=1}^{D} p_{i,j}\mathcal{D}\left(\mathbf{x};\boldsymbol{\alpha}+\tau_i\,\mathbf{e}_i+\tau_j\,\mathbf{e}_j\right). \tag{7.3}$$

Imposing the matrix $\mathbf{P}$ symmetric, as in the DFD case, this mixture has $\dfrac{D(D+1)}{2}$ components. The additional τ_r's allow to break both the constraint above described, as can be seen in Figure 7.2.

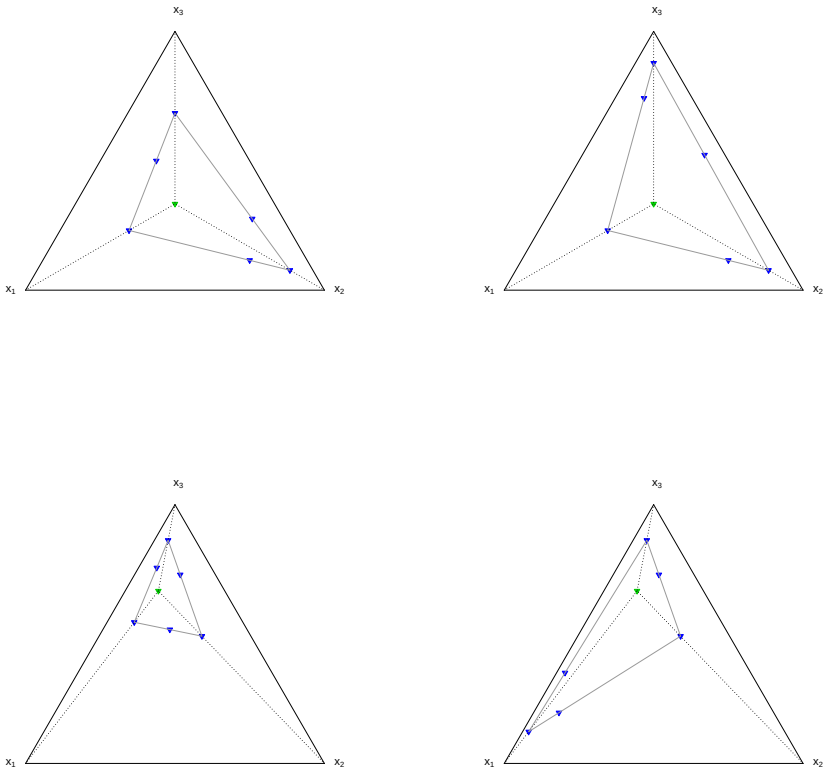

Fig. 7.2: EDFD cluster means structure. *Top-Left:* $\boldsymbol{\alpha} = (3,3,3)^\intercal$, $\boldsymbol{\tau} = (2,15,5)^\intercal$. *Top-Right:* $\boldsymbol{\alpha} = (3,3,3)^\intercal$, $\boldsymbol{\tau} = (2,15,20)^\intercal$. *Bottom-Left:* $\boldsymbol{\alpha} = (10,5,30)^\intercal$, $\boldsymbol{\tau} = (5,8,32)^\intercal$. *Bottom-Right:* $\boldsymbol{\alpha} = (10,5,30)^\intercal$, $\boldsymbol{\tau} = (100,8,32)^\intercal$.

The new cluster mean vectors can be expressed as weighted mean of three points: $\bar{\boldsymbol{\alpha}}$, $\mathbf{e}_i$ and $\mathbf{e}_j$:

$$\begin{aligned}\boldsymbol{\mu}_{i,j}^{EDFD} &= \frac{\boldsymbol{\alpha} + \tau_i\,\mathbf{e}_i + \tau_j\,\mathbf{e}_j}{\alpha^+ + \tau_i + \tau_j} \\ &= \left(\frac{\alpha^+}{\alpha^+ + \tau_i + \tau_j}\right)\bar{\boldsymbol{\alpha}} + \left(\frac{\tau_i}{\alpha^+ + \tau_i + \tau_j}\right)\mathbf{e}_i + \left(\frac{\tau_j}{\alpha^+ + \tau_i + \tau_j}\right)\mathbf{e}_j\end{aligned} \tag{7.4}$$

In this way the vector $\boldsymbol{\mu}_{i,j}$, $i \neq j$, is not forced to be at the midpoint of $\boldsymbol{\mu}_{i,i}$ and $\boldsymbol{\mu}_{j,j}$: it is still located on the segment joining these two vectors but its position on this line depends on τ_i and τ_j. In order to fit this model to real data, one must estimate D α_r's, D τ_r's and $\frac{D(D+1)}{2} - 1$'s $p_{i,j}$, that is a very high number. It is important to note that this number is equal to the number of parameters of a DFD model plus $(D-1)$ (that is the number of new τ_r's).

The major flexibility of the EDFD over the DFD can be seen even in the univariate case (Figure 7.3). It shows the density function of the DFD and the EDFD in different parametric configurations. Blue lines represent the position of the first element of cluster means. The matrix $\mathbf{P}$ is fixed for the four scenarios: $\mathbf{P} = \begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.5 \end{bmatrix}$.

In the DFD's panels, the distance between the first two cluster means' position ($\boldsymbol{\mu}_{1,1}$ and $\boldsymbol{\mu}_{1,2}$) is the same as the distance between the second and the third's ones ($\boldsymbol{\mu}_{1,2}$ and $\boldsymbol{\mu}_{2,2}$). This does not hold anymore for the EDFD model, where the first two green lines are closer than the second and the third ones.

EDFD as a starting point for regression models

Several proposals have been made in order to model proportions as function of a set of covariates [53]. The first attemp was in the Beta regression [35, 73], that is a good model for a large class of phenomena, excluding heavy tailed and multimodality ones. In order to overcome these limitations, the Beta Rectangular has been proposed [47]. This distribution is defined as a finite mixture an Uniform and a Beta components and therefore it fit data better than the Beta: thanks to the Uniform component, it allows for heavy tails. The Beta Rectanuglar can be used to define a regression model (BR regression model [13]). More recently, Migliorati et al [62] took advtantage of the univariate version of the Flexible Dirichlet (named "Flexible Beta", FB) and developed a regression model on it: the FB regression. The

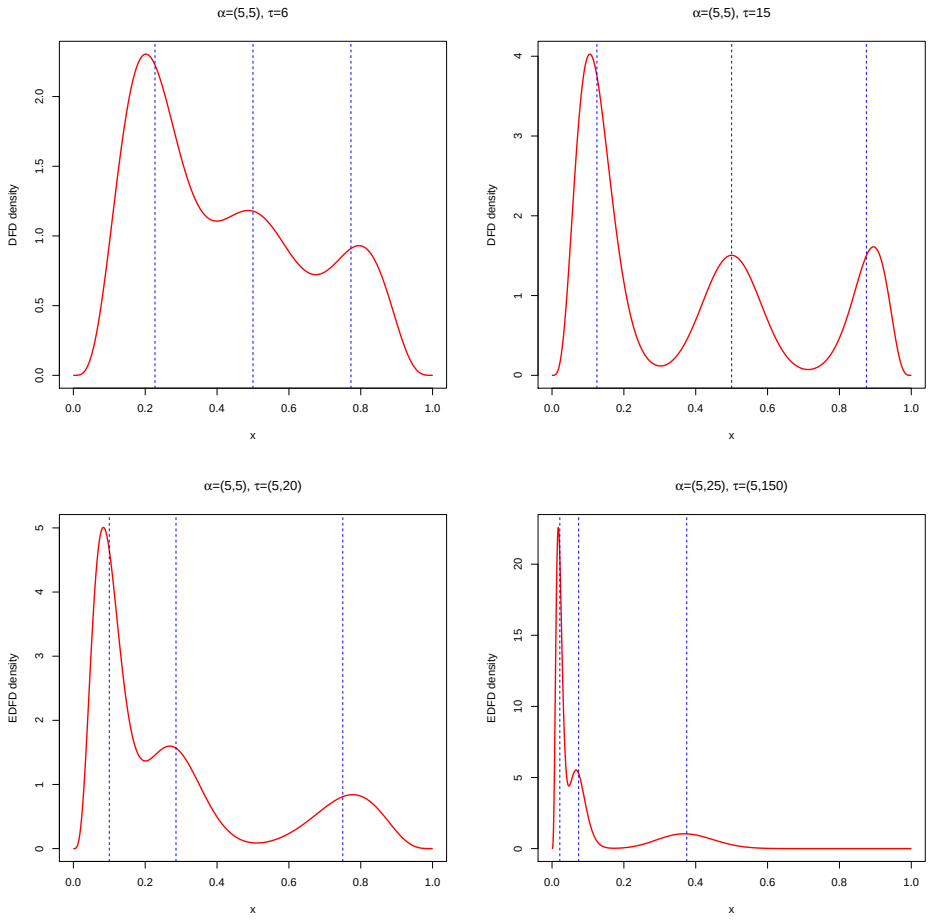

Fig. 7.3: DFD (*Top*) and EDFD *(Bottom)* univariate density functions. The matrix $\mathbf{P}$ has elements $p_{1,1} = 0.2$, $p_{1,2} = p_{2,1} = 0.15$ and $p_{2,2} = 0.5$. Blue lines indicate cluster means' position.

advantage of the FB with respect to the BR is that the first allows for bimodality forms, as well as asymmetric and heavy tails ones.

The univariate version of the EDFD could bring some advantage compared also to the FB: with just one more parameter ($p_{1,2}$), it allows for the presence of a third cluster and then the model (as well as the regression) have more flexibility.

7.1.3 Initialization methods for the EM algorithm in the Extended Flexible Dirichlet-Multinomial scenario

It has already been said that the choice of a good starting point is a crucial aspect for the EM algorithm. It holds also in the EFDM context. As in Section 3.4.3, some ad hoc initialization procedures should be developed. A very simple idea is to rely on the initialization obtained for the EFD distribution, with the same algorithm:

- Transform every 0 data in some small positive integers, according to the order of magnitude of data. This is necessary because the EFD (as most of simplex distribution) is not defined on the boundary of the simplex, that can be expressed as the set of compositions with at least one null component.

- Treat the transformed data as a basis and close it.

- Apply the initialization procedure described in 3.4.3 to this composition.

This is a very naif proposal that does not consider the count nature of the data, but it leads to good starting values for the EM algorithm. New and more reliable algorithms should be found.

8 Appendix

8.1 Bayesian estimation procedure

8.1.1 Other parameter configurations

The simulation study developed in Section 3.3.2 regards five parameters configurations, with different structure of the clusters. These configurations are the following:

	α_1	α_2	α_3	τ	p_1	p_2	p_3
ID 1	10	10	10	17	1/3	1/3	1/3
ID 2	7	26	15	17	1/3	1/3	1/3
ID 3	14	18	21	5.5	1/3	1/3	1/3
ID 4	10	10	10	12	0.2	0.45	0.35
ID 5	8	20	35	23	0.2	0.45	0.35

Tab. 8.1: Parameter configurations - old parameterization.

	μ_1	μ_2	μ_3	p_1	p_2	p_3	ϕ	w
ID 1	1/3	1/3	1/3	1/3	1/3	1/3	47	0.3617
ID 2	0.1949	0.4872	0.3179	1/3	1/3	1/3	65	0.4474
ID 3	0.271	0.339	0.390	1/3	1/3	1/3	58.5	0.1158
ID 4	0.295	0.367	0.338	0.2	0.45	0.35	42	0.3506
ID 5	0.1465	0.3529	0.5005	0.2	0.45	0.35	86	0.3651

Tab. 8.2: Parameter configurations - new parameterization.

ID 1: Well-separated clusters

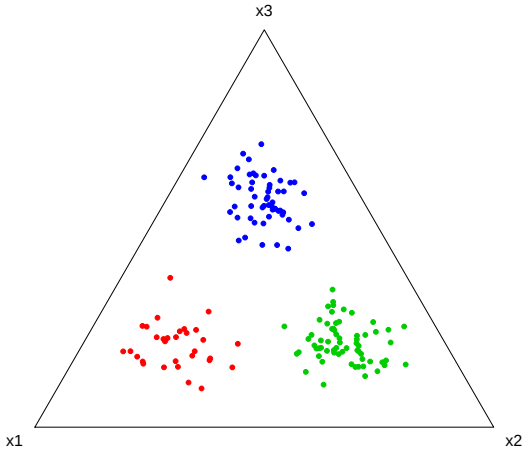

Fig. 8.1: Ternary Plot ID 1.

	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.333	0.3338	0.3336	0.3329	0.0146	0.3337	0.0145
μ_2	0.333	0.3352	0.3350	0.3349	0.0147	0.3353	0.0150
μ_3	0.333	0.3310	0.3307	0.3304	0.0145	0.3310	0.0146
p_1	0.333	0.3341	0.3334	0.3319	0.0384	0.3340	0.0388
p_2	0.333	0.3373	0.3365	0.3344	0.0385	0.3374	0.0388
p_3	0.333	0.3286	0.3278	0.3262	0.0381	0.3286	0.0392
ϕ	47	47.2354	47.1335	46.9404	3.8691	47.8244	3.8269
w	0.3617	0.3612	0.3613	0.3614	0.0088	0.3897	0.0175

Tab. 8.3: Simulation results - ID 1.

ID 2: Well-separated but closer clusters

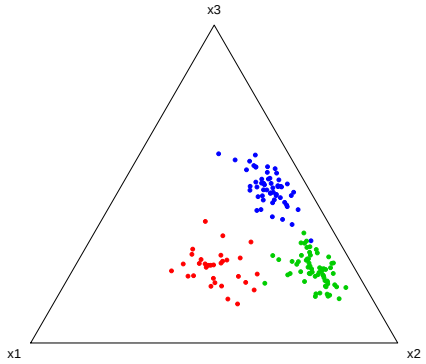

Fig. 8.2: Ternary Plot ID 2.

	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.1949	0.1953	0.1951	0.1947	0.01	0.1950	0.0109
μ_2	0.4872	0.4883	0.4882	0.4879	0.01	0.4885	0.0111
μ_3	0.3179	0.3164	0.3162	0.3157	0.01	0.3166	0.0108
p_1	0.333	0.3350	0.3342	0.3333	0.0374	0.3336	0.0410
p_2	0.333	0.3367	0.3360	0.3361	0.0387	0.3371	0.0395
p_3	0.333	0.3284	0.3276	0.3251	0.0387	0.3293	0.0381
ϕ	65	65.0841	64.9359	64.8278	5.4555	65.8584	5.1893
w	0.4474	0.2610	0.2610	0.2613	0.01	0.4456	0.0352

Tab. 8.4: Simulation results - ID 2.

ID 3: Overlapped clusters

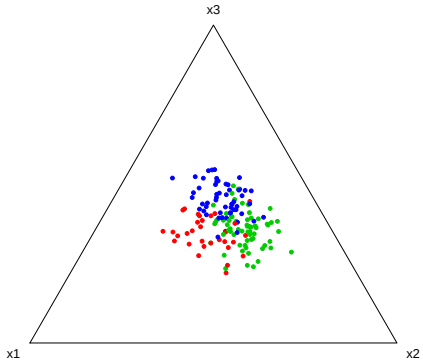

Fig. 8.3: Ternary Plot ID 3.

	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.271	0.2705	0.2705	0.2705	0.0058	0.2706	0.0058
μ_2	0.339	0.3402	0.3402	0.3403	0.0061	0.3402	0.0061
μ_3	0.390	0.3892	0.3892	0.3891	0.0063	0.3891	0.0062
p_1	0.333	0.3681	0.3413	0.3232	0.2054	0.3368	0.1549
p_2	0.333	0.3307	0.3013	0.2748	0.1990	0.3444	0.1624
p_3	0.333	0.3012	0.2692	0.2442	0.1975	0.3187	0.1712
ϕ	58.5	48.7262	47.5953	45.1559	8.7155	59.3322	9.9501
w	0.1158	0.0659	0.0692	0.0932	0.0305	0.1523	0.0496

Tab. 8.5: Simulation results - ID 3.

ID 4: Closed clusters and different weights

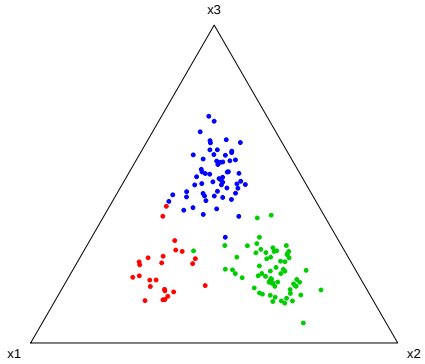

Fig. 8.4: Ternary Plot ID 4.

	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.295	0.2966	0.2963	0.2957	0.01	0.2961	0.0108
μ_2	0.367	0.3656	0.3656	0.3656	0.0141	0.3658	0.0123
μ_3	0.338	0.3377	0.3376	0.3379	0.0141	0.3381	0.0119
p_1	0.2	0.2035	0.2021	0.1987	0.0332	0.2019	0.0339
p_2	0.45	0.4467	0.4465	0.4477	0.0424	0.4470	0.0426
p_3	0.35	0.3498	0.3490	0.3464	0.04	0.3511	0.0392
ϕ	42	42.0453	41.9520	41.6533	3.6528	42.6291	3.0885
w	0.3506	0.2848	0.2849	0.2850	0.01	0.3508	0.0237

Tab. 8.6: Simulation results - ID 4.

ID 5: Well-separated clusters and different weights

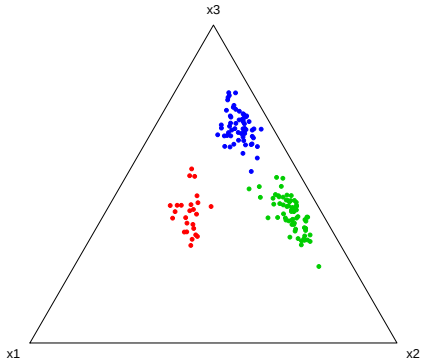

Fig. 8.5: Ternary Plot ID 5.

	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.1465	0.1483	0.1480	0.1473	0.01	0.1475	0.0098
μ_2	0.3529	0.3515	0.3514	0.3511	0.01	0.3520	0.0113
μ_3	0.5005	0.5002	0.5001	0.4997	0.01	0.5005	0.0106
p_1	0.2	0.2053	0.2040	0.2012	0.0332	0.2023	0.0340
p_2	0.45	0.4451	0.4448	0.4438	0.04	0.4471	0.0410
p_3	0.35	0.3496	0.3490	0.3473	0.0387	0.3506	0.0374
ϕ	86	86.4743	86.2843	89.5219	7.1449	87.6166	6.2640
w	0.3651	0.2671	0.2671	0.2670	0.01	0.3753	0.0301

Tab. 8.7: Simulation results - ID 5.

8.1.2 Robustness analysis on the prior for ϕ

In this section two cluster structure are considered: one with well-separated clusters and one with overlapped clusters. They correspond to ID 1 and 3 of Table 8.1.

First prior

$\phi \sim \text{Gamma}(g_2, g_2)$, with $g_2 = 0.0001$.

- Well-separated clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.333	0.334	0.333	0.333	0.015	0.334	0.015
μ_2	0.333	0.335	0.335	0.335	0.015	0.335	0.015
μ_3	0.333	0.331	0.331	0.330	0.015	0.331	0.015
p_1	0.333	0.334	0.333	0.332	0.038	0.334	0.039
p_2	0.333	0.337	0.337	0.336	0.039	0.337	0.039
p_3	0.333	0.329	0.328	0.326	0.038	0.329	0.039
ϕ	47	47.237	47.135	47.072	3.872	47.824	3.827
w	0.3617	0.361	0.361	0.361	0.009	0.390	0.018

- Overlapped clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.271	0.271	0.270	0.271	0.006	0.271	0.006
μ_2	0.339	0.340	0.340	0.340	0.006	0.340	0.006
μ_3	0.390	0.389	0.389	0.389	0.006	0.389	0.006
p_1	0.333	0.366	0.339	0.298	0.206	0.337	0.155
p_2	0.333	0.335	0.304	0.271	0.203	0.344	0.162
p_3	0.333	0.299	0.267	0.244	0.198	0.319	0.171
ϕ	58.5	48.684	47.526	44.997	8.720	59.332	9.950
w	0.1158	0.066	0.069	0.097	0.031	0.152	0.050

Second prior

$\phi \sim \text{Gamma}(k \cdot g_2, g_2)$, with $k = 40$ and $g_2 = 0.0001$.

- Well-separated clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.333	0.334	0.333	0.333	0.015	0.334	0.015
μ_2	0.333	0.335	0.335	0.334	0.015	0.335	0.015
μ_3	0.333	0.331	0.331	0.331	0.015	0.331	0.015
p_1	0.333	0.334	0.333	0.331	0.038	0.334	0.039
p_2	0.333	0.338	0.337	0.336	0.039	0.337	0.039
p_3	0.333	0.329	0.328	0.326	0.038	0.329	0.039
ϕ	47	47.239	47.135	46.974	3.872	47.824	3.827
w	0.3617	0.361	0.361	0.362	0.009	0.390	0.018

- Overlapped clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.271	0.271	0.270	0.270	0.006	0.271	0.006
μ_2	0.339	0.340	0.340	0.340	0.006	0.340	0.006
μ_3	0.390	0.389	0.389	0.389	0.006	0.389	0.006
p_1	0.333	0.367	0.341	0.305	0.208	0.337	0.155
p_2	0.333	0.332	0.301	0.274	0.203	0.344	0.162
p_3	0.333	0.300	0.268	0.244	0.199	0.319	0.171
ϕ	58.5	48.664	47.489	45.216	8.718	59.332	9.950
w	0.1158	0.066	0.069	0.095	0.031	0.152	0.050

Third prior

$\phi \sim$ Gamma(g_2, g_2), with $k = 200000$ and $g_2 = 0.0001$. This is a very extreme scenario: both the prior expectation and the prior variance are very large. This is a vague prior, with a (little) mass peak in $\phi = k$.

- Well-separated clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.333	0.334	0.334	0.333	0.015	0.334	0.015
μ_2	0.333	0.335	0.335	0.335	0.015	0.335	0.015
μ_3	0.333	0.331	0.331	0.331	0.015	0.331	0.015
p_1	0.333	0.334	0.333	0.331	0.038	0.334	0.039
p_2	0.333	0.337	0.337	0.334	0.039	0.337	0.039
p_3	0.333	0.329	0.328	0.328	0.038	0.329	0.039
ϕ	47	53.583	53.479	53.274	4.122	47.824	3.827
w	0.3617	0.363	0.363	0.363	0.008	0.390	0.018

- Overlapped clusters:

Parameter	True	Post. Mean	Post. Median	Post. Mode	Post. SD	MLE Exp.	MLE SE
μ_1	0.271	0.271	0.271	0.270	0.006	0.271	0.006
μ_2	0.339	0.340	0.340	0.340	0.006	0.340	0.006
μ_3	0.390	0.389	0.389	0.389	0.006	0.389	0.006
p_1	0.333	0.346	0.343	0.338	0.081	0.337	0.155
p_2	0.333	0.336	0.332	0.324	0.083	0.344	0.162
p_3	0.333	0.318	0.314	0.309	0.082	0.319	0.171
ϕ	58.5	77.339	77.265	80.071	8.870	59.332	9.950
w	0.1158	0.108	0.109	0.111	0.010	0.152	0.050

8.2 EFD: Conditional expectation

In this section will be shown the conditional expectation $\mathbb{E}\left[\mathbf{S}_1 | X_4 = x_4\right]$, varying the value of x_4. $\mathbf{S}_1$ is the 3-dimensional subcomposition originated by the first 3 elements of the 4-part compositions $\mathbf{X} = (X_1, X_2, X_3, X_4)^\intercal$.

$$\begin{cases} \boldsymbol{\alpha} = (2,2,2,2)^\intercal \\ \boldsymbol{\tau} = (4,4,4,4)^\intercal \\ \mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^\intercal \end{cases}$$

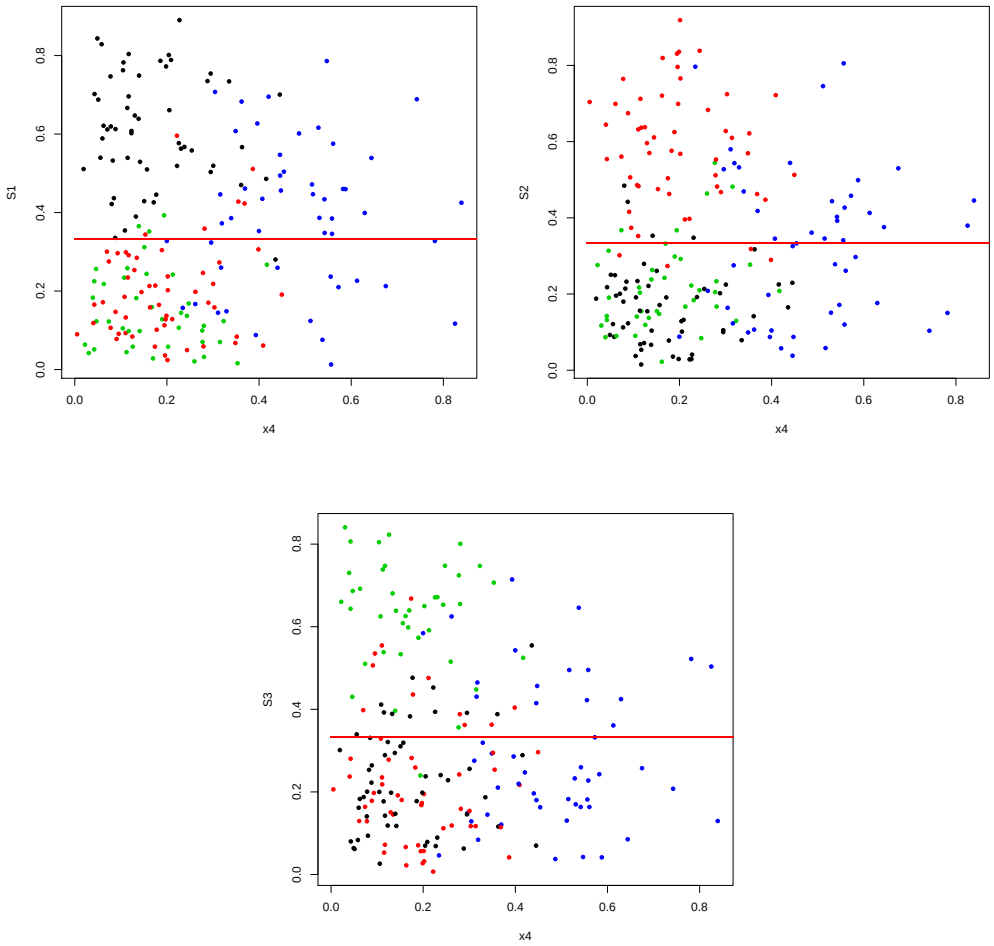

Fig. 8.6: EFD Conditional Expectation - First Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2, 10, 1, 6)^\intercal \\ \boldsymbol{\tau} = (10, 10, 10, 10)^\intercal \\ \mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^\intercal \end{cases}$$

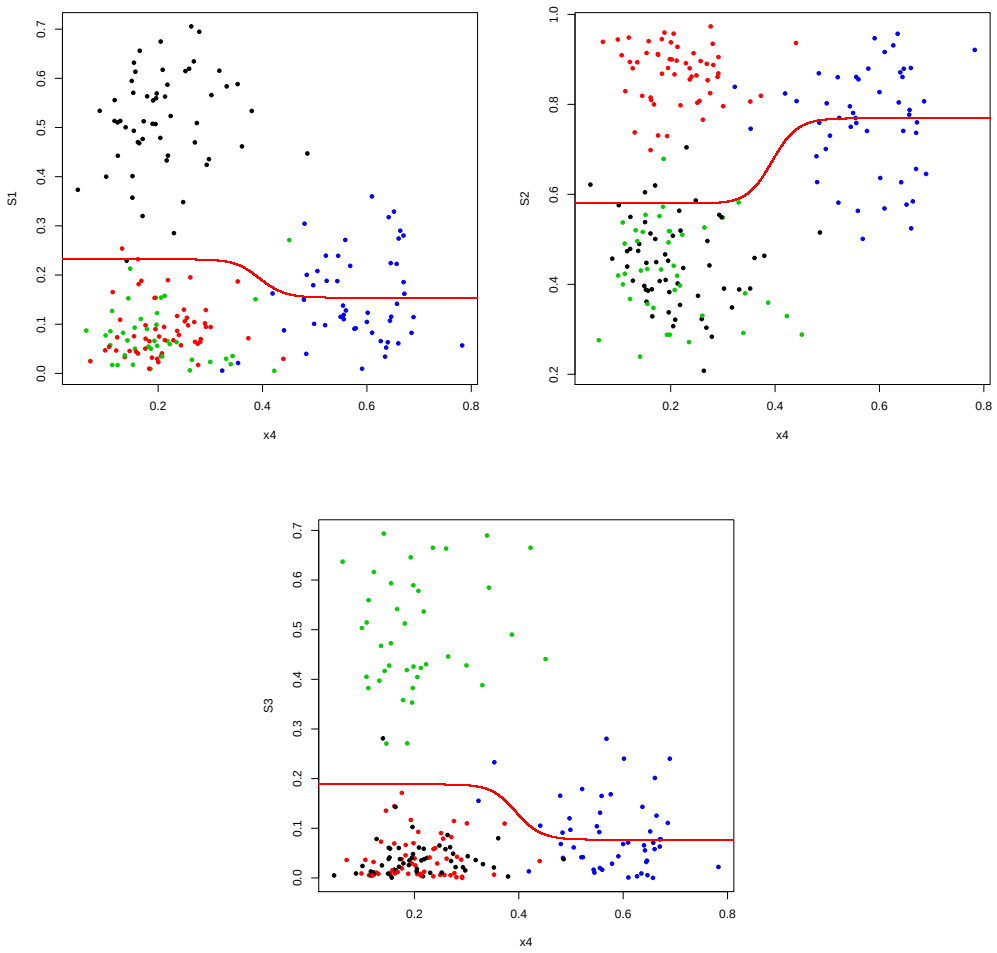

Fig. 8.7: EFD Conditional Expectation - Second Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2,2,2,2)^{\intercal} \\ \boldsymbol{\tau} = (1,4,70,3)^{\intercal} \\ \mathbf{p} = \left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)^{\intercal} \end{cases}$$

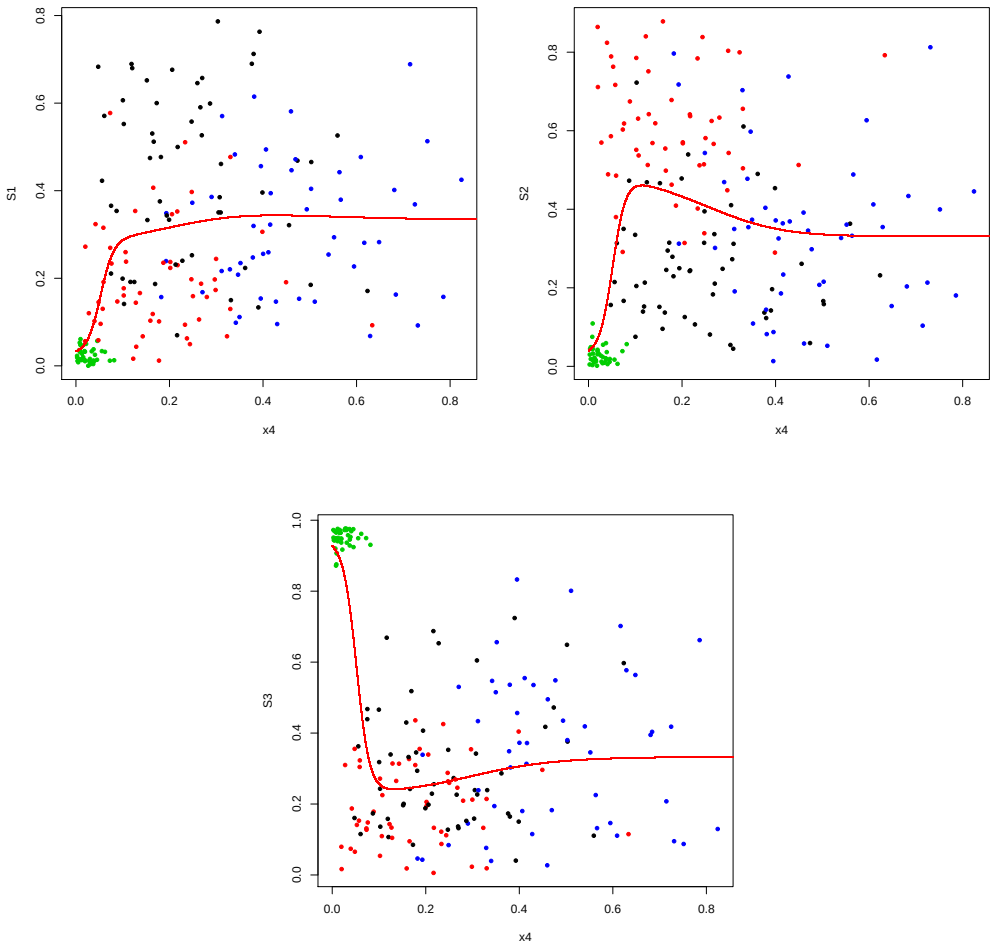

Fig. 8.8: EFD Conditional Expectation - Third Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2, 2, 2, 2)^\intercal \\ \boldsymbol{\tau} = (1, 4, 70, 3)^\intercal \\ \mathbf{p} = (0.1, 0.5, 0.05, 0.35)^\intercal \end{cases}$$

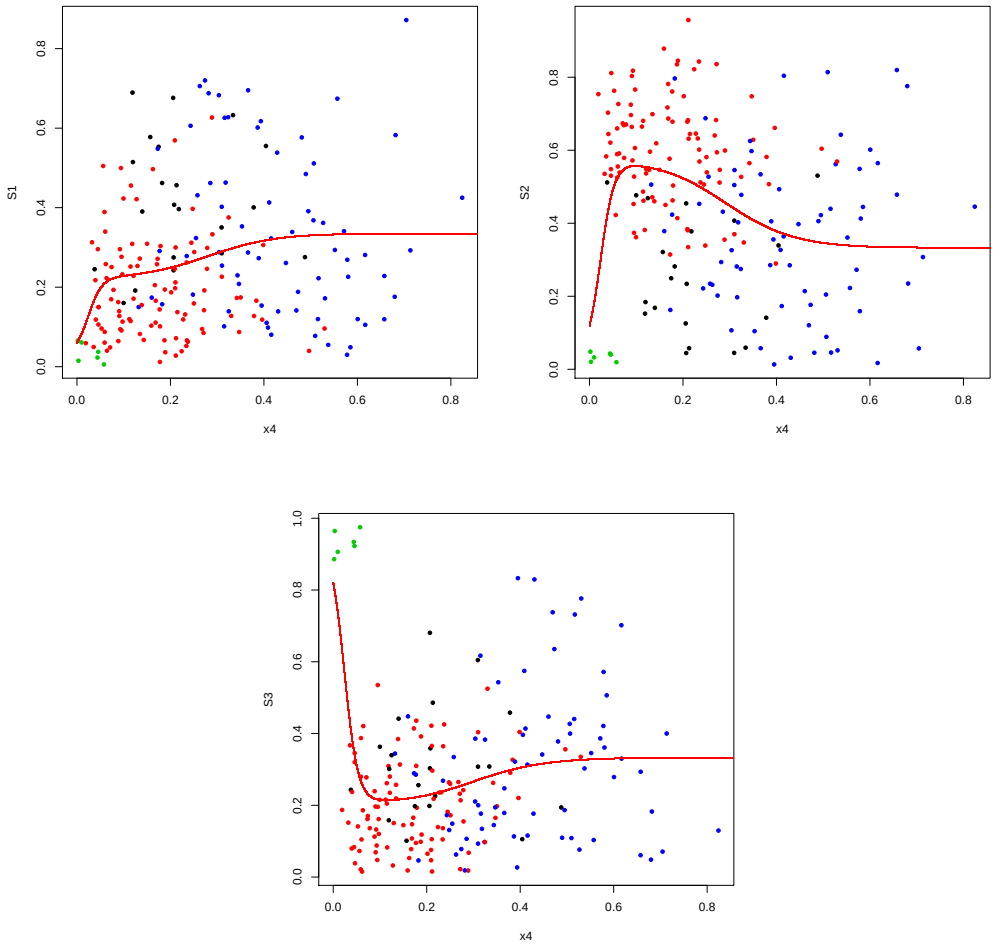

Fig. 8.9: EFD Conditional Expectation - Forth Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2,2,2,2)^{\intercal} \\ \boldsymbol{\tau} = (2,2,2,10)^{\intercal} \\ \mathbf{p} = \left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)^{\intercal} \end{cases}$$

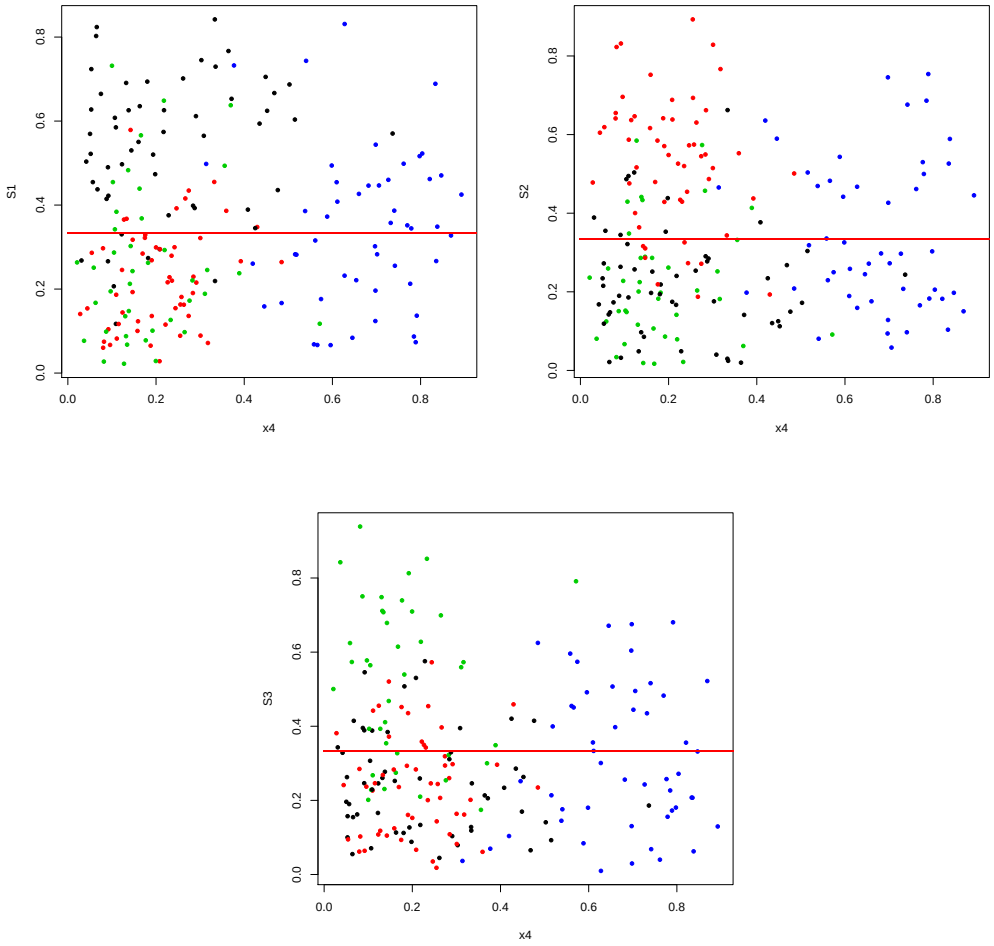

Fig. 8.10: EFD Conditional Expectation - Fifth Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2, 2, 2, 2)^\intercal \\ \boldsymbol{\tau} = (2, 2, 30, 2)^\intercal \\ \mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^\intercal \end{cases}$$

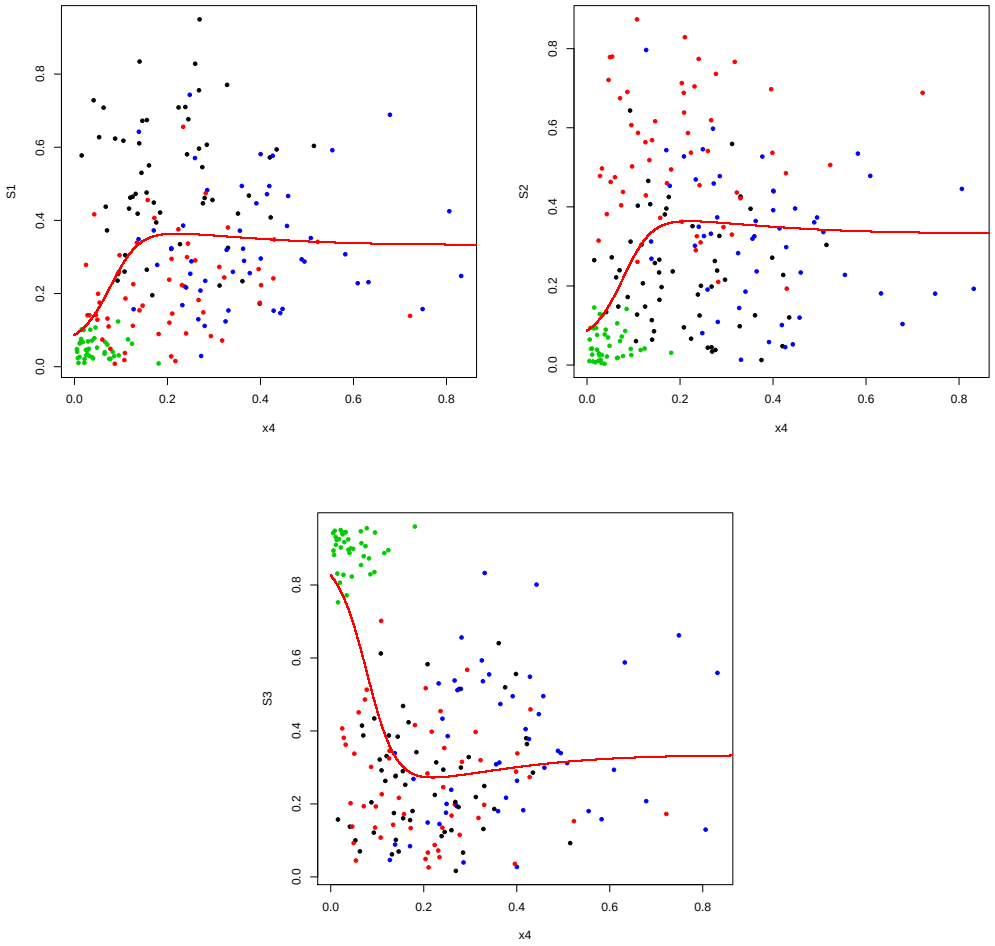

Fig. 8.11: EFD Conditional Expectation - Sixth Scenario.

$$\begin{cases} \boldsymbol{\alpha} = (2, 10, 1, 6)^\intercal \\ \boldsymbol{\tau} = (5, 5, 5, 3)^\intercal \\ \mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right)^\intercal \end{cases}$$

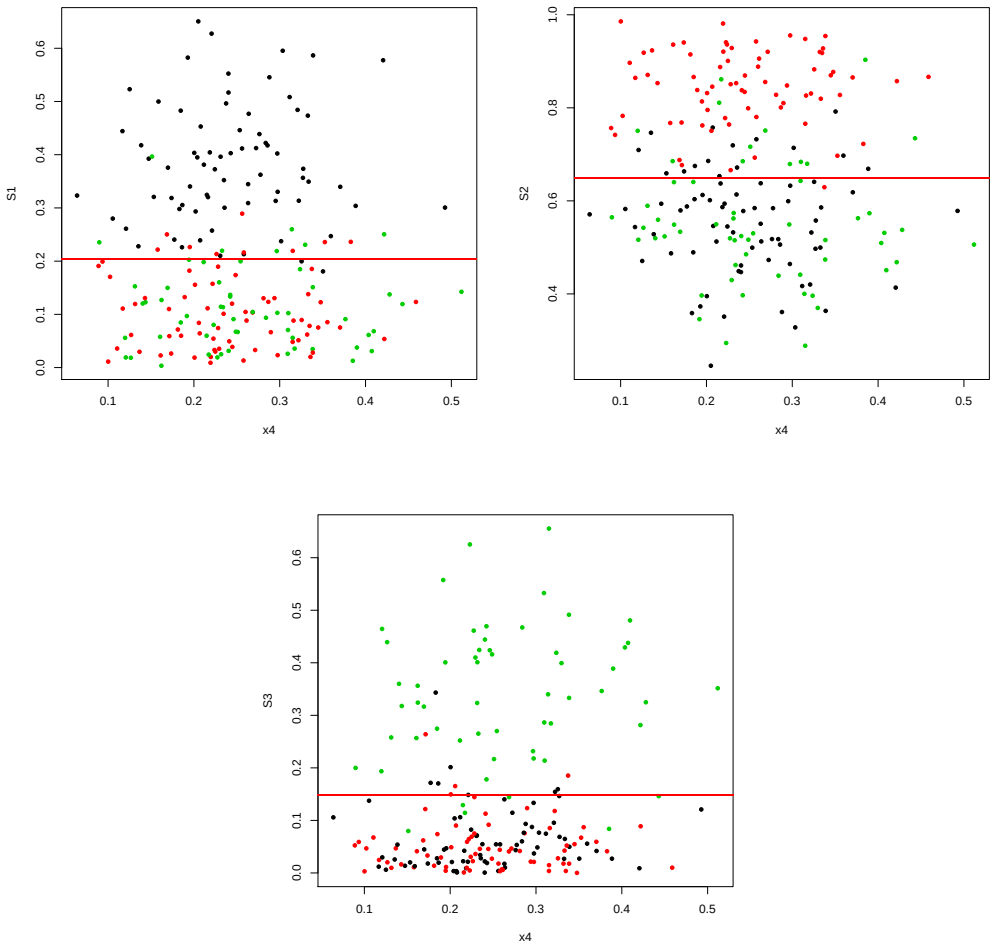

Fig. 8.12: EFD Conditional Expectation - Seventh Scenario.

8.3 EFD: MLE performance simulation

The following parameter configurations have been investigated:

ID	α_1	α_2	α_3	τ_1	τ_2	τ_3	p_1	p_2	p_3	$d_{SKL}(f_1, f_2)$	$d_{SKL}(f_1, f_3)$	$d_{SKL}(f_2, f_3)$
1	15	15	15	20	20	20	1/3	1/3	1/3	34.666	34.666	34.666
2	15	15	15	20	20	20	0.05	0.65	0.3	34.666	34.666	34.666
3	15	15	15	10	40	80	1/3	1/3	1/3	45.042	97.358	187.393
4	15	15	15	10	40	80	0.05	0.65	0.3	45.042	97.358	187.393
5	10	20	30	20	20	20	1/3	1/3	1/3	36.770	33.005	24.467
6	10	20	30	20	20	20	0.05	0.65	0.3	36.770	33.005	24.467
7	10	20	30	10	40	80	1/3	1/3	1/3	41.041	63.336	136.022
8	10	20	30	10	40	80	0.05	0.65	0.3	41.041	63.336	136.022
9	15	15	15	10	30	100	1/3	1/3	1/3	32.627	124.111	193.885
10	10	40	80	20	20	20	0.2	0.6	0.2	30.847	27.142	12.681
11	10	40	80	5	30	25	1/3	1/3	1/3	14.801	6.176	23.627
12	50	50	50	5	30	25	1/3	1/3	1/3	10.945	8.267	24.292
13	50	50	50	5	30	25	0.2	0.6	0.2	10.945	8.267	24.292
14	5	30	70	20	45	15	0.1	0.75	0.15	70.964	36.579	37.899
15	5	30	70	20	45	15	0.75	0.1	0.15	70.964	36.579	37.899
16	5	30	70	20	45	15	0.1	0.15	0.75	70.964	36.579	37.899
17	5	30	70	20	15	45	0.1	0.75	0.15	39.814	51.744	21.913
18	5	30	70	45	20	15	0.1	0.75	0.15	113.591	104.026	13.077
19	5	30	70	15	45	20	0.1	0.75	0.15	56.933	26.819	42.170
20	15	15	15	10	20	15	1/3	1/3	1/3	20.892	15.456	27.581
21	5	30	70	10	25	15	0.1	0.75	0.15	25.180	14.400	17.472

Tab. 8.8: Parameter configurations considered in the EFD's simulation studies.

- MLE Results ID = 1:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	15	15	15	20	20	20
MLE Mean	0.333	0.332	15.561	15.549	15.6	20.757	20.841	20.828
MLE sd	0.047	0.047	1.655	1.674	1.673	2.793	2.784	2.798
SE mean	0.047	0.047	1.618	1.617	1.623	2.744	2.753	2.749
arb	0.028	0.029	0.080	0.083	0.083	0.080	0.079	0.080
Coverage	0.951	0.952	0.946	0.943	0.942	0.944	0.947	0.951

Tab. 8.9: MLE Results for ID = 1.

- MLE Results ID = 2:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.05	0.65	15	15	15	20	20	20
MLE Mean	0.048	0.621	15.561	15.644	15.565	21.071	20.700	20.877
MLE sd	0.022	0.050	1.591	1.723	1.639	5.743	2.684	2.844
SE mean	0.021	0.048	1.581	1.734	1.622	5.214	2.600	2.846
arb	0.177	0.032	0.078	0.080	0.079	0.220	0.080	0.082
Coverage	0.862	0.902	0.953	0.956	0.951	0.941	0.948	0.947

Tab. 8.10: MLE Results for ID = 2.

- MLE Results ID = 3:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	15	15	15	10	40	80
MLE Mean	0.335	0.331	15.574	15.553	15.575	10.386	41.524	83.164
MLE sd	0.047	0.048	1.643	1.642	1.651	1.866	4.942	9.222
SE mean	0.047	0.047	1.608	1.605	1.608	1.749	4.808	9.115
arb	0.028	0.033	0.083	0.083	0.084	0.094	0.084	0.081
Coverage	0.948	0.951	0.945	0.949	0.948	0.943	0.954	0.953

Tab. 8.11: MLE Results for ID = 3.

- MLE Results ID = 4:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.05	0.65	15	15	15	10	40	80
MLE Mean	0.048	0.619	15.519	15.551	15.552	10.752	41.503	82.855
MLE sd	0.021	0.047	1.634	1.781	1.639	4.107	4.716	9.754
SE mean	0.021	0.048	1.565	1.708	1.609	3.722	4.597	9.242
arb	0.173	0.034	0.089	0.090	0.081	0.231	0.083	0.094
Coverage	0.870	0.909	0.949	0.937	0.949	0.945	0.958	0.936

Tab. 8.12: MLE Results for ID = 4.

- MLE Results ID = 5:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	10	20	30	20	20	20
MLE Mean	0.336	0.330	10.420	20.806	31.269	20.855	20.899	20.850
MLE sd	0.047	0.048	1.131	2.210	3.399	2.574	3.028	3.436
SE mean	0.047	0.047	1.095	2.168	3.255	2.525	2.889	3.336
arb	0.030	0.031	0.084	0.080	0.088	0.081	0.087	0.078
Coverage	0.953	0.946	0.954	0.947	0.941	0.952	0.944	0.947

Tab. 8.13: MLE Results for ID = 5.

- MLE Results ID = 6:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.05	0.65	10	20	30	20	20	20
MLE Mean	0.050	0.618	10.419	20.949	31.314	20.919	20.753	21.085
MLE sd	0.022	0.049	1.057	2.351	3.208	4.614	2.850	4.029
SE mean	0.022	0.049	1.071	2.343	3.293	4.271	2.899	3.765
arb	0.173	0.022	0.078	0.081	0.080	0.197	0.073	0.098
Coverage	0.882	0.907	0.950	0.951	0.961	0.957	0.965	0.931

Tab. 8.14: MLE Results for ID = 6.

- MLE Results ID = 7:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	10	20	30	10	40	80
MLE Mean	0.334	0.336	10.395	20.762	31.142	10.393	41.530	83.101
MLE sd	0.047	0.046	1.130	2.208	3.313	1.549	5.022	9.838
SE mean	0.047	0.047	1.076	2.133	3.194	1.488	4.837	9.431
arb	0.029	0.035	0.091	0.086	0.086	0.088	0.087	0.089
Coverage	0.955	0.966	0.940	0.941	0.937	0.937	0.946	0.950

Tab. 8.15: MLE Results for ID = 7.

- MLE Results ID = 8:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.05	0.65	10	20	30	10	40	80
MLE Mean	0.048	0.619	10.300	20.629	30.966	10.664	41.303	82.738
MLE sd	0.021	0.048	1.062	2.322	3.306	3.281	4.943	10.389
SE mean	0.021	0.048	1.039	2.268	3.198	2.951	4.780	9.927
arb	0.175	0.022	0.082	0.086	0.085	0.228	0.084	0.095
Coverage	0.867	0.915	0.950	0.944	0.949	0.941	0.951	0.951

Tab. 8.16: MLE Results for ID = 8.

- MLE Results ID = 9:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	15	15	15	10	30	100
MLE Mean	0.333	0.336	15.542	15.554	15.572	10.384	31.254	103.673
MLE sd	0.046	0.046	1.649	1.592	1.650	1.743	3.992	11.289
SE mean	0.047	0.047	1.609	1.610	1.613	1.761	3.776	11.292
arb	0.036	0.038	0.083	0.077	0.081	0.075	0.092	0.080
Coverage	0.949	0.961	0.943	0.956	0.941	0.961	0.948	0.953

Tab. 8.17: MLE Results for ID = 9.

- MLE Results ID = 10:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.2	0.6	10	40	80	20	20	20
MLE Mean	0.200	0.598	10.404	41.719	83.406	20.929	20.885	20.843
MLE sd	0.041	0.053	1.120	4.624	9.091	2.736	3.416	6.445
SE mean	0.040	0.053	1.115	4.565	8.926	2.701	3.402	6.503
arb	0.068	0.051	0.079	0.081	0.090	0.088	0.127	0.107
Coverage	0.933	0.944	0.954	0.945	0.956	0.955	0.960	0.948

Tab. 8.18: MLE Results for ID = 10.

- MLE Results ID = 11:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	10	40	80	5	30	25
MLE Mean	0.336	0.331	10.491	41.982	83.890	5.388	31.506	26.901
MLE sd	0.073	0.051	1.259	5.166	10.007	1.461	4.545	6.742
SE mean	0.07	0.05	1.228	5.006	9.834	1.376	4.500	6.491
arb	0.168	0.048	0.083	0.080	0.078	0.135	0.075	0.105
Coverage	0.919	0.936	0.939	0.942	0.950	0.935	0.947	0.931

Tab. 8.19: MLE Results for ID = 11.

- MLE Results ID = 12:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	50	50	50	5	30	25
MLE Mean	0.330	0.335	52.293	52.308	52.224	5.419	31.377	26.394
MLE sd	0.064	0.053	6.134	6.254	6.141	2.802	4.769	4.459
SE mean	0.065	0.053	6.050	6.117	6.081	2.896	4.696	4.377
arb	0.116	0.060	0.077	0.078	0.076	0.105	0.075	0.081
Coverage	0.932	0.940	0.940	0.942	0.942	0.964	0.951	0.937

Tab. 8.20: MLE Results for ID = 12.

- MLE Results ID = 13:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.2	0.6	50	50	50	5	30	25
MLE Mean	0.201	0.595	52.496	52.643	52.555	5.808	31.499	26.607
MLE sd	0.058	0.057	6.091	6.458	6.211	3.814	4.539	5.080
SE mean	0.056	0.054	5.900	6.155	5.907	3.750	4.437	4.966
arb	0.166	0.076	0.085	0.088	0.088	0.198	0.087	0.112
Coverage	0.922	0.933	0.946	0.943	0.937	0.962	0.941	0.930

Tab. 8.21: MLE Results for ID = 13.

- MLE Results ID = 14:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	20	45	15
MLE Mean	0.099	0.752	5.186	31.225	72.773	20.843	46.683	15.981
MLE sd	0.029	0.044	0.554	3.598	7.662	2.911	5.596	6.723
SE mean	0.030	0.043	0.529	3.524	7.419	2.859	5.427	6.842
arb	0.109	0.049	0.089	0.086	0.085	0.100	0.084	0.099
Coverage	0.932	0.958	0.944	0.951	0.948	0.955	0.951	0.955

Tab. 8.22: MLE Results for ID = 14.

- MLE Results ID = 15:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.75	0.1	5	30	70	20	45	15
MLE Mean	0.752	0.100	5.205	31.247	72.925	20.863	47.031	15.571
MLE sd	0.043	0.030	0.671	3.113	7.290	2.210	6.224	5.511
SE mean	0.043	0.030	0.664	3.161	7.388	2.256	6.385	5.364
arb	0.092	0.079	0.078	0.047	0.108	0.079	0.098	0.106
Coverage	0.950	0.935	0.957	0.955	0.949	0.951	0.967	0.950

Tab. 8.23: MLE Results for ID = 15.

- MLE Results ID = 16:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.15	5	30	70	20	45	15
MLE Mean	0.101	0.149	5.174	31.084	72.827	20.889	46.997	15.217
MLE sd	0.029	0.035	0.531	3.203	8.159	3.140	6.815	5.270
SE mean	0.030	0.035	0.530	3.188	8.224	2.977	6.987	5.168
arb	0.114	0.082	0.080	0.079	0.084	0.113	0.098	0.084
Coverage	0.941	0.940	0.948	0.944	0.950	0.943	0.960	0.953

Tab. 8.24: MLE Results for ID = 16.

- MLE Results ID = 17:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	20	15	45
MLE Mean	0.100	0.749	5.225	31.440	73.180	20.996	15.628	47.191
MLE sd	0.030	0.044	0.540	3.654	7.709	3.178	2.981	10.636
SE mean	0.030	0.044	0.543	3.611	7.622	2.909	2.821	10.171
arb	0.108	0.050	0.080	0.087	0.079	0.127	0.092	0.111
Coverage	0.935	0.933	0.952	0.952	0.951	0.949	0.941	0.948

Tab. 8.25: MLE Results for ID = 17.

- MLE Results ID = 18:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	45	20	15
MLE Mean	0.099	0.747	5.205	31.374	72.992	47.142	20.813	16.354
MLE sd	0.030	0.048	0.585	3.932	8.185	6.185	3.480	8.429
SE mean	0.029	0.048	0.554	3.698	7.840	5.848	3.363	7.886
arb	0.107	0.079	0.092	0.102	0.090	0.106	0.096	0.142
Coverage	0.928	0.944	0.942	0.943	0.939	0.944	0.938	0.936

Tab. 8.26: MLE Results for ID = 18.

- MLE Results ID = 19:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	15	45	20
MLE Mean	0.100	0.751	5.156	31.097	72.474	15.601	46.520	20.921
MLE sd	0.031	0.042	0.525	3.462	7.403	2.433	5.492	7.087
SE mean	0.030	0.043	0.528	3.530	7.419	2.354	5.421	7.318
arb	0.111	0.053	0.080	0.087	0.080	0.117	0.082	0.104
Coverage	0.922	0.955	0.953	0.961	0.958	0.955	0.951	0.967

Tab. 8.27: MLE Results for ID = 19.

- MLE Results ID = 20:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.333	0.333	15	15	15	10	20	15
MLE Mean	0.333	0.333	15.626	15.624	15.626	10.408	20.935	15.700
MLE sd	0.050	0.047	1.702	1.695	1.718	1.991	2.914	2.454
SE mean	0.050	0.048	1.691	1.700	1.694	1.941	2.876	2.411
arb	0.044	0.038	0.078	0.075	0.079	0.078	0.077	0.076
Coverage	0.950	0.944	0.958	0.949	0.943	0.952	0.950	0.95

Tab. 8.28: MLE Results for ID = 20.

- MLE Results ID = 21:

	p_1	p_2	α_1	α_2	α_3	τ_1	τ_2	τ_3
True	0.1	0.75	5	30	70	10	25	15
MLE Mean	0.100	0.750	5.212	31.598	73.177	10.531	25.942	16.764
MLE sd	0.031	0.045	0.582	4.021	8.177	2.193	3.820	8.670
SE mean	0.032	0.045	0.555	3.827	7.878	2.048	3.761	8.232
arb	0.132	0.059	0.090	0.098	0.086	0.164	0.086	0.147
Coverage	0.936	0.936	0.943	0.940	0.945	0.928	0.957	0.941

Tab. 8.29: MLE Results for ID = 21.

8.4 Proof of Theorem 1

Theorem (Identifiability of the DFD model)**.** *Let* $\boldsymbol{X} \sim DFD(\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \tau, \boldsymbol{P})^\intercal$ *and* $\boldsymbol{X}' \sim DFD(\boldsymbol{\theta}')$, $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \tau', \boldsymbol{P}')^\intercal$. *Then, if* $\boldsymbol{P}$ *and* $\boldsymbol{P}'$ *are not diagonal matricies,* $\boldsymbol{X} \sim \boldsymbol{X}'$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

Proof. It is obvious that if $\boldsymbol{\theta} = \boldsymbol{\theta}'$ then $\mathbf{X} \sim \mathbf{X}'$. Focusing on the converse, if $\mathbf{X} \sim \mathbf{X}'$, then the marginals have the same distribution: $X_k \sim X'_k$, $k = 1, \ldots, D$. Thanks to the closure under amalgamation property, the probability density function of X_k can be obtained as:

$$g_k(x; \boldsymbol{\theta}) = x^{\alpha_k - 1}(1-x)^{\alpha^+ - \alpha_k - 1} \cdot \Bigg\{ p_{k,k}\, a_k(\boldsymbol{\alpha}, \tau) x^{2\tau} + \\ + \left(\sum_{i \neq k} \sum_{j \neq k} p_{i,j} \right) b_k(\boldsymbol{\alpha}, \tau)(1-x)^{2\tau} + 2 \left(\sum_{i \neq k} p_{i,k} \right) c_k(\boldsymbol{\alpha}, \tau) \cdot [x^\tau (1-x)^\tau] \Bigg\}, \tag{8.1}$$

where $x \in (0, 1)$ and:

- $a_k(\boldsymbol{\alpha}, \tau) = \dfrac{\Gamma(\alpha^+ + 2\tau)}{\Gamma(\alpha_k + 2\tau)\Gamma(\alpha^+ - \alpha_k)}$
- $b_k(\boldsymbol{\alpha}, \tau) = \dfrac{\Gamma(\alpha^+ + 2\tau)}{\Gamma(\alpha_k)\Gamma(\alpha^+ - \alpha_k + 2\tau)}$
- $c_k(\boldsymbol{\alpha}, \tau) = \dfrac{\Gamma(\alpha^+ + 2\tau)}{\Gamma(\alpha_k + \tau)\Gamma(\alpha^+ - \alpha_k + \tau)}$

If $X_k \sim X'_k$, then $g_k(x;\boldsymbol{\theta}) = g_k(x;\boldsymbol{\theta'})$ a.s., $k = 1, \ldots, D$. As $g_k(x;\boldsymbol{\theta})$ and $g_k(x;\boldsymbol{\theta'})$ are two continuous density functions on the interval $(0,1)$, the previous equality must hold identically for any $x \in (0,1)$. In particular, $\lim\limits_{x\to 0^+} \dfrac{g_x(x;\boldsymbol{\theta})}{x^{\alpha_k - 1}} = \lim\limits_{x\to 0^+} \dfrac{g_x(x;\boldsymbol{\theta'})}{x^{\alpha_k - 1}}$.

$$\lim_{x\to 0^+} \frac{g_x(x;\boldsymbol{\theta})}{x^{\alpha_k - 1}} = \left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau)$$

$$\lim_{x\to 0^+} \frac{g_x(x;\boldsymbol{\theta'})}{x^{\alpha_k - 1}} = \left(\lim_{x\to 0^+} \frac{x^{\alpha'_k - 1}}{x^{\alpha_k - 1}}\right) \left(\sum_{i\neq k}\sum_{j\neq k} p'_{i,j}\right) b_k(\boldsymbol{\alpha'}, \tau')$$

In order to satisfy the equality, the quantity $\left(\lim\limits_{x\to 0^+} \dfrac{x^{\alpha'_k - 1}}{x^{\alpha_k - 1}}\right)$ must be finite and positive.

$$\left(\lim_{x\to 0^+} \frac{x^{\alpha'_k - 1}}{x^{\alpha_k - 1}}\right) = \begin{cases} 0, \text{ if } \alpha'_k > \alpha_k \\ 1, \text{ if } \alpha'_k = \alpha_k \\ +\infty, \text{ if } \alpha'_k < \alpha_k \end{cases}$$

Then,

$$\begin{cases} \alpha_k = \alpha'_k \ \ \forall\ k \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha'} \\ \left(\sum\limits_{i\neq k}\sum\limits_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau) = \left(\sum\limits_{i\neq k}\sum\limits_{j\neq k} p'_{i,j}\right) b_k(\boldsymbol{\alpha'}, \tau') \end{cases} \tag{8.2}$$

The equality $\dfrac{g_k(x;\boldsymbol{\theta})}{g_k(x;\boldsymbol{\theta'})} = 1$ can be written for any $x \in (0,1)$ with $\boldsymbol{\alpha} = \boldsymbol{\alpha'}$ and (8.2):

$$\frac{p_{k,k}\ a_k(\boldsymbol{\alpha}, \tau)x^{2\tau} + \left(\sum\limits_{i\neq k}\sum\limits_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau)(1-x)^{2\tau} + 2\left(\sum\limits_{i\neq k} p_{i,k}\right) c_k(\boldsymbol{\alpha}, \tau) \cdot [x^{\tau}(1-x)^{\tau}]}{p'_{k,k}\ a_k(\boldsymbol{\alpha}, \tau')x^{2\tau'} + \left(\sum\limits_{i\neq k}\sum\limits_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau)(1-x)^{2\tau'} + 2\left(\sum\limits_{i\neq k} p'_{i,k}\right) c_k(\boldsymbol{\alpha}, \tau') \cdot \left[x^{\tau'}(1-x)^{\tau'}\right]} = 1$$

The $\lim\limits_{x\to 1^-}$ of the above fraction must be equal to 1 for the a.s. equality. It follows that:

$$p_{k,k}\ a_k(\boldsymbol{\alpha}, \tau) = p'_{k,k}\ a_k(\boldsymbol{\alpha}, \tau') \tag{8.3}$$

The equality $\dfrac{g_k(x;\boldsymbol{\theta})}{g_k(x;\boldsymbol{\theta'})} = 1$ for any $x \in (0,1)$ now appears like:

$$\frac{p_{k,k}\ a_k(\boldsymbol{\alpha},\tau)x^{2\tau} + \left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha},\tau)(1-x)^{2\tau} + 2\left(\sum_{i\neq k} p_{i,k}\right) c_k(\boldsymbol{\alpha},\tau)\left[x^{\tau}(1-x)^{\tau}\right]}{p_{k,k}\ a_k(\boldsymbol{\alpha},\tau)x^{2\tau'} + \left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha},\tau)(1-x)^{2\tau'} + 2\left(\sum_{i\neq k} p'_{i,k}\right) c_k(\boldsymbol{\alpha},\tau')\left[x^{\tau'}(1-x)^{\tau'}\right]} = 1$$

Deriving both the numerator and the denominator with respect to x and dividing them:

$$\frac{2\tau p_{k,k}\ a_k(\boldsymbol{\alpha},\tau)x^{2\tau-1} - 2\tau\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha},\tau)(1-x)^{2\tau-1} + 2\left(\sum_{i\neq k} p_{i,k}\right) c_k(\boldsymbol{\alpha},\tau)\tau x^{\tau}(1-x)^{\tau}\left[x^{-1}-(1-x)^{-1}\right]}{2\tau' p_{k,k}\ a_k(\boldsymbol{\alpha},\tau)x^{2\tau'-1} - 2\tau'\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha},\tau)(1-x)^{2\tau'-1} + 2\left(\sum_{i\neq k} p'_{i,k}\right) c_k(\boldsymbol{\alpha},\tau')\tau' x^{\tau'}(1-x)^{\tau'}\left[x^{-1}-(1-x)^{-1}\right]} = 1 \tag{8.4}$$

Since the above equality must hold for any $x \in (0,1)$, Equation (8.4) can be evaluated in $x = 0.5$:

$$\frac{2\tau 0.5^{2\tau-1}}{2\tau' 0.5^{2\tau'-1}} = 1 \qquad \Longrightarrow \qquad \frac{\tau}{4^{\tau}} = \frac{\tau'}{4^{\tau'}}. \tag{8.5}$$

This means that if $\mathbf{X} \sim \mathbf{X}'$ with $\tau \neq \tau'$, then τ and τ' must satify the equality (8.5). Studying the function $f(\tau) = \dfrac{\tau}{4^{\tau}}$, with $\tau > 0$, one can conclude that:

- $f(\tau) = \dfrac{\tau}{4^{\tau}} > 0 \Longleftrightarrow \tau > 0$
- $\lim\limits_{\tau\to 0^+} \dfrac{\tau}{4^{\tau}} = 0; \qquad \lim\limits_{\tau\to+\infty} \dfrac{\tau}{4^{\tau}} = 0$
- $f'(\tau) = \dfrac{\partial}{\partial\tau}\dfrac{\tau}{4^{\tau}} = \dfrac{1-\tau\cdot\ln 4}{4^{\tau}} \geq 0 \Longleftrightarrow \tau \leq \dfrac{1}{\ln 4}$
- $f''(\tau) = \dfrac{\partial}{\partial\tau} f'(\tau) = \dfrac{\ln 4\,(\tau\ln 4 - 2)}{4^{\tau}}$
- $f''\left(\dfrac{1}{\ln 4}\right) = -\dfrac{\ln 4}{4^{\frac{1}{\ln 4}}} < 0$

Then $\tau = \dfrac{1}{\ln 4}$ is the maximum of $f(\tau)$; in Figure (8.13) it is possible to see its plot. It is easy to see that the same value of $\frac{\tau}{4^{\tau}}$ can be reached at most with two different values of τ.

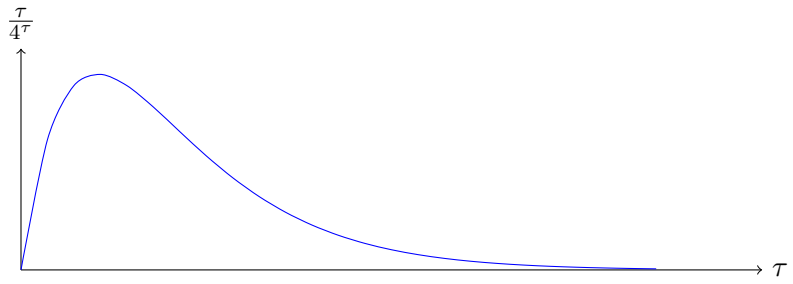

Fig. 8.13: Plot of $\frac{\tau}{4^\tau}$.

Thus, two different scenarios arise: one with $\tau = \tau'$ and one with $\tau \neq \tau'$. If $\tau = \tau'$, one can show, from equations (8.2) and (8.3), that the equality of the distributions is possibile only if $\mathbf{P} = \mathbf{P}'$ and this means that $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

The case $\tau \neq \tau'$ needs to be studied. Looking at equation (8.4), $\lim_{x\to 0^+}$ of both sides can be computed, considering if $\tau \notin \left\{1, \frac{1}{2}\right\}$ and $\tau' \notin \left\{1, \frac{1}{2}\right\}$:

$$\lim_{x\to 0^+}(8.4) = \frac{-2\tau\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau)}{-2\tau'\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, \tau)} = 1 \quad \Longrightarrow \quad \boxed{\tau = \tau'}$$

It is interesting to note that $\tau = 1$ and $\tau' = \frac{1}{2}$ are "connected" with respect to constraint (8.5): $\frac{\tau}{4^\tau} - \frac{\tau'}{4^{\tau'}} = \frac{1}{4} - \frac{1}{2\sqrt{4}} = 0$. This means that studying only this scenario is enough: the case where only one between τ and τ' belongs to $\left\{1, \frac{1}{2}\right\}$ can be avoided.

Evaluating (8.4) with $\tau = 1$ and $\tau' = \frac{1}{2}$ brings to the equality:

$$\frac{2p_{k,k}\, a_k(\boldsymbol{\alpha}, 1)x^1 - 2\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, 1)(1-x)^1 + 2\left(\sum_{i\neq k} p_{i,k}\right) c_k(\boldsymbol{\alpha}, 1)\,[1-2x]}{p_{k,k}\, a_k(\boldsymbol{\alpha}, 1) - \left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, 1) + 2\left(\sum_{i\neq k} p'_{i,k}\right) c_k(\boldsymbol{\alpha}, 0.5)\left[\frac{1}{2}\left(\frac{1-x}{x}\right)^{0.5} - \frac{1}{2}\left(\frac{x}{1-x}\right)^{0.5}\right]} = 1 \tag{8.6}$$

$$\lim_{x\to 0^+} \text{Num} = -2\left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, 1) + 2\left(\sum_{i\neq k} p_{i,k}\right) c_k(\boldsymbol{\alpha}, 1) \tag{8.7}$$

$$\lim_{x\to 0^+} \text{Den} = p_{k,k}\; a_k(\boldsymbol{\alpha}, 1) - \left(\sum_{i\neq k}\sum_{j\neq k} p_{i,j}\right) b_k(\boldsymbol{\alpha}, 1) + \\ + 2\left(\sum_{i\neq k} p'_{i,k}\right) c_k(\boldsymbol{\alpha}, 0.5) \cdot \lim_{x\to 0^+}\left[\frac{1}{2}\left(\frac{1-x}{x}\right)^{0.5} - \frac{1}{2}\left(\frac{x}{1-x}\right)^{0.5}\right] \tag{8.8}$$

Note that $\lim_{x\to 0^+}\left[\frac{1}{2}\left(\frac{1-x}{x}\right)^{0.5} - \frac{1}{2}\left(\frac{x}{1-x}\right)^{0.5}\right] = +\infty$. Since $\lim_{x\to 0^+}$ Num is finite and it must be equal to $\lim_{x\to 0^+}$ Den, $\left(\sum_{i\neq k} p'_{i,k}\right)$ must be equal to 0 for any $k = 1, \ldots, D$. It follows that $\mathbf{P}'$ must be a diagonal matrix.

Combining these conclusions with constraint (8.3) brings that:

$$p_{k,k}\; a_k(\boldsymbol{\alpha}, 1) = p'_{k,k}\; a_k(\boldsymbol{\alpha}, 0.5) \qquad \Longrightarrow \qquad p'_{k,k} = p_{k,k}\; \frac{\alpha^+ + 1}{\alpha_k + 1},$$

where $a_k(\boldsymbol{\alpha}, 1) = \dfrac{(\alpha^+ + 1)\alpha^+\Gamma(\alpha^+)}{(\alpha_k + 1)\alpha_k\Gamma(\alpha_k)\Gamma(\alpha^+ - \alpha_k)}$ and $a_k(\boldsymbol{\alpha}, 0.5) = \dfrac{\alpha^+\Gamma(\alpha^+)}{\alpha_k\Gamma(\alpha_k)\Gamma(\alpha^+ - \alpha_k)}$.

Finally, if $\mathbf{P}'$ is diagonal, then:

$$\sum_{k=1}^{D} p'_{k,k} = 1 \qquad \Longrightarrow \qquad \sum_{k=1}^{D} p_{k,k}\; \frac{\alpha^+ + 1}{\alpha_k + 1} = 1 \qquad \Longrightarrow \qquad \sum_{k=1}^{D} \frac{p_{k,k}}{\alpha_k + 1} = \frac{1}{\alpha^+ + 1}. \tag{8.9}$$

It is easy to see that $p_{k,k} = \dfrac{\alpha_k(\alpha_k + 1)}{\alpha^+(\alpha^+ + 1)}$ satisfies this constraint. Showing that this is the only $p_{k,k}$ satisfying (8.9) is much more complicated and in order to have an identifiable model, it is sufficient to exclude every diagonal matrix $\mathbf{P}$ from the parametric space. □

Bibliography

[1] J. Aitchison. “A General Class of Distributions on the Simplex”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 47.1 (1985), pp. 136–146 (cit. on p. 26).

[2] J. Aitchison. “On criteria for measures of compositional difference”. In: *Mathematical Geology* 24.4 (1992), pp. 365–379 (cit. on p. 13).

[3] J. Aitchison. *The Statistical Analysis of Compositional data*. London: The Blackburn Press, 2003 (cit. on pp. vii, 2, 3, 8, 9, 17, 22–24).

[4] J. Aitchison. “The Statistical Analysis of Compositional Data (with discussion)”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 139–177 (cit. on pp. 3, 15, 25).

[5] J. Aitchison and S. M. Shen. “Logistic-Normal Distributions: Some Properties and Uses”. In: *Biometrika* 67.2 (1980), pp. 261–272 (cit. on pp. 3, 22).

[6] H. Akaike. “Information Theory and an Extension of the Maximum Likelihood Principle”. In: *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, 1973, pp. 199–213 (cit. on p. 102).

[7] R. Ascari, S. Migliorati, and A. Ongaro. “A special Dirichlet mixture model in a Bayesian perspective”. In: *Accepted for the Springer Book of CLAssification and Data Analysis Group (CLADAG), 13-15 September 2017. Milan, Italy* (cit. on pp. 4, 33).

[8] R. Ascari, S. Migliorati, and A. Ongaro. “The Extended Flexible Dirichlet model: a simulation study”. In: *Applied Stochastic Models and Data Analysis (ASMDA), 6-9 June 2017. London, UK* (cit. on pp. vii, 4).

[9] M. Avetisyan and J. P. Fox. “The Dirichlet-Multinomial Model for Multivariate Randomized Response Data and Small Samples.” In: *Psicologica: International Journal of Methodology and* 33 (2012), pp. 362–390 (cit. on pp. 149, 150).

[10] A. Azzalini and N. Torelli. “Clustering via nonparametric density estimation”. In: *Statistics and Computing* 17.1 (2007), pp. 71–80. arXiv: 1301.6559 (cit. on pp. 4, 133).

[11] Adelchi Azzalini and Giovanna Menardi. “Clustering via Nonparametric Density Estimation: The {R} Package {pdfCluster}”. In: *Journal of Statistical Software* 57.11 (2014), pp. 1–26 (cit. on p. 133).

[12] O. E. Barndorff-Nielsen and B. Jørgensen. “Some parametric models on the simplex”. In: *Journal of Multivariate Analysis* 39.1 (1991), pp. 106–116 (cit. on p. 21).

[13]C. L. Bayes, J. L. Bazán, and C. Garća. “A new robust regression model for proportions”. In: *Bayesian Analysis* 7.4 (2012), pp. 841–866. arXiv: arXiv:1011.1669v3 (cit. on p. 182).

[14]M. Betancourt. “A Conceptual Introduction to Hamiltonian Monte Carlo”. 2018 (cit. on p. 172).

[15]C. Biernacki, G. Celeux, and G. Govaert. “Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models”. In: *Computational Statistics & Data Analysis* 41 (2003), pp. 561–575 (cit. on pp. 33, 55, 95, 100).

[16]D. Billheimer, P. Guttorp, and W. F. Fagan. “Statistical analysis and interpretation of discrete compositional data”. In: *National Center for Statistics and the Environment* 011 (1998) (cit. on p. 10).

[17]D. Billheimer, P. Guttorp, W. F. Fagan, et al. “Statistical Interpretation of Species Composition”. In: 96.456 (2001), pp. 1205–1214 (cit. on pp. 149, 150, 154).

[18]K. G. van den Boogaart and R. Tolosana-Delgado. *Analyzing Compositional Data with R*. Springer, 2011 (cit. on pp. 15, 23).

[19]N. Bouguila. “Clustering of count data using generalized dirichlet multinomial distributions”. In: *IEEE Transactions on Knowledge and Data Engineering* 20.4 (2008), pp. 462–474 (cit. on p. 149).

[20]R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. “A Limited Memory Algorithm for Bound Constrained Optimization”. In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208. arXiv: arXiv:1011.1669v3 (cit. on pp. 55, 57).

[21]G. Celeux and G. Govaert. “A classification EM algorithm for clustering and two stochastic versions”. In: *Computational Statistics & Data Analysis - Special issue on optimization techniques in statistics* 14.3 (1992), pp. 315–332 (cit. on pp. 55, 100).

[22]G. Celeux, D. Chauveau, and J. Diebolt. *On Stochastic Versions of the EM Algorithm On Stochastic Versions of the EM Algorithm*. Tech. rep. INRIA, 1995 (cit. on pp. 55, 100).

[23]R. Connor and J. E. Mosimann. “Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution”. In: *Journal of the American Statistical Association* 64.325 (1969), pp. 194–206 (cit. on pp. 15, 21).

[24]J. N. Darroch and I. R. James. “F-Independence and Null Correlation of Continuous, Bounded-Sum, Positive Variables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.3 (1974), pp. 467–483 (cit. on p. 15).

[25]J. N. Darroch and D. Ratcliff. “A Characterization of the Dirichiet Distribution”. In: 66.335 (1971), pp. 641–643 (cit. on p. 15).

[26]J. N. Darroch and D. Ratcliff. “No-association of proportions”. In: *Journal of the International Association for Mathematical Geology* 10.4 (1978), pp. 361–368 (cit. on p. 15).

[27]J. N. Darroch and D. Ratcliff. “Null correlation for proportions - II”. In: *Mathematical Geology* 2.3 (1970), pp. 307–312 (cit. on p. 15).

[28]A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society Series B Methodological* 39.1 (1977), pp. 1–38. arXiv: 0710.5696v2 (cit. on pp. viii, 33, 54, 94).

[29]J. Diebolt and E. H. S. Ip. “Stochastic EM: method and application”. In: *Markov Chain Monte Carlo in practice*. 1996, pp. 259–273 (cit. on pp. 33, 56, 95).

[30]J. Diebolt and C. P. Robert. “Estimation of Finite Mixture Distributions through Bayesian Sampling”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 56.2 (1994), pp. 363–375 (cit. on p. 34).

[31]S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. “Hybird Monte Carlo”. In: *Physics Letters B* 195.2 (1987), pp. 216–222 (cit. on p. 172).

[32]J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. “Isometric Logratio Transformations for Compositional Data Analysis”. In: *Mathematical Geology* 35.3 (2003), pp. 279–300 (cit. on p. 25).

[33]P. J. Farrell, M. Salibian-Barrera, and K. Naczk. “On tests for multivariate normality and associated simulation studies”. In: *Journal of Statistical Computation and Simulation* 77.12 (2007), pp. 1065–1080 (cit. on p. 23).

[34]S. Favaro, G. Hadjicharalambous, and I. Prünster. “On a class of distributions on the simplex”. In: *Journal of Statistical Planning and Inference* 141.9 (2011), pp. 2987–3004 (cit. on p. 21).

[35]S. L. P. Ferrari and F. Cribari-Neto. “Beta regression for modelling rates and proportions”. In: *Journal of Applied Statistics* 31.7 (2004), pp. 799–815 (cit. on p. 182).

[36]M. Forina, Armanino C., S. Lanteri, and E. Tiscornia. “Classication of olive oils from their fatty acid composition , in Food Research and Data Analysis”. In: *Food Research and Data Analysis* (1983) (cit. on pp. 4, 133).

[37]B. Frigyik, A. Kapila, and M. R. Gupta. “Introduction to the Dirichlet Distribution and Related Processes”. In: *Electrical Engineering* 27.206 (2010), pp. 46–48,50–52 (cit. on p. 17).

[38]S. Frühwirth-Schnatter. *Finite mixture and markov switching models*. New York: Springer, 2006. arXiv: `arXiv:1011.1669v3` (cit. on pp. 3, 4, 28, 33, 34, 52, 105).

[39]A. Gelman, J. B. Carlin, H. S. Stern, et al. *Bayesian Data Analysis*. Third. CRC Press, 2014 (cit. on pp. 3, 33, 34).

[40]M. Grigoletto, F. Lisi, and S. Petrone, eds. *Complex Models and Computational Methods in Statistics*. 1st ed. Springer-Verlag Mailand, 2013 (cit. on p. 4).

[41]A. K. Gupta and D Song. “Generalized Liouville Distribution”. In: 32.2 (1996), pp. 103–109 (cit. on p. 21).

[42]R. D. Gupta and D. St. P. Richards. “Multivariate Liouville distributions”. In: *Journal of Multivariate Analysis* 23 (1987), pp. 233–256 (cit. on p. 21).

[43]R. D. Gupta and D. St. P. Richards. “Multivariate Liouville distributions, II”. In: *Probability and Mathematical Statistics* 12 (1991), pp. 291–309 (cit. on p. 21).

[44]R. D. Gupta and D. St. P. Richards. “Multivariate Liouville distributions, III”. In: *Journal of Multivariate Analysis* 43 (1992), pp. 29–57 (cit. on p. 21).

[45]R. D. Gupta and D. St. P. Richards. “Multivariate Liouville distributions, IV”. In: *Journal of Multivariate Analysis* 54.1 (1995), pp. 1–17 (cit. on p. 21).

[46]R. D. Gupta and D. St. P. Richards. “Multivariate Liouville distributions, V”. In: *Advances in the Theory and Practice of Statistics: a volume in honour of Samuel Kotz*. Ed. by N. L. Johnson and N. Balakrishnan. New York: Wiley, 1997, pp. 377–396 (cit. on p. 21).

[47]E. D. Hahn. “Mixture densities for project management activity times: A robust approach to PERT”. In: *European Journal of Operational Research* 188.2 (2008), pp. 450–459 (cit. on p. 182).

[48]L. J. Halliwell. *The Lognormal Random Multivariate*. 2015 (cit. on p. 23).

[49]I. Holmes, K. Harris, and C. Quince. “Dirichlet multinomial mixtures: Generative models for microbial metagenomics”. In: *PLoS ONE* 7.2 (2012) (cit. on pp. 149, 150).

[50]R. J. Howarth. “Sources for a history of the ternary diagram”. In: *The British Journal for the History of Science* 29.03 (1996), p. 337 (cit. on p. 14).

[51]G. Hughes, G. P. Munkvold, and S. Samita. “Application of the logistic-normal-binomial distribution to the analysis of Eutypa dieback disease incidence”. In: *International Journal of Pest Management* 44.1 (1998), pp. 35–42 (cit. on p. 150).

[52]I. R. James and J. E. Mosimann. “A new characterization of the dirichlet distribution through neutrality”. In: *The Annals of Statistics* 8.1 (1980), pp. 183–189 (cit. on p. 15).

[53]R. Kieschnick and B. D. Mccullough. “Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions”. In: *Statistical Modeling* 3.3 (2003), pp. 193–213 (cit. on p. 182).

[54]S. Kullback and R. A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. arXiv: `1511.00860` (cit. on pp. 26, 52).

[55]J. D. Leckenby and S. Kishi. “The Dirichlet multinomial distribution as a magazine exposure model”. In: *Journal of Marketing Research* XXI.February (1984), pp. 100–106 (cit. on p. 149).

[56]Thomas A. Louis. “Finding the Observed Information Matrix when Using the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 226–233 (cit. on p. 56).

[57]D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. “WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility”. In: *Statistics and Computing* 10 (2000), pp. 325–337. arXiv: `arXiv:1011.1669v3` (cit. on p. 37).

[58]M. Markatou. “Mixture models, robustness, and the weighted likelihood methodology”. In: *Biometrics* 56.2 (2000), pp. 483–486 (cit. on p. 4).

[59]A. W. Marshall and I Olkin. *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Pr., 1979 (cit. on p. 21).

[60]G. McLachlan and D. Peel. *Finite Mixture Models*. 2005 (cit. on pp. 4, 52, 105).

[61]S. Migliorati and A. Ongaro. “The Extended Flexible Dirichlet model : some theoretical and computational issues”. In: *Applied Stochastic Models and Data Analysis (ASMDA), 30 June – 4 July 2015. Piraeus, Greece*. 2015 (cit. on pp. vii, 4).

[62]S. Migliorati, A. M. Di Brisco, and A. Ongaro. “A new regression model for bounded responses”. In: *Bayesian Analysis* 13.3 (2018), pp. 845–872 (cit. on pp. 35, 182).

[63]S. Migliorati, A. Ongaro, and G. S. Monti. “A structured Dirichlet mixture model for compositional data: inferential and applicative issues”. In: *Statistics and Computing* 27.4 (2017), pp. 963–983 (cit. on pp. vii, 3, 29, 33, 56, 57, 133).

[64]J. E. Mosimann. “On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions”. In: *Biometrika* 49.1/2 (1962), pp. 65–82 (cit. on p. 149).

[65]K. P. Murphy. *Machine Learning: A probabilistic perspective*. The MIT Press, 2012. arXiv: 0-387-31073-8 (cit. on p. 153).

[66]R. M. Neal. *An improved acceptance procedure for the hybrid monte carlo algorithm*. 1994. arXiv: 9208011 [hep-lat] (cit. on p. 172).

[67]K. W. Ng, M. L. Tang, M. Tan, and G. L. Tian. “Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis”. In: *Journal of Multivariate Analysis* 99.3 (2008), pp. 490–509 (cit. on p. 21).

[68]K. W. Ng, M. L. Tang, G. L. Tian, and M. Tan. “The Nested Dirichlet distribution and incomplete categorical data analysis”. In: *Statistica Sinica* 19 (2009), pp. 251–271 (cit. on p. 21).

[69]I. Ntzoufras. *Bayesian Modeling Using WinBUGS*. Vol. 698. 2009, p. 592 (cit. on p. 37).

[70]A. O’Hagan, T. B. Murphy, and I. C. Gormley. “Computational aspects of fitting mixture models via the expectation-maximization algorithm”. In: *Computational Statistics and Data Analysis* 56.12 (2012), pp. 3843–3864 (cit. on pp. 33, 95).

[71]A. Ongaro and S. Migliorati. “A Dirichlet Mixture Model for Compositions Allowing for Dependence on the Size”. In: *Advances in Latent Variables*. November 2014. 2014, pp. 101–111 (cit. on pp. vii, 4, 21).

[72]A. Ongaro and S. Migliorati. “A generalization of the dirichlet distribution”. In: *Journal of Multivariate Analysis* 114.1 (2013), pp. 412–426 (cit. on pp. vii, 3, 21, 26, 27).

[73]Philip Paolino. “Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables”. In: *Political Analysis* 9.04 (2001), pp. 325–346 (cit. on p. 182).

[74]V. Pawlowsky-Glahn and J. J. Egozcue. “BLU estimators and compositional data”. In: *Mathematical Geology* 34.3 (2002), pp. 259–274 (cit. on p. 13).

[75]V. Pawlowsky-Glahn and J. J. Egozcue. “Geometric approach to statistical analysis on the simplex”. In: *Stochastic Environmental Research and Risk Assessment* 15.5 (2001), pp. 384–398 (cit. on p. 13).

[76]K. Pearson. “Mathematical Contributions to the Theory of Evolution . On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs”. In: *Proceedings of the Royal Society of London* 60 (1897), pp. 489–498 (cit. on pp. 3, 14).

[77]R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018 (cit. on p. 35).

[78]W. S. Rayens and C. Srinivasan. “Dependence properties of generalized Liouville distributions on the simplex”. In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1465–1470 (cit. on p. 21).

[79]C. P. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. 2009. arXiv: arXiv:1011.1669v3 (cit. on p. 35).

[80]R. T. Rust and R. P. Leone. "The Mixed-Media Dirichlet Multinomial Distribution: A Model for Evaluating Television-Magazine Advertising Schedules". In: *Journal of Marketing Research* 21.1 (1984), pp. 89–99 (cit. on p. 150).

[81]H. Scheffé. "Experiments With Mixtures". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2 (1958), pp. 344–360 (cit. on p. 149).

[82]G. Schwarz. "Estimating the Dimension of a Model". In: *Annals of Statistics* 2 (6), pp. 461–464 (cit. on p. 102).

[83]B. D. Sivazlian. "On a Multivariate Extension of the Gamma and Beta Distributions". In: *SIAM Journal on Applied Mathematics* 41.2 (1981), pp. 205–209 (cit. on p. 21).

[84]J. G. Skellam. "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10.2 (1948), pp. 257–261 (cit. on p. 149).

[85]B. Smith and W. Rayens. "Conditional generalized liouville distributions on the simplex". In: *Statistics* 36.2 (2002), pp. 185–194 (cit. on pp. 21, 51, 52).

[86]Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. 2017. arXiv: 1404.4866 (cit. on p. 172).

[87]W. Stuetzle. "Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample". In: *Journal of Classification* 20 (2003), pp. 25–47 (cit. on pp. 4, 133).

[88]F. Xia, J. Chen, W. K. Fung, and H. Li. "A logistic normal multinomial regression model for microbiome compositional data analysis". In: *Biometrics* 69.4 (2013), pp. 1053–1063 (cit. on pp. 149, 150).

[89]Z. Xu and R. Akella. "A new probabilistic retrieval model based on the dirichlet compound multinomial distribution". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08* (2008), p. 427 (cit. on pp. 149, 150).

[90]J. Yin and J. Wang. "A dirichlet multinomial mixture model-based approach for short text clustering". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (2014), pp. 233–242 (cit. on p. 150).

[91]Y. Zhang and H. Zhou. *MGLM: Multivariate Response Generalized Linear Models*. 2018 (cit. on p. 154).

[92]Y. Zhang, H. Zhou, J. Zhou, and W. Sun. "Regression Models for Multivariate Count Data". In: *Journal of Computational and Graphical Statistics* 26.1 (2017), pp. 1–13 (cit. on p. 154).

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.