INTERNSHIP REPORT

Outlier detection using ICS for compositional data

Author:
Camille MONDON
DMA
École normale supérieure

Supervisor:
Anne RUIZ-GAZEN
UMR-TSE-R
Toulouse School of Economics

A report submitted in fulfillment of the requirements for the first year of Master's degree

in the

Département de Mathématiques et Applications
École normale supérieure

May 2021 – September 2021

Abstract

A multivariate data set for which each variable can be interpreted as part of a whole is called a compositional data set. In that case, it is preferable to decrease by one the dimension and consider it as embedded in the simplex, which is equipped with its own Euclidian structure. Thus, adapting a method such as Invariant Coordinate Selection (ICS) to compositional data is an interesting challenge since ICS has proved itself useful on classical multivariate data sets to reveal hidden structures such as outliers or groups, but cannot be applied as such on the simplex.
This internship report summarises the main aspects of compositional data analysis and algebra, in order to adapt the ICS method to outlier detection for compositional data.

Keywords: Invariant Coordinate Selection, compositions, outliers, multivariate data analysis, kurtosis, automotive market.

Acknowledgements

I would like to express my gratitude to my internship supervisor Pr. Anne Ruiz-Gazen, for putting great effort into providing guidance whenever needed. I am also very grateful to Pr. Christine Thomas-Agnan and Thibault Laurent, with whom I have worked all along this internship.

Contents

Notations

- $\mathcal{S}^D = \left\{(x_1, \ldots, x_D)^\top \in {\mathbb{R}_+^*}^D, \sum_{1 \leq i \leq D} x_i = 1\right\}$ is the simplex in dimension D, and its elements are called compositions.
- If $\mathbf{z} \in {\mathbb{R}_+^*}^D$, $\mathcal{C}(\mathbf{z})$ is the unique composition which is collinear to $\mathbf{z}$.
- $\mathbf{x} = (x_1, \ldots, x_D)^\top$ and $\mathbf{y} = (y_1, \ldots, y_D)^\top$ are two compositions.
- $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left((x_1 y_1, \ldots, x_D y_D)^\top\right)$ is the Aitchison addition (or perturbation).
- If $\alpha \in \mathbb{R}$, $\alpha \odot \mathbf{x} = \mathcal{C}\left({x_1}^\alpha, \ldots, {x_D}^\alpha)^\top\right)$ is the Aitchison scalar multiplication (or powering).
- $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{1 \leq i < j \leq D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)$ is the Aitchison inner product.
- $\mathcal{B} = (\mathbf{e}_1, \ldots, \mathbf{e}_{D-1})$ is an orthonormal basis of $\mathcal{S}^D$, ilr and V are respectively its associated transformation and its contrast matrix.
- $\mathbf{1}_D = (1, \ldots, 1)^\top \in \mathbb{R}^D$ is the vector whose entries are all ones, I_D denotes the $D \times D$ identity matrix, and G_D denotes the $D \times D$-matrix $\mathrm{I}_D - \frac{1}{D} \mathbf{1}_D {\mathbf{1}_D}^\top$.
- $\mathbf{X}$ is a random variable on $\mathbb{R}^D$, whose distribution function is $F_{\mathbf{X}} : \mathcal{B}(\mathbb{R}^D) \to \mathbb{R}$, and if absolutely continuous with respect to the Lebesgue measure, its probability density function is $f_{\mathbf{X}} : \mathbb{R}^D \to \mathbb{R}$.
- $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ are independent random variables such that each one follows the distribution $F_{\mathbf{X}}$, and that represent the statistical observations.
- If θ is a parameter of the distribution function of $\mathbf{X}$, we will use as often as possible the notation $\hat{\theta}$ for an estimator of θ, except in particular for $\mathbb{E}[\mathbf{X}]$ (estimated by the sample mean $\overline{\mathbf{X}_j}$) and the variance $\mathbb{V}[\mathbf{X}]$ (estimated by the sample variance matrix $\hat{\Sigma} = \hat{\Sigma}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$). We will omit the dependence on $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as often as possible.

Introduction

Context & goals

This report summarises the advancements that have been made during my research internship at TSE-R, between May 2021 and August 2021, under the supervision of Pr. Anne Ruiz-Gazen, with the collaboration of Pr. Christine Thomas-Agnan and Thibault Laurent.

The aim of my internship was to prepare the foundation for the redaction of a publication on the following matter: *'Outlier detection using the Invariant Coordinate Selection method for compositional data'*.

The first step was to learn the basics of compositional algebra and data analysis, which are Christine Thomas-Agnan's research interests; the second step was to understand the Invariant Coordinate Selection method, which relies on the concepts of scatter operators and kurtosis, and was conceived during Anne Ruiz-Gazen's thesis; the third step was to adapt the latter to the former, and to develop a good sense of what could be adapted, what needed to be changed and what could not work when moving from general multivariate data to compositional data.

Application to the `BDDSegX` data set

Once the foundations are laid, the idea was to apply the newly conceived method to a real data set (by designing - for instance - an R package called `ICSCoDa`) to test the validity of the results that were found on this particular data set, and to interpret this compositional data set thanks to the Invariant Coordinate Selection method.

The `BDDSegX` data set on the automotive market was chosen to illustrate the notions and compare the methods that were exposed. This data set was generated by Christine Thomas-Agnan and Joanna Morais and contains (among other variables) the monthly evolution from 2003 to 2015 of each volume of the five-segment car classification according to their size (from the A-segment regrouping city cars to the E-segment regrouping executive cars), with a total of 152 observations.

It is relevant to see this data set as compositional data, provided of course that one is not interested in studying the variation of the total volume of the automotive market.

Chapter 1

Prerequisites

1.1 Multivariate outlier detection and ICS

A common problem in multivariate data analysis is that of detecting outliers in a contaminated data set. If we denote $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ the set of independent observations of a random variable $\mathbf{X}$ on $\mathbb{R}^D$, we need to model a contamination process from the original uncontaminated random variable $\mathbf{X}_0$ to the contaminated $\mathbf{X}$.

There are many ways to design contaminated data, among them to model the data generating process (DGP) by a convex combination (called a mixture) of the original distribution function with other distributions modelling the outliers:

$$F_{\mathbf{X}} = (1-\varepsilon)F_0 + \sum_{k=1}^{k=q} \varepsilon_k F_k$$

where $F_0 = F_{\mathbf{X}_0}$, $\varepsilon = \sum_{k=1}^{k=q} \varepsilon_k$ is the theoretical proportion of outliers in the data and if $1 \leq k \leq q$, F_k denotes the distribution function of the kth cloud of outliers.

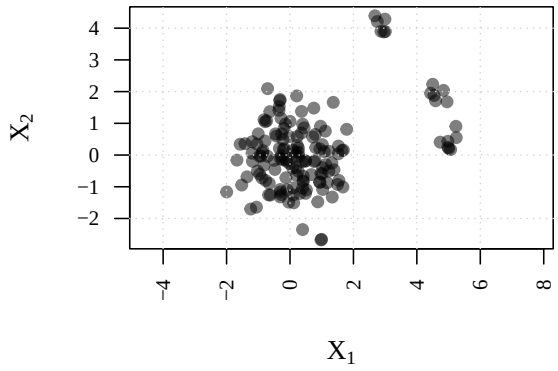


Figure 1.1: $n = 150$ points generated from a Gaussian mixture model with $D = 2$, $q = 3$, $(\varepsilon_1, \varepsilon_2) = (0.05, 0.05)$.

Example. Figure 1.1 gives a scatter plot of $n = 150$ observations generated in $D = 2$ dimensions from a mixture of $q = 3$ Gaussian distributions with $\varepsilon = 0.1$, $\varepsilon_1 = 0.05$ and $\varepsilon_2 = 0.05$. The two groups of outliers are clearly visible on this plot. But this may not be the case when the dimension D is larger than 3.

1.1.1 Elliptical distributions

The outlier detection method studied in this report relies on the lack of elliptical symmetry of the contaminated distribution, so we have to properly define elliptical distributions, which are a generalization of normal distributions.

Definition 1.1 (Elliptical distributions)**.** Let μ be a vector of $\mathbb{R}^D$ and $\Sigma \in \mathbb{R}^{D\times D}$ a positive semi-definite symmetric matrix. A random variable $\mathbf{X}$ is said to follow a (μ,Σ)-elliptical distribution if there exists a random variable such that:

(i) $\mathbf{X} = \Sigma^{\frac{1}{2}}\mathbf{Z} + \mu$;

(ii) $\mathbf{Z}$ has a spherical distribution, meaning that for every orthogonal matrix Q the equality of distribution functions $\mathbf{Z} \sim \mathrm{Q}\,\mathbf{Z}$ holds.

Examples.

- The multivariate Gaussian distribution with mean μ and covariance matrix Σ is the most well-known elliptical distribution.
  Figure 1.2 gives a scatter plot of $n = 150$ observations generated in two dimensions with parameters $\mu = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. The ellipse corresponds to a quantile of order 0.98.
- Other distributions such as the multivariate Student distributions also belong to the family of elliptical distributions.

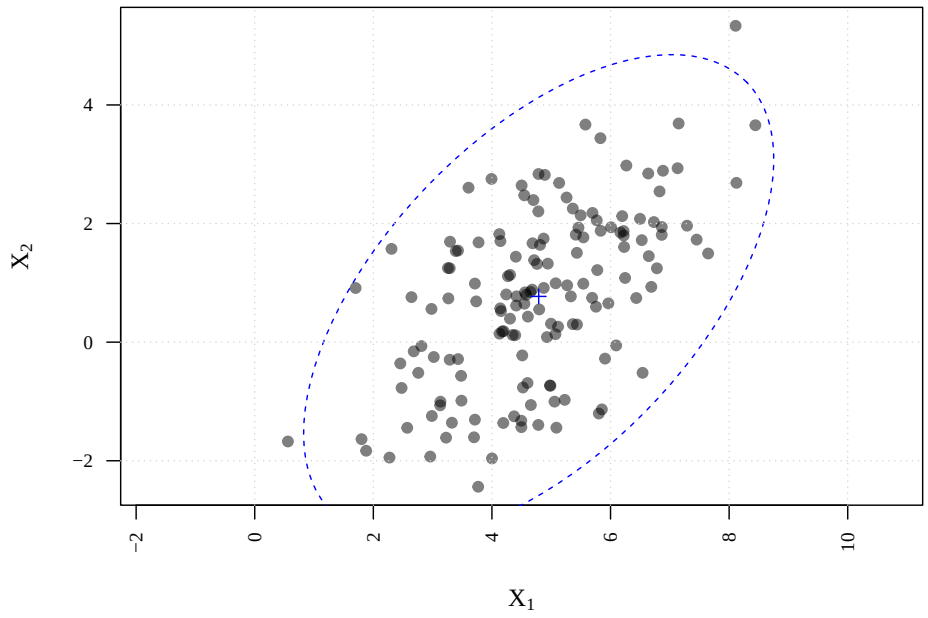


Figure 1.2: $n = 150$ points generated from a normal distribution with $\mu = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

Remark. For an elliptical distribution $\mathbf{X}$, the vector parameter and the matrix parameter are uniquely defined by the distribution function of $\mathbf{X}$, and in particular are functions of $\mathbf{X}$ denoted respectively by $\mu[\mathbf{X}]$ and $\Sigma[\mathbf{X}]$.
This fact can be simply illustrated in the case of absolutely continuous elliptical distributions, since the density function is determined by its values on the Σ-shaped ellipses around μ (see the following result, cited in Tyler et al. 2009).

Proposition 1.2. *An absolutely continuous (μ, Σ)-elliptical random variable* $\mathbf{X}$ *on* $\mathbb{R}^D$ *has a density function verifying, for every* $\mathbf{x}$ *in* $\mathbb{R}^D$*:*

$$f_{\mathbf{X}}(\mathbf{x}) = C_D \det(\Sigma)^{-\frac{1}{2}} g\big((\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\big)$$

where $g : \mathbb{R}_+ \to \mathbb{R}_+$ *is a non-negative function and* $C_D \in \mathbb{R}_+^*$ *is a constant.*

1.1.2 Location operators and scatter functionals

As elliptical distributions generally do not admit first and second moments, it is necessary to define generalized notions of expectation (location) and variance (scatter).

Definition 1.3 (Location and scatter)**.** A location m is an operator from a subspace of random variables on $\mathbb{R}^D$ onto $\mathbb{R}^D$ which verifies the following two properties of the usual expectation $\mathbb{E}$, for all random variables $\mathbf{X}$ and $\mathbf{Y}$ in the considered subspace, and every non-singular matrix $\mathrm{A} \in \mathbb{R}^{D\times D}$ and $b \in \mathbb{R}^D$:

(i) if $\mathbf{X} \sim \mathbf{Y}$, $m[\mathbf{X}] = m[\mathbf{Y}]$ (dependent on the distribution only);

(ii) $m[\mathrm{A}\mathbf{X} + \mathbf{b}] = \mathrm{A}\, m[\mathbf{X}] + \mathbf{b}$ (affine equivariant).

A scatter S is a functional from a subspace of random variables on $\mathbb{R}^D$ onto the convex cone of positive-definite matrices $\mathbb{R}^D$ which verifies the following two properties of the usual (co-)variance $\mathbb{V}$, for all random variables $\mathbf{X}$ and $\mathbf{Y}$ in the considered subspace, and every non-singular matrix $\mathrm{A} \in \mathbb{R}^{D\times D}$ and $b \in \mathbb{R}^D$:

(i) if $\mathbf{X} \sim \mathbf{Y}$, $S[\mathbf{X}] = S[\mathbf{Y}]$ (dependent on the distribution only);

(ii) $S[\mathrm{A}\mathbf{X} + \mathbf{b}] = \mathrm{A}\, S[\mathbf{X}]\, \mathrm{A}^\top$ (affine equivariant).

Naturally, the subspace of elliptical distributions is stable under any affine transformation of $\mathbb{R}^D$ and on this space, the vector parameter μ defines a location operator while the matrix parameter Σ defines a scatter functional. There is a converse result (cited in Tyler et al. 2009):

Proposition 1.4. *A location operator* m *and a scatter functional* S *defined on a subspace containing the elliptically distributed random variables on* $\mathbb{R}^D$*, have to verify for every elliptically distributed random variable* $\mathbf{X}$*:*

$$m[\mathbf{X}] = \mu[\mathbf{X}] \text{ and } S[\mathbf{X}] = \lambda\, \Sigma[\mathbf{X}]$$

where $\lambda \in \mathbb{R}_+$ *is a constant depending on the distribution function of* $\mathbf{X}$*.*

Proof. Let $\mathbf{X}$ be a (μ, Σ)-elliptically distributed random variable on $\mathbb{R}^D$, and $\mathbf{Z}$ a random variable as in Definition 1.1. Then if Q is an orthogonal matrix:

- $m[\mathbf{Z}] = m[\mathrm{Q}\mathbf{Z}] = \mathrm{Q}\, m[\mathbf{Z}]$ i.e. $m[\mathbf{Z}] = 0$ (for $\mathrm{Q} = -\mathrm{I}_D$), so $m[\mathbf{X}] = m[\Sigma^{\frac{1}{2}}\mathbf{Z} + \mu] = \mu$.
- $S[\mathbf{Z}] = S[\mathrm{Q}\,\mathbf{Z}] = \mathrm{Q}\, S[\mathbf{Z}]\, \mathrm{Q}^\top$ i.e. $S[\mathbf{X}]\,\mathrm{Q} = \mathrm{Q}\, S[\mathbf{X}]$ i.e. $S[\mathbf{Z}] = \lambda\, \mathrm{I}_D$ for a $\lambda \in \mathbb{R}_+$ (by showing for instance that $S[\mathbf{Z}]$ must stabilize every line), so $S[\mathbf{X}] = S[\Sigma^{\frac{1}{2}}\mathbf{Z} + \mu] = \Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}})^\top = \Sigma$. □

Remark. Note that for a given elliptical distribution (generally the Gaussian distribution) and a given scatter matrix, the constant λ can be calculated and used to define a scatter matrix equal to the scatter parameter (see the example below).

Example. An important scatter functional, other that $\mathbb{V}$, is the scatter matrix of fourth order moments:

$$\mathbb{V}_4[\mathbf{X}] = \frac{1}{D+2}\mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^\top \mathbb{V}[\mathbf{X}]^{-1}(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^\top\right]$$

which is defined (for instance by Archimbaud, Nordhausen, and Ruiz-Gazen 2018a) for random variables admitting the first four moments, and will be particularly useful for the ICS outlier detection method.
Note that the constant $D+2$ has been calculated in order to verify that $\mathbb{V}_4[\mathbf{X}] = \Sigma[\mathbf{X}]$ for a Gaussian distribution.

As usual in statistics, we consider a sample $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ which is a set of independent and identically distributed random variables. We work conditionally to the observations $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$, which in that case amounts to treating them as constants. We will use estimated location operators $\hat{m}(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ and estimated scatter functionals $\hat{S}(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ which can be defined from the particular case of Definition 1.3 when the random variables $\mathbf{X}$ and $\mathbf{Y}$ are discrete (their probability distribution function is the empirical distribution function of the observations).

1.1.3 Whitening matrices and Mahalanobis distance

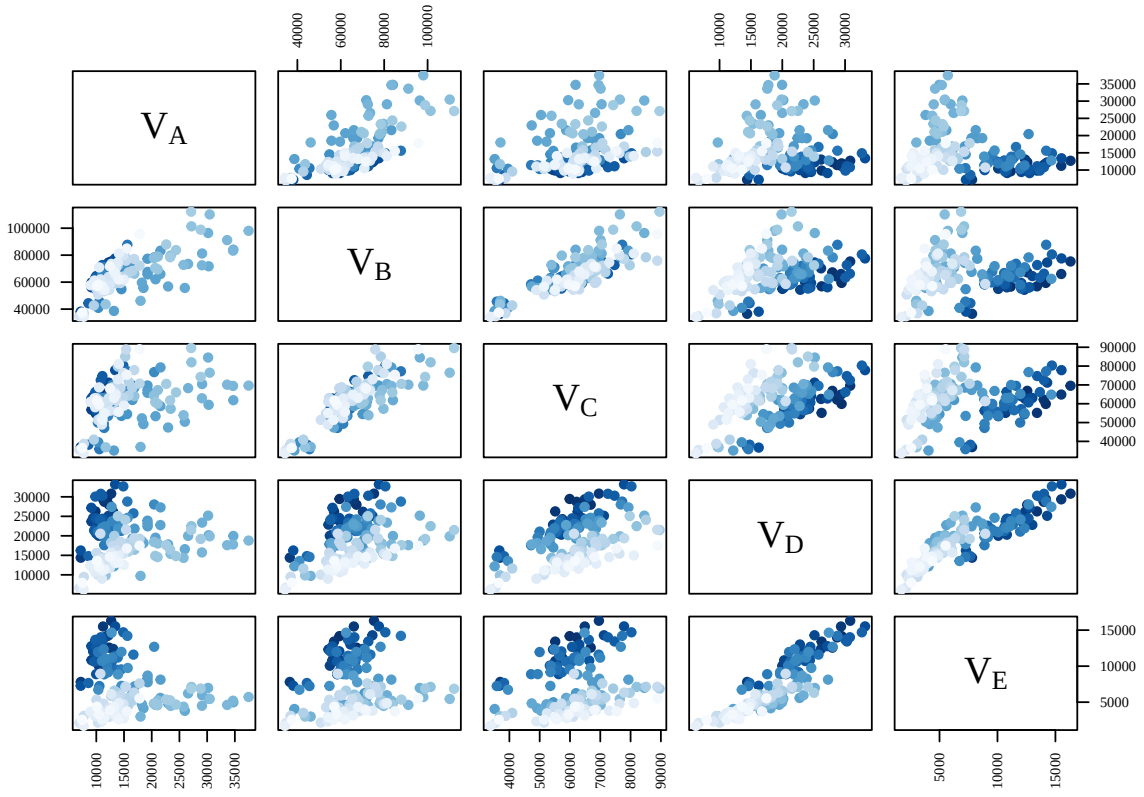


Figure 1.3: `BDDSegX` data set (volumes), original.

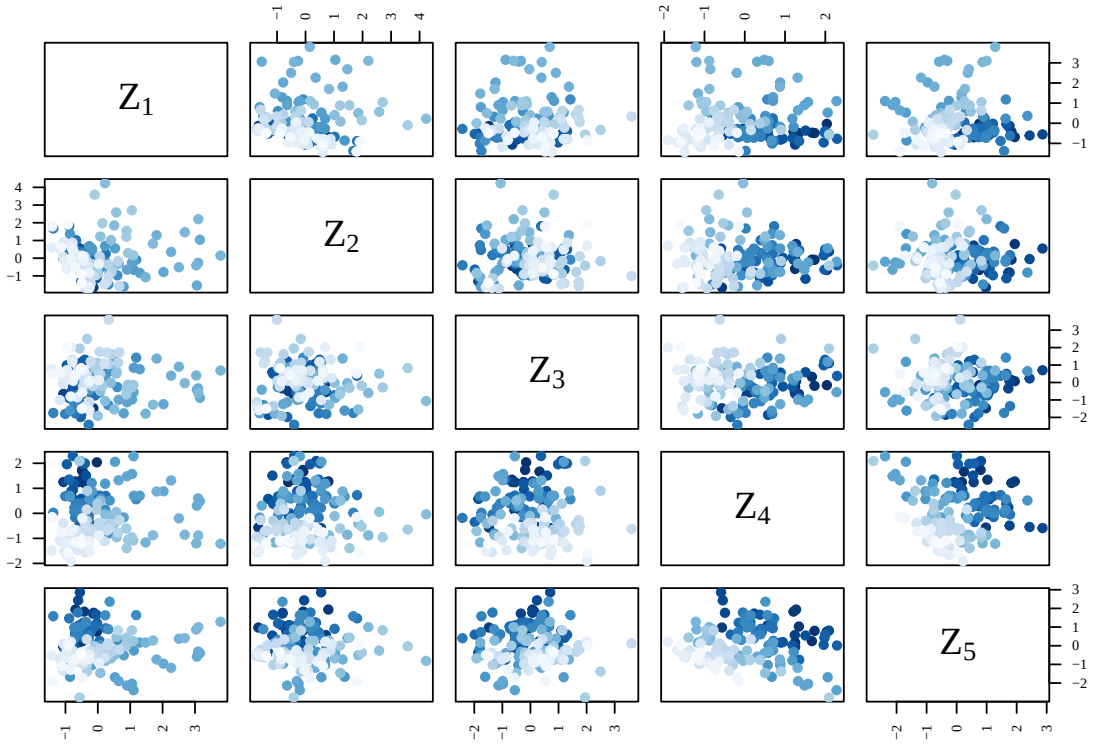


Figure 1.4: `BDDSegX` data set (volumes), whitened.

A first method to detect outliers is to center and whiten the data, which means applying a non-singular matrix $\hat{\mathrm{W}}$ such that the transformed data set:

$$\left(\mathbf{Z}_1 \,\middle|\, \dots \,\middle|\, \mathbf{Z}_n \right) = \hat{\mathrm{W}} \left(\mathbf{Y}_1 \,\middle|\, \dots \,\middle|\, \mathbf{Y}_n \right) \text{ (where } \mathbf{Y}_j = \mathbf{X}_j - \hat{m}(\mathbf{X}_1, \dots, \mathbf{X}_n), 1 \leq j \leq n)$$

has an estimated location $\hat{m}(\mathbf{Z}_1, \dots, \mathbf{Z}_n) = 0$ and an estimated scatter $\hat{S}(\mathbf{Z}_1, \dots, \mathbf{Z}_n) = \mathrm{I}_D$ (standardized and decorrelated).

If $\hat{\mathrm{W}}$ is a positive-definite matrix, applying this transformation corresponds to a change of inner product and of Euclidian norm, meaning for $\mathbf{x}$ in $\mathbb{R}^D$:

$$\left\| \hat{\mathrm{W}}\mathbf{x} \right\| = \sqrt{\left(\hat{\mathrm{W}}\mathbf{x}\right)^\top \left(\hat{\mathrm{W}}\mathbf{x}\right)} = \sqrt{\mathbf{x}^\top \hat{\mathrm{W}}^2 \mathbf{x}}$$

The most common choice is $\hat{\mathrm{W}} = \hat{\mathrm{S}}^{-\frac{1}{2}}$ (the inverse square root of a scatter functional evaluated on the observations) which corresponds to what we will call the 'Mahalanobis transformation'. The distance associated to this transformation is called the 'Mahalanobis distance' (from the work of Mahalanobis 1936).

This transformation can be applied to an absolutely continuous mixture model of elliptical distributions, for which the outliers will be the observations $\mathbf{X}_j, 1 \leq j \leq n$ such that:

$$\|\mathbf{Z}_j\| = \left\| \hat{\mathrm{S}}^{-\frac{1}{2}} \left(\mathbf{X}_j - \hat{m} \right) \right\| \geq Q(p)$$

where Q is the quantile function of the uncontaminated elliptical distribution X_0 (i.e. the generalized inverse of the cumulative distribution function associated with the function g defined in Proposition 1.2) and $p \in (0,1)$ is the level of confidence required.

When second order moments exist, we can choose for $\hat{S}$ the sample (co-)variance matrix $\hat{\Sigma}$ and for $\hat{m}$ the sample mean. This allows an interpretation of the newly defined variables $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ by computing the (co-)variance matrix with the original variables $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$:

$$\hat{\Sigma}_{(\mathbf{Z}_i, \mathbf{X}_j)} = \hat{W}^{-1} = \hat{\Sigma}^{\frac{1}{2}} \text{ (since } \hat{\Sigma}(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = I_D \text{)}$$

or the correlation matrix:

$$\hat{\rho}_{(\mathbf{Z}_i, \mathbf{X}_j)} = \left(\frac{\left(\hat{W}^{-1}\right)_{i,j}}{\sqrt{\sum_{1 \le k \le D} \left(\hat{W}^{-1}\right)^2_{k,j}}} \right)$$

Example. Figures 1.3 and 1.4 give matrix scatter plots for the `BDDSegX` automotive market example. The original volumes $\mathtt{V}_1, \ldots, \mathtt{V}_5$ are plotted on the upper graph and exhibit some correlations, while the whitened date $\mathtt{Z}_1, \ldots, \mathtt{Z}_5$ can be found on the lower graph and have a zero mean vector and a sample (co-)variance matrix equal to the identity.
Note that the dark blue corresponds to the beginning of the period (2003) and becomes lighter as the years go by in order to visualize also the temporal aspect of the data.

As the outliers are isolated due to their large deviation with respect to the estimated location, two problems emerge.

(1) The first problem is the impact of the outliers on the computation of the estimated location and scatter (and of the quantile function of the unknown uncontaminated distribution). This problem can be solved by using robust estimators (see Tyler et al. 2009) but is not considered further in the present report.

(2) The second problem is that this method implicitly assumes that outliers are the observations deviating most from the average. If outliers were defined according to another model, such as the elliptical mixture model, it would produce false positives (observations that were part of the uncontaminated distribution are detected as outliers) or false negatives (observations that were part of the outlier clouds are not detected).

1.1.4 The Invariant Coordinate Selection method

The general idea of ICS is to exploit the differences between two scatter operators evaluated at $\mathbf{X}$ to deduce a lack of ellipticity of the contaminated distribution, and then find projection directions that helps highlighting the outliers.

Figure 1.5 illustrates the fact that for a Gaussian distribution the two scatter matrices $\mathbb{V}$ and $\mathbb{V}_4$ are the same (*left panel*). But when considering a contaminated distribution, such as the one of Figure 1.1, the two scatter matrices differ as can be noticed on the *right panel* of Figure 1.5 where the ellipse associated with $\mathbb{V}_4$ (*in red*) is elongated in the area where the outliers lie, compared to the ellipse associated with $\mathbb{V}$ (*in blue*).

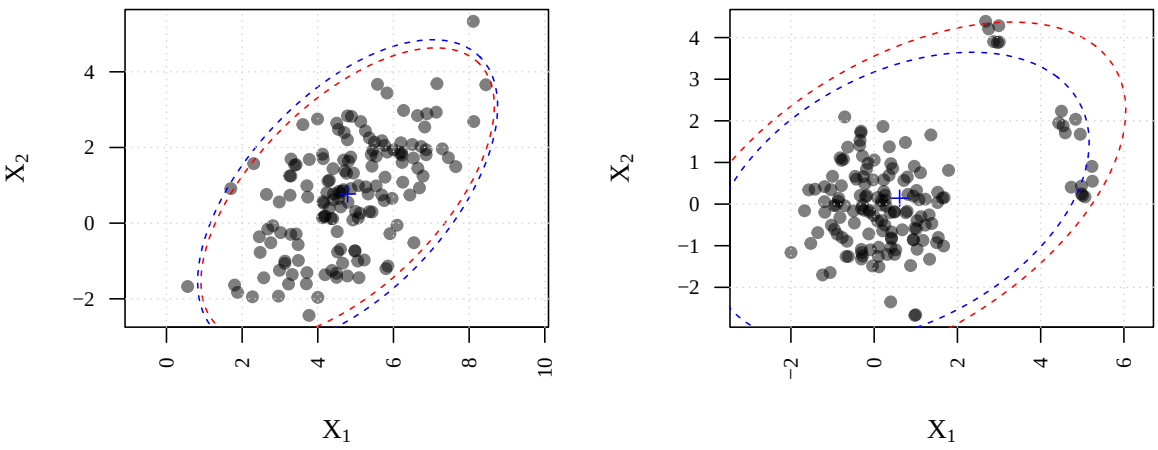


Figure 1.5: Computation of COV (blue) and COV_4 (red) ellipses for an elliptical (*left*) and a non-elliptical (*right*) distribution.

For distribution functions admitting the first four moments, the most common choice as the scatter pair is indeed $(\mathbb{V}, \mathbb{V}_4)$, which corresponds to the Fourth Order Blind Identification (FOBI) algorithm known since 1989 in the Blind Source Separation literature (see Cardoso 1989).

Generally speaking, comparing two estimated scatter matrices:

$$\left(\hat{\mathrm{S}}_1, \hat{\mathrm{S}}_2\right) = \left(\hat{S}_1(\mathbf{X}_1, \ldots, \mathbf{X}_n), \hat{S}_2(\mathbf{X}_1, \ldots, \mathbf{X}_n)\right)$$

is done by simultaneous diagonalization, which is closely related to the diagonalization of $\hat{\mathrm{S}}_1^{-1}\hat{\mathrm{S}}_2$ (as we will see at the end of this paragraph).

In order to do this, the natural idea is to symmetrize $\hat{\mathrm{S}}_1^{-1}\hat{\mathrm{S}}_2$ through conjugation by $\hat{\mathrm{S}}_1^{-\frac{1}{2}}$ which results in the similar positive-definite matrix $\mathrm{A} = \hat{\mathrm{S}}_1{}^{-\frac{1}{2}}\hat{\mathrm{S}}_2\,\hat{\mathrm{S}}_1^{-\frac{1}{2}}$ that can be diagonalized as usual:

$$\mathrm{Q}^\top \mathrm{A}\, \mathrm{Q} = \Delta$$

where Q is an orthogonal matrix and Δ is diagonal with decreasing eigenvalues $(\rho_1, \ldots, \rho_D)$, so that when defining the non-singular matrix $\mathrm{H} = \hat{\mathrm{S}}_1^{-\frac{1}{2}}\mathrm{Q}$:

$$\mathrm{H}^\top \hat{\mathrm{S}}_2\, \mathrm{H} = \Delta \text{ and } \mathrm{H}^\top \hat{\mathrm{S}}_1\, \mathrm{H} = \mathrm{I}_D$$

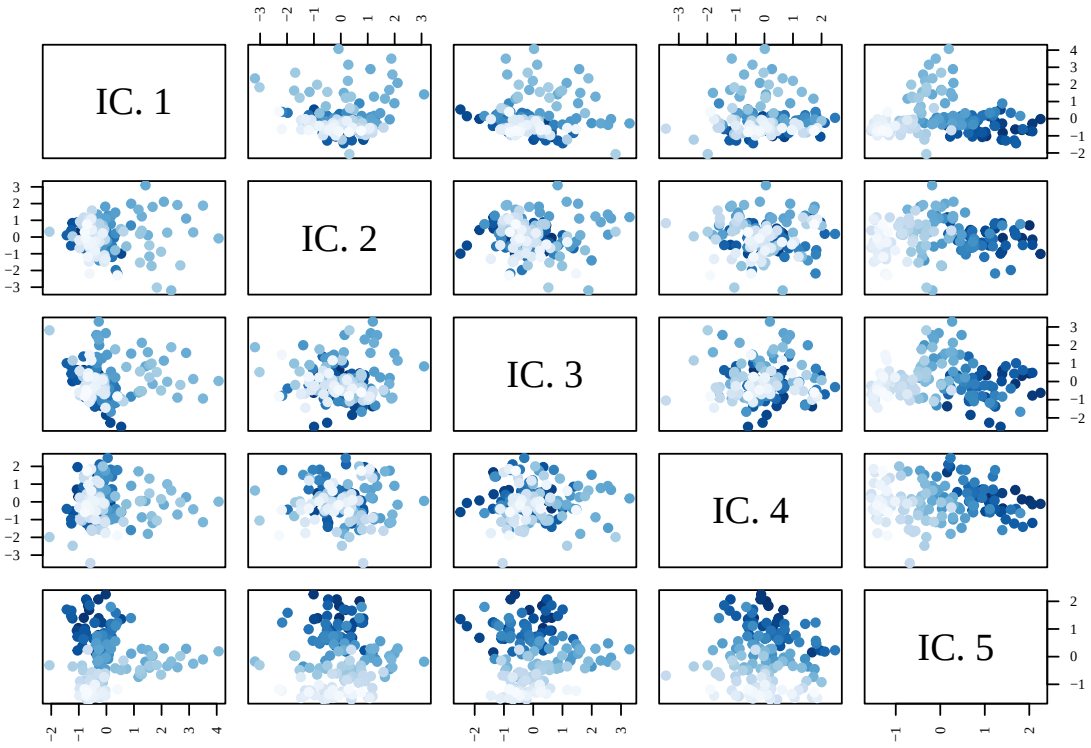
Remark. Since H is not orthogonal, this is not a proper diagonalization of matrices, but must be understood in terms of bilinear forms associated to the matrices $\hat{\mathrm{S}}_1$ and $\hat{\mathrm{S}}_2$: this simultaneous diagonalization can be reinterpreted as replacing the canonical inner product by that associated to $\hat{\mathrm{S}}_1$ (for which H is now orthogonal and $\hat{\mathrm{S}}_2$ is still positive-definite).

What is gained compared to the Mahalanobis transformation is that after performing the change of inner product according to a scatter operator, another scatter is diagonalized according to the new inner product.

Computing the Invariant Coordinates

The change of $\hat{\mathrm{S}}_1$-orthogonal basis:

$$\left(\ \mathbf{Z}_1 \ \middle|\ \ldots \ \middle|\ \mathbf{Z}_n\ \right) = \mathrm{H}^\top \left(\ \mathbf{X}_1 - \hat{m} \ \middle|\ \ldots \ \middle|\ \mathbf{X}_n - \hat{m}\ \right)$$



	IC.1	IC.2	IC.3	IC.4	IC.5
V_A	0.88	0.06	0.39	−0.26	−0.04
V_B	0.56	−0.55	0.27	−0.54	0.11
V_C	0.16	−0.40	0.13	−0.88	0.13
V_D	−0.06	−0.11	0.18	−0.37	0.90
V_E	−0.16	−0.14	−0.18	−0.12	0.95

Figure 1.6: Invariant Coordinates of the `BDDSegX` data set (volumes), and the correlation matrix of the original data with the transformed data.

is the first step of the method, and the coordinates of the $\mathbf{Z}_j, 1 \leq j \leq n$ are called 'Invariant Coordinates' because even though H is not unique (as eigenvectors are not unique), once chosen a way to compute one solution H, the transformed data set is affine invariant up to the sign of the $\mathbf{Z}_j, 1 \leq j \leq n$.

Selecting the Principal Components

According to the value of the spectrum $(\rho_1, \ldots, \rho_D)$ of $\hat{S}_1^{-1}\hat{S}_2$, we determine the components on which outliers must be looked for.

There are several ways to proceed, but the general idea is that since the spectrum is a measure of kurtosis of the distribution in various directions, we must keep the directions where the eigenvalues are substantially different from 1 (which means the distribution departs from an elliptical distributions in these directions). Eigenvalues larger than 1 are usually associated with invariant coordinates that help to detect small clusters of outliers,

$$\begin{bmatrix} \texttt{IC.1} & \texttt{IC.2} & \texttt{IC.3} & \texttt{IC.4} & \texttt{IC.5} \\ 1.41 & 1.23 & 1.11 & 1.01 & 0.65 \end{bmatrix}$$

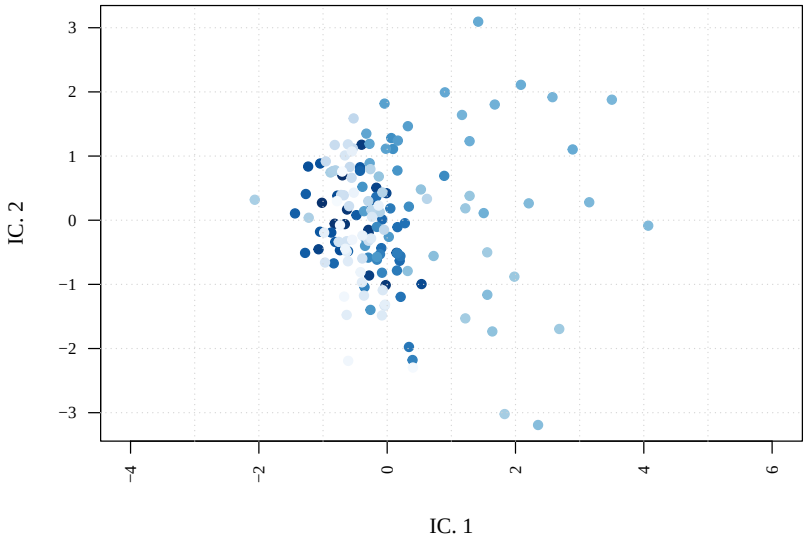


Figure 1.7: Eigenvalues of $\mathbb{V}[\mathbf{X}]^{-1}\mathbb{V}_4[\mathbf{X}]$ and projection on `IC.1` and `IC.2` of the Invariant Coordinates of the `BDDSegX` data set (volumes).

while eigenvalues smaller than 1 are associated with large clusters of outliers (see Theorem 3 from Tyler et al. 2009 for a theoretical justification of the method under a particular Gaussian mixture model).

The `ICSOutlier` R package by Archimbaud, Nordhausen, and Ruiz-Gazen 2018b suggests using normality tests such as D'Agostino's tests, which are based on estimation of the skewness and the kurtosis of the invariant components.

This is the step where the ICS method differs from the method presented in Subsection 1.1.3. Without projecting on a strict subspace, the next step will be identical to the Mahalanobis criterion for outlier detection.

Detecting the outliers

Just as for the Mahalanobis distance, outliers can be detected by computing the ICS distances (i.e. the norms of the transformed observations).

Example. Let us illustrate shortly the ICS procedure on the data set `BDDSegX`. On Figure 1.6, we can see that some observations look quite far from the main bulk of the data on the first component `IC.1` (see also the zoomed-in Figure 1.7). And this component to an eigenvalue larger than 1. Plotting the ICS distances calculated with `IC.1` leads to Figure 1.8 where outliers are all identified around 2010. The correlations of `IC.1` with original volumes are given on Figure 1.6. `IC.1` is positively correlated with the small volumes $\texttt{V}_\texttt{A}$ and $\texttt{V}_\texttt{B}$ meaning that because of the economic crisis of 2008, people have been more inclined to buy small cars.

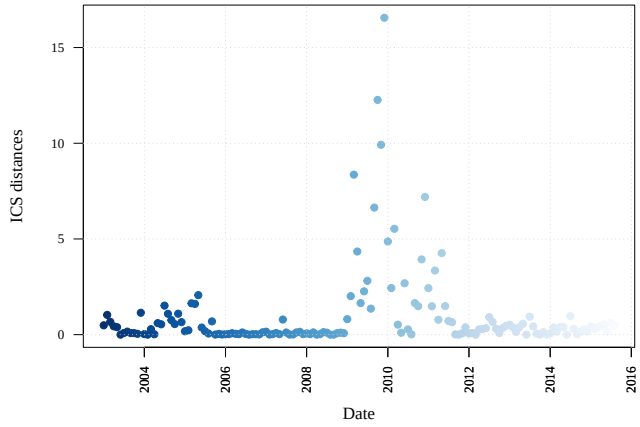


Figure 1.8: Outliers in selected Invariant Coordinates of the `BDDSegX` data set (volumes).

1.2 Compositional data analysis

This section summarises the basics of compositional linear algebra and data analysis, as presented in Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015.

1.2.1 Simplex of compositions, structure and isometries

Definition 1.5 (Simplex $\mathscr{S}^D$)**.** We define

$$\mathscr{S}^D = \left\{(x_1, \ldots, x_D)^\top \in {\mathbb{R}_+^*}^D, \sum_{1 \leq i \leq D} x_i = 1\right\}$$

as the simplex in dimension D, and its elements are called compositions.

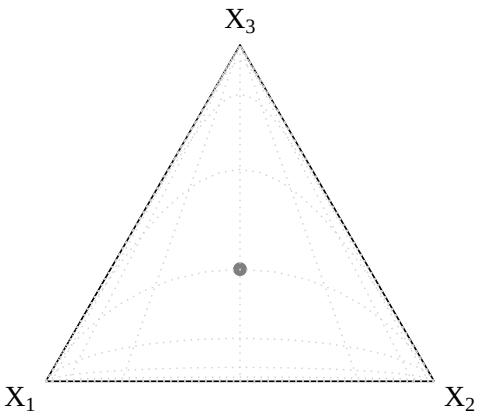


Figure 1.9: The triangle $\mathscr{S}^3$.

It is equipped with the following structure, for $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{S}^D$ and α in $\mathbb{R}$:

- $\mathbf{x} \oplus \mathbf{y} = \mathscr{C}\left((x_1 y_1, \ldots, x_D y_D)^\top\right)$ (Aitchison perturbation)
- If $\alpha \odot \mathbf{x} = \mathscr{C}\left({x_1}^\alpha, \ldots, {x_D}^\alpha)^\top\right)$ (Aitchison powering)

- $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{1 \le i < j \le D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)$ (Aitchison inner product)

where $\mathscr{C}$ is denotes the closure operator, which maps a $\mathbf{z} \in {\mathbb{R}_+^*}^D$ to the unique composition which is collinear to $\mathbf{z}$.

Due to the interpretation of compositions' coordinates as ratios of a whole, the classical metric is not the most relevant. That explains why we define these particular operations, and use a logarithmic transformation to measure distances.
The next theorem is fundamental because it provides two kinds of isometries between the simplex and Euclidian spaces of dimension $D-1$ that we will be using extensively; it was proved in Chapter 4 of the book by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

Theorem 1.6 (Structure & isometries)**.** *The simplex $\mathscr{S}^D$ with Aitchison's structure is an Euclidian space of dimension $D-1$, thus it is isometric to any Euclidian space of dimension $D-1$, for these particular isometries, respectively the 'centered logratio' and the 'isometric logratio':*

- $\mathrm{clr}: \begin{array}{ccc} \mathscr{S}^D & \to & \mathscr{H} = \ker\{\mathbf{x} \mapsto \sum_{i=1}^{i=D} x_i\} \\ \mathbf{x} & \mapsto & \left(\ln\left(\frac{x_1}{g(\mathbf{x})}\right), \dots, \ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right)^\top \end{array}$ *, where* $g: \mathbf{x} \mapsto \sqrt[D]{\prod_{i=1}^{i=D} x_i}$
- $\mathrm{ilr}_{\mathscr{B}}: \begin{array}{ccc} \mathscr{S}^D & \to & \mathbb{R}^{D-1} \\ \mathbf{x} = \sum_{i=1}^{i=D-1} {x_i}^* \mathbf{e}_i & \mapsto & \mathbf{x}^* = ({x_1}^*, \dots, {x_{D-1}}^*) \end{array}$ *, where* $\mathscr{B} = (\mathbf{e}_1, \dots, \mathbf{e}_{D-1})$ *is a given orthonormal basis of* $\mathscr{S}^D$*.*

The clr isometry is canonical in the sense that it does not require constructing an orthonormal basis (the inner product on $\mathscr{S}^D$ has been designed to match the clr transformation), but it is not surjective from $\mathscr{S}^D$ onto $\mathbb{R}^{D-1}$.
On the contrary, the $\mathrm{ilr}_{\mathscr{B}}$ isometries are indeed bijective from $\mathscr{S}^D$ to $\mathbb{R}^{D-1}$, but they are not canonical since there is no natural choice for an orthonormal basis of $\mathscr{S}^D$.

We must then verify that every construction we make using the $\mathrm{ilr}_{\mathscr{B}}$ transformation is independent of the basis $\mathscr{B}$ and also meaningful when considered directly on the simplex. The contrast matrices will be a helpful tool especially to prove this kind of results:

Definition 1.7 (Contrast matrix)**.** For a basis $\mathscr{B} = (\mathbf{e}_1, \dots, \mathbf{e}_{D-1})$ of the simplex $\mathscr{S}^D$, we define the contrast matrix $\mathrm{V}_{\mathscr{B}}$ as the matrix associated to the linear mapping clr from basis $\mathscr{B}$ to the canonical basis of $\mathbb{R}^D$, i.e. the matrix whose columns are the $\mathrm{clr}(\mathbf{e}_i), 1 \le i \le D-1$.

Proposition 1.8. *Let* V *be the contrast matrix of an orthonormal basis $\mathscr{B}$.*

- V *has dimensions* $D \times (D-1)$ *and its rank is* $D-1$.
- *We have:* $\mathrm{V}^\top \mathrm{V} = \mathrm{I}_{D-1}$ *and* $\mathrm{V}\mathrm{V}^\top = \mathrm{I}_D - \frac{1}{D} \mathbf{1}_D \mathbf{1}_D{}^\top$*, where* $\mathbf{1}_D = (1, \dots, 1)^\top \in \mathbb{R}^D$.
- *For all* $\mathbf{x}$ *in* $\mathscr{S}^D$*, we have:* $\mathrm{clr}(\mathbf{x}) = \mathrm{V}\,\mathrm{ilr}(\mathbf{x})$ *and* $\mathrm{ilr}(\mathbf{x}) = \mathrm{V}^\top \mathrm{clr}(\mathbf{x})$.

Note that there also exists another transformation called alr:

Definition 1.9 (alr transformation)**.**

$$\mathrm{alr}: \begin{array}{ccc} \mathscr{S}^D & \to & \mathbb{R}^{D-1} \\ \mathbf{x} & \mapsto & \left(\ln\left(\frac{x_1}{x_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right)^\top \end{array}$$

We define the $(D-1)\times D$-matrix F and the $D\times(D-1)$-matrix K by:

$$\mathrm{F}=\left(\begin{array}{c|c} \mathrm{I}_{D-1} & -\mathbf{1}_{D-1} \end{array}\right) \text{ and } \mathrm{K}=\begin{pmatrix} \mathrm{I}_{D-1}-\frac{1}{D}\mathbf{1}_{D-1}\mathbf{1}_{D-1}{}^{\top} \\ -\frac{1}{D}\mathbf{1}_{D-1}{}^{\top} \end{pmatrix}$$

Proposition 1.10.

- *We have:* $\mathrm{F}\,\mathrm{K}=\mathrm{I}_{D-1}$ *and* $\mathrm{K}\,\mathrm{F}=\mathrm{G}_D=\mathrm{I}_D-\frac{1}{D}\mathbf{1}_D\mathbf{1}_D{}^{\top}$.
- *For all* $\mathbf{x}$ *in* $\mathscr{S}^D$*, we have:* $\operatorname{alr}(\mathbf{x})=\mathrm{F}\operatorname{clr}(\mathbf{x})$ *and* $\operatorname{clr}(\mathbf{x})=\mathrm{K}\operatorname{alr}(\mathbf{x})$.

The alr transformation is a linear mapping, but is not an isometry, and is not symmetric in the coordinates. For these reasons, we will not be able to use it in this report. However, it important to note that it is the easiest one to compute.

1.2.2 Matrix-vector product on the simplex

The Invariant Coordinate Selection method is based on a linear transformation of the data set, so for our purposes we need to study the basic of linear algebra on the simplex.
As usual, in order to get a matrix description of linear mappings on a vector space we first need to define a matrix-vector product.

But in the case of the simplex, this leads naturally to defining two matrix-vector products, since when adapting the classical product to simplex operations, we lose the commutative property of the multiplication. In this report, we will only use the right-side product, which is defined for a $D\times D$-matrix A and a vector $\mathbf{x}\in\mathscr{S}^D$ by:

$$\mathrm{A}\boxdot\mathbf{x}=\mathscr{C}\left(\left(\prod_{j=1}^{j=D}x_j^{a_{1,j}},\ldots,\prod_{j=1}^{j=D}x_j^{a_{D,j}}\right)^{\top}\right)$$

The relevance of this new product is given by the following proposition, which is Proposition 11.3.9 in Egozcue et al. 2011.

Proposition 1.11. *For every* $\mathrm{A}\in\mathbb{R}^{D\times D}$ *and* $\mathbf{x}\in\mathscr{S}^D$ *:*

$$\operatorname{clr}(\mathrm{A}\boxdot\mathbf{x})=\mathrm{A}\operatorname{clr}(\mathbf{x})$$

This compatibility with the clr isometry explains why the compositional matrix-vector product $\boxdot$ will be useful to describe the algebra $\mathscr{L}(\mathscr{S}^D)$ of linear mappings of the simplex $\mathscr{S}^D$, provided that it is linear (which, as explained in Egozcue et al. 2011 is not the case for every matrix $\mathrm{A}\in\mathbb{R}^{D\times D}$). We will come back to this in section 2.1.

1.2.3 Statistics on the simplex

Through an ilr isometry, we can transport the Borel σ-field of $\mathbb{R}^{D-1}$ and the Lebesgue measure to $\mathscr{S}^D$, and thus construct random variables on the simplex (their distribution functions are defined as the push-forward measure of the distribution of $\mathbf{X}^*$ on $\mathbb{R}^{D-1}$). Then we have to show that the constructed probability space and the random variables do not depend on the choice of isometry.

Centre & geometrical mean

Definition 1.12 (Centre)**.** The centre of a random variable $\mathbf{X}$ on $\mathscr{S}^D$ is defined as:

$$\mathbf{g}[\mathbf{X}] = \mathrm{clr}^{-1}\big(\mathbb{E}[\mathrm{clr}(\mathbf{X})]\big) = \mathrm{ilr}_{\mathscr{B}}^{-1}\big(\mathbb{E}[\mathrm{ilr}_{\mathscr{B}}(\mathbf{X})]\big)$$

for any orthonormal basis $\mathscr{B}$ of $\mathscr{S}^D$, provided that it verifies $\mathbb{E}[\|\mathrm{clr}(\mathbf{X})\|] < \infty$.

A consistent estimator of $\mathbf{g}[\mathbf{X}]$ is given by the geometric mean :

$$\hat{\mathbf{g}} = \tfrac{1}{n} \odot (\mathbf{X}_1 \oplus \cdots \oplus \mathbf{X}_n)$$

Notions of variance on the simplex

In this subsection, we make a summary of the common notions of variance that can be defined on the simplex.

The first object (defined in particular in the R package `compositions` by van den Boogaart, Tolosana-Delgado, and Bren 2021) will be the most useful when applying the ICS method to compositional data:

- The (co-)variance clr-matrix:

$$\mathbb{V}[\mathbf{X}] = \mathbb{V}[\mathrm{clr}(\mathbf{X})] = \mathrm{V}_{\mathscr{B}}\, \mathbb{V}[\mathrm{ilr}_{\mathscr{B}}(\mathbf{X})]\, \mathrm{V}_{\mathscr{B}}^{\top} \text{ for any orthonormal basis } \mathscr{B}$$

The following three objects are defined by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015. We considered using them when adapting ICS to compositional data, but due to their lack of interpretation as linear mappings we preferred the (co-)variance clr-matrix.

- Aitchison's variation matrix:

$$\mathrm{T} = \left(\mathbb{V}\Big[\ln\left(\frac{X_i}{X_j}\right)\Big]\right)_{1 \le i,j \le D-1}$$

- The variability according to a direction $\mathbf{z} \in \mathscr{S}^D$:

$$\mathrm{var}[\mathbf{X}, \mathbf{z}] = \mathbb{E}\big[\|\mathbf{X} \ominus \mathbf{z}\|_a^{\,2}\big]$$

- The total variance:

$$\mathrm{totvar}[\mathbf{X}] = \min_{\mathbf{z} \in \mathscr{S}^D} \mathrm{var}[\mathbf{X}, \mathbf{z}] = \mathbb{E}\big[\|\mathbf{X} \ominus \mathbf{g}(\mathbf{X})\|_a^{\,2}\big]$$

Chapter 2

Algebra and statistics embedded in the simplex

In the previous section, we reminded helpful algebra and statistics tools to study compositional data which often made use of the isometric logratio transformations.
In this section, we will develop a more intrinsic point of view on compositional linear algebra and statistics, in order to prepare for Chapter 3, where we will suggest an adapted ICS method for compositional data.

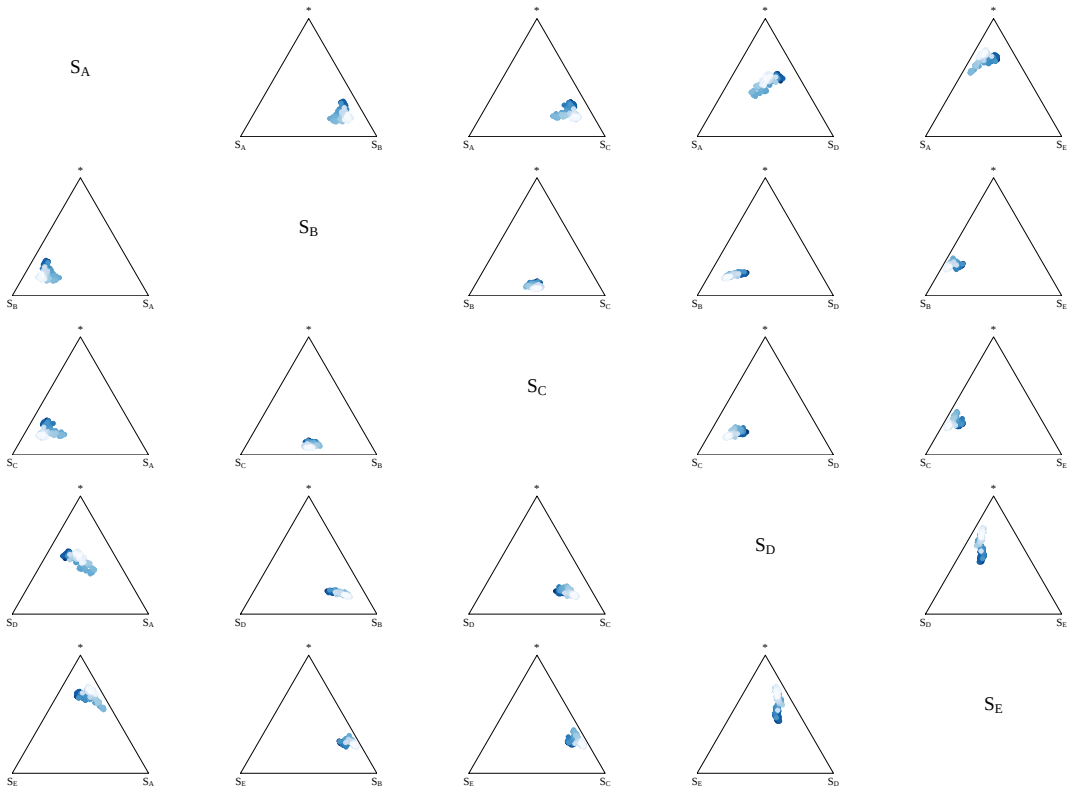


Figure 2.1: Pairwise ternary diagrams of the `BDDSegX` data set (compositions).

From now on, we transform the volumes from the data set `BDDSegX` in compositions denoted by $\mathtt{S_A}, \mathtt{S_B}, \mathtt{S_C}, \mathtt{S_D}$ and $\mathtt{S_E}$ by using the closure operator. Note that in the case of $D = 3$, it is possible to plot compositional data by using a ternary representation. When $D > 3$, as in the

example `BDDSegX` where $D = 5$, we can plot a matrix of ternary diagrams by considering in turn two coordinates and aggregating the other coordinates of the composition (see Figure 2.1).

2.1 Linear algebra in the simplex

Since $\mathcal{S}^D$ is canonically isometric to the clr-space $\mathcal{H}$, we can naturally identify linear mappings on the simplex and on $\mathcal{H}$, associating to a mapping $u \in \mathcal{L}(\mathcal{S}^D)$ the mapping:

$$\Psi(u): \begin{array}{ccc} \mathcal{H} & \to & \mathcal{H} \\ \mathbf{y} & \mapsto & \mathrm{clr}(u(\mathrm{clr}^{-1}(\mathbf{y}))) \end{array} \in \mathcal{L}(\mathcal{H}) \tag{2.1}$$

To get a matrix description of the algebra $\mathcal{L}(\mathcal{H})$, it is more convenient to see it as embedded into $\mathcal{L}(\mathbb{R}^D)$, through the isomorphism:

$$\begin{array}{ccc} \left\{ v \in \mathcal{L}(\mathbb{R}^D) \middle| v(\mathcal{H}) \subseteq \mathcal{H}, v(\mathbf{1}_D) = 0 \right\} & \to & \mathcal{L}(\mathcal{H}) \\ v & \mapsto & v|_{\mathcal{H}} \end{array}$$

because then the conditions to belong to $\mathcal{L}(\mathcal{H})$ can be translated into matrix conditions (called 'zero-sum property' in Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

These intuitions can be summed up into the following theorem:

Theorem 2.1 (Matrix description of linear mappings on the simplex)**.**

- *The algebra $\mathcal{L}(\mathcal{S}^D)$ of linear mappings on $\mathcal{S}^D$ is isomorphic to the algebra:*

$$\mathcal{M} = \left\{ \mathrm{A} \in \mathbb{R}^{D \times D} \middle| \mathbf{1}_D{}^\top \mathrm{A} = 0 \text{ and } \mathrm{A}\mathbf{1}_D = 0 \right\}$$

  of $D \times D$-matrices verifying the zero-sum property (row-wise and column-wise), called 'clr-matrices', on which the identity element with respect to the multiplication is the matrix:

$$\mathrm{G}_D = \mathrm{I}_D - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D{}^\top$$

  and the multiplicative inverse of a rank $D-1$ matrix $A \in \mathcal{M}$ is:

$$A^{-1} = (\mathrm{A} + \mathbf{1}_D\mathbf{1}_D{}^\top)^{-1} - \mathbf{1}_D\mathbf{1}_D{}^\top = \mathrm{ilr}_{\mathcal{B}}^{-1}\left(\mathrm{ilr}_{\mathcal{B}}(\mathrm{A})^{-1}\right) \in \mathcal{M}$$

  for any basis $\mathcal{B}$.

- *The following mapping is an isomorphism:*

$$\Phi: \begin{array}{ccc} \mathcal{M} & \to & \mathcal{L}(\mathcal{S}^D) \\ \mathrm{A} & \mapsto & \left(\mathbf{x} \mapsto \mathrm{A} \boxdot \mathbf{x}\right) \end{array}$$

  and its inverse is:

$$\Phi^{-1}: \begin{array}{ccc} \mathcal{L}(\mathcal{S}^D) & \to & \mathcal{M} \\ u & \mapsto & \left(\mathrm{clr}(u(\varepsilon_1)), \ldots, \mathrm{clr}(u(\varepsilon_D))\right) \end{array}$$

  where $(\varepsilon_1, \ldots, \varepsilon_D)$ is the generating set of $\mathcal{S}^D$ obtained when applying clr^{-1} to the canonical basis of $\mathbb{R}^D$.

Proof.

- We prove that Φ is an isomorphism by writing $\Phi = \Phi_2 \circ \Phi_1$ thanks to the compatibility of the matrix-vector product on the simplex from Proposition 1.11, where $\Phi_2 = \Psi^{-1}$ (as defined in Equation (2.1)) is clearly an algebra isomorphism and:

$$\Phi_1 : \begin{array}{ccc} \mathscr{M} & \to & \mathscr{L}(\mathscr{H}) \\ \mathrm{A} & \mapsto & \left(\mathbf{x} \mapsto \mathrm{A}\,\mathbf{x}\right) \end{array}$$

  is well defined because if $\mathrm{A} \in \mathscr{M}$ and $u = \Phi_1(\mathrm{A}) \in \mathscr{L}(\mathbb{R}^D)$, we have $\mathrm{A}\,\mathbf{1}_D = \mathbf{0}_D$ so $u(\mathbf{1}_D) = \mathbf{0}_D$ and ${\mathbf{1}_D}^\top \mathrm{A} = \mathbf{0}_D$ so for every $\mathbf{x} \in \mathbb{R}^D$, $\langle \mathbf{1}_D, u(\mathbf{x}) \rangle = 0$ i.e. $u(\mathbf{x}) \in {\mathbf{1}_D}^\perp = \mathscr{H}$ so $\Phi_1(\mathrm{A}) \in \mathscr{L}(\mathscr{H})$. Then Φ_1 is the restriction/corestriction of an injective algebra homomorphism and the subspaces $\mathscr{M}$ and $\mathscr{L}(\mathscr{H})$ have the same dimension $(D-1)^2$.

- To prove the expression of Φ^{-1}, it is sufficient to write it as $\Phi_1^{-1} \circ \Phi_2^{-1}$.

- We have $\Phi^{-1}\left(\mathrm{Id}_{\mathscr{S}^D}\right) = \mathrm{G}_D$.

- If $\mathrm{A} \in \mathscr{M}$ is a rank $D-1$ clr-matrix, we consider the linear mapping $v \in \mathscr{L}(\mathbb{R}^D)$ associated with $\mathrm{A} + \mathbf{1}_D {\mathbf{1}_D}^\top$ which stabilizes $\mathscr{H}$ and $\mathbb{R}\mathbf{1}_D$, is invertible and verifies $v_{|\mathscr{H}} = \Phi_1(\mathrm{A})$ and $v_{|\mathbb{R}\mathbf{1}_D} = \mathrm{Id}_1$, from where we deduce the formula. □

Remarks.

- The isomorphisms from the theorem are canonical, in the sense that they do not rely on a choice of orthonormal basis of $\mathscr{S}^D$.

- This theorem can be seen as a particular case of theorem 11.3.10 in Egozcue et al. 2011 in which the linear mappings are endomorphisms. But it is more precise because it gives algebras isomorphisms and a formula to compute the inverse of a rank $D-1$ clr-matrix.

- One could as easily prove a formula such as, for every $\mathrm{A} \in \mathscr{M}$:

$$\det(\mathrm{A}^*) = \det\left(\Phi(\mathrm{A})\right) = \det\left(\mathrm{A} + \mathbf{1}_D {\mathbf{1}_D}^\top\right)$$

2.2 Whitening data in the simplex

Thanks to the formula of the inverse of a rank $D-1$ clr-matrix $\mathrm{A} \in \mathscr{M}$:

$$\mathrm{A}^{-1} = (\mathrm{A} + \mathbf{1}_D {\mathbf{1}_D}^\top)^{-1} - \mathbf{1}_D {\mathbf{1}_D}^\top$$

it is possible to compute the inverse square root of the (co-)variance clr-matrix $\mathbb{V}[\mathbf{X}]^{-\frac{1}{2}}$, in order to whiten a compositional random variable admitting the first two moments without computing any isometric logratio transformation:

$$\mathbf{Z} = \mathbb{V}[\mathbf{X}]^{-\frac{1}{2}} \boxdot \left(\mathbf{X} \ominus \mathbf{g}[\mathbf{X}]\right) = \mathrm{clr}^{-1}\left(\mathbb{V}[\mathrm{clr}(\mathbf{X})]^{-\frac{1}{2}}\left(\mathrm{clr}(\mathbf{X}) - \mathbb{E}[\mathrm{clr}(\mathbf{X})]\right)\right)$$

with $\mathbb{V}[\mathbf{Z}] = \mathrm{G}_D$ and $\mathbf{g}[\mathbf{Z}] = \frac{1}{D}\mathbf{1}_D$.

Example. Figure 2.2 illustrates the impact of whitening data in the simplex on the `BDDSegX` data set. We consider $\mathrm{S}_\mathtt{A}$ and $\mathrm{S}_\mathtt{B}$ and aggregate the other coordinates in order to get compositions in 3 dimensions. The mean (geometric mean) of the data points is plotted as a small circle while the scatter is represented by an 'ellipse' transformed back to the simplex.

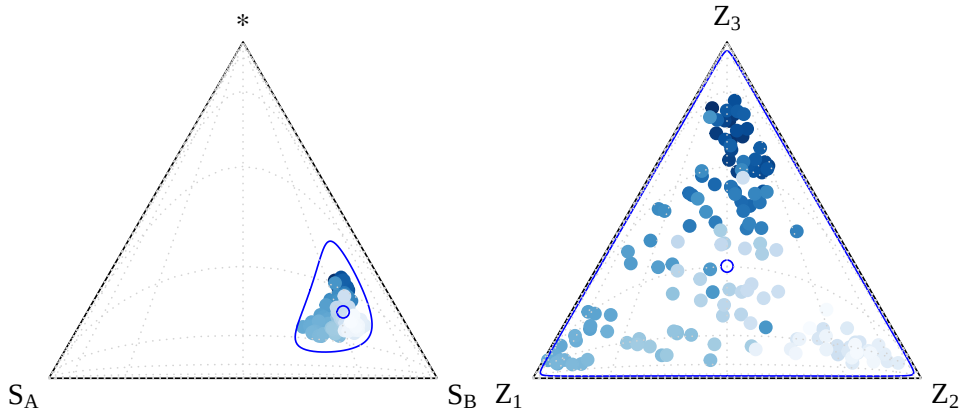


Figure 2.2: Original (*left*) and whitened (*right*) subcompositions of `BDDSegX` data set (compositions).

2.3 Elliptical distributions on the simplex

In this subsection, we will define elliptical distributions on the simplex, generalizing the definition of the Gaussian distribution on the simplex in Theorem 6.4 by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015.

Example. On Figure 2.3, we generated $n = 100$ observations following a Gaussian distribution in $\mathscr{S}^3$ and plot the associated ternary diagram with an 'ellipse' contour on the left plot. The same data are plotted on the right part of the figure after whitening.

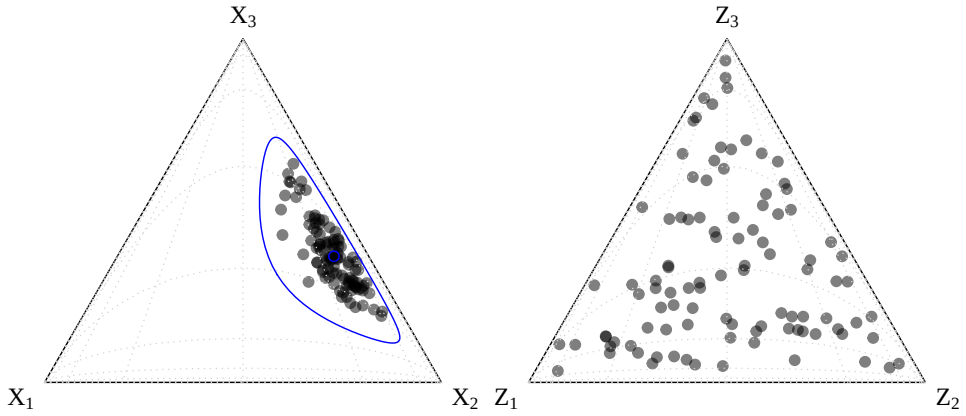


Figure 2.3: Gaussian data in $\mathscr{S}^3$, before (left plot) and after whitening (right plot).

Proposition 2.2. *Consider two orthonormal bases $\mathscr{B}_1$ and $\mathscr{B}_2$ of the simplex $\mathscr{S}^D$ with associated contrast matrices V_1 and V_2 respectively and a random composition $\mathbf{X}$. For i in $\{1,2\}$, let $\mathbf{X}_i{}^* = \mathrm{ilr}_i(\mathbf{X})$ represent its orthonormal coordinates according to $\mathscr{B}_i$. If $\mathbf{X}_1{}^*$ follows a (μ_1, Σ_1)-elliptical distribution, then $\mathbf{X}_2{}^*$ follows a (μ_2, Σ_2)-elliptical distribution, where $\mu_2 = V_2{}^\top V_1 \mu_1$ and $\Sigma_2 = V_1{}^\top V_2 \Sigma_1 V_2{}^\top V_1$.*

Proof. As we know that $\mathrm{ilr}_2(\mathbf{x}) = {V_2}^\top V_1 \mathrm{ilr}_1(\mathbf{x})$, it is enough to prove that ${V_2}^\top V_1$ is a $(D-1) \times (D-1)$-orthogonal matrix, because then we only need to perform an orthonormal change of basis in the expression of the distribution of $\mathbf{X}$ to obtain the result.
Let us verify that it is indeed true:

$$\begin{aligned}\left({V_2}^\top V_1\right)^\top {V_2}^\top V_1 = {V_1}^\top V_2 {V_2}^\top V_1 &= {V_1}^\top \left(I_D - \tfrac{1}{D}\mathbf{1}_D {\mathbf{1}_D}^\top\right) V_1 \\ &= {V_1}^\top V_1 - \tfrac{1}{D}{V_1}^\top \mathbf{1}_D {\mathbf{1}_D}^\top V_1 = I_{D-1}\end{aligned}$$

since ${V_1}^\top \mathbf{1}_D$ is the vector of the sums of columns of V_1 and is equal to $\mathbf{0}_{D-1}$. □

Remark. In this proof, we use the fact that the matrix ${V_2}^\top V_1$ is a $(D-1) \times (D-1)$-orthogonal matrix, which is natural since a change of contrast matrix corresponds to a different choice of orthonormal basis of $\mathscr{S}^D$ (which is isometric to $\mathbb{R}^{D-1}$), but also very helpful to prove this kind of ilr-independence results.

This allows us to define the elliptical distributions as push-forward measures:

Definition 2.3 (Elliptical distributions)**.** If $\mu \in \mathscr{S}^D$ and $\Sigma \in \mathscr{M}$, a random composition $\mathbf{X}$ on $\mathscr{S}^D$ is said to follow a (μ, Σ)-elliptical distribution if the random vector $\mathrm{ilr}(\mathbf{X})$ (where ilr is any isometric logratio transformation) follows a (μ^*, Σ^*)-elliptical distribution on $\mathbb{R}^{D-1}$.

In particular the Gaussian distribution on the simplex $\mathscr{N}_{\mathscr{S}}(\mu, \Sigma)$ is defined as the push-forward measure by any ilr^{-1} of the Gaussian distribution $\mathscr{N}(\mu^*, \Sigma^*)$.

Chapter 3

Outlier detection using ICS for compositional data

3.1 ICS on the ilr-space

Given what has been said earlier about outlier detection and compositional data, one way to treat the subject would be to send isometrically the compositional data set to the ilr-space and apply ICS under the elliptical mixture model to ilr($\mathbf{X}$).

Applying results like Theorem 3 from the article by Tyler et al. 2009, we would know that under some assumptions the outliers would be highlighted by an orthogonal projection on the first or the last ilr-coordinates.

However, many questions arise from this approach:

- What is the dependence of this model on the choice of isometry ilr?
- Why use an ilr transformation and not the clr transformation?
- How to interpret (if it is unique) the ICS transformation back in the simplex?
- Are elliptical distributions on the simplex and the results of the ICS method meaningful and can they model real data?

The last question can already be partially answered, since elliptical distributions are a generalization of the normal distribution on the simplex, for which the central limit theorem (Theorem 6.24 in Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015) still applies.
This suggests that the family of distributions with elliptical symmetry is still a worthy candidate to model real compositional data.

First, let us apply the elliptical mixture model after the isometric logratio transformation, and define an intrinsic model on the simplex.

3.2 ICS on the simplex: what is possible – what is not

3.2.1 Elliptical mixture models on the simplex

Firstly, if the distribution function $F_{\mathbf{X}}$ of $\mathbf{X}$ verifies:

$$F_{\mathbf{X}^*} = (1-\varepsilon)F_0^* + \sum_{k=1}^{k=q} \varepsilon_k F_k^*$$

in the ilr-space then, by definition of distribution functions on $\mathscr{S}^D$ (as push-forward measures using the measurable function ilr^{-1}):

$$F_{\mathbf{X}} = (1-\varepsilon)F_0 + \sum_{k=1}^{k=q} \varepsilon_k F_k$$

where, for $0 \leq k \leq q$ we define $F_k = F_k^* \circ \mathrm{ilr}$ (which does not rely on ilr).

Now, let us apply the Invariant Coordinate Selection method on the ilr coordinates and try to get rid of the dependence on the isometric logratio transformation.

3.2.2 ICS from the ilr-space back to the simplex

As we saw earlier when we whitened compositional data, the ilr-linear transformation corresponds, back in the simplex, to:

$$\mathbf{Z} = \mathrm{H}^\top \boxdot \big(\mathbf{X} \ominus \mathbf{g}[\mathbf{X}]\big)$$

and $\mathrm{H}^* = {\mathrm{S}_1^*}^{-\frac{1}{2}}\mathrm{Q}^*$ is associated with $\mathrm{H} = {\mathrm{S}_1}^{-\frac{1}{2}}\mathrm{Q}$ (where the inverse square root exists because the linear mapping associated with the clr-matrix S_1 is positive-definite). So applying $\mathrm{H}^\top$ means first applying S_1 (i.e. changing the Aitchison inner product) and then applying the orthogonal matrix Q: the interpretation of the ICS transformation remains the same on the simplex.

What we need to keep in mind is that the ICS method is based on the interpretation of the eigenvalues of ${\mathrm{S}_1}^{-1}\mathrm{S}_2$ as a measure of kurtosis according to each axis, and that a high kurtosis according to a direction suggests the existence of outliers in that particular direction: one way or another, we must choose H^* so that this interpretation is not lost when isometrically coming back to the simplex.

Here, one could imagine an alternative way to find an orthogonal matrix Q by diagonalizing $\mathrm{A} = {\mathrm{S}_1}^{-\frac{1}{2}}\mathrm{S}_2\,{\mathrm{S}_1}^{-\frac{1}{2}}$ directly in the clr-space. Naturally, the eigenvalues (omitting the 0 associated to the zero-sum property) would be preserved, but the eigenvectors would form a basis of $\mathbb{R}^D$ instead of the clr-space $\mathscr{H}$. However, when diagonalizing in an ilr-space then computing $\mathrm{Q} = \mathrm{V}\mathrm{Q}^*\mathrm{V}^\top$, the matrix Q does not diagonalize A (it is not even orthogonal as a $D \times D$-matrix).

Example. Before analyzing in detail the dependence of the method on the choice of ilr transformation (which will be the object of the next section), we present an application of this compositional ICS method to a data set of size $n = 150$ generated according to a

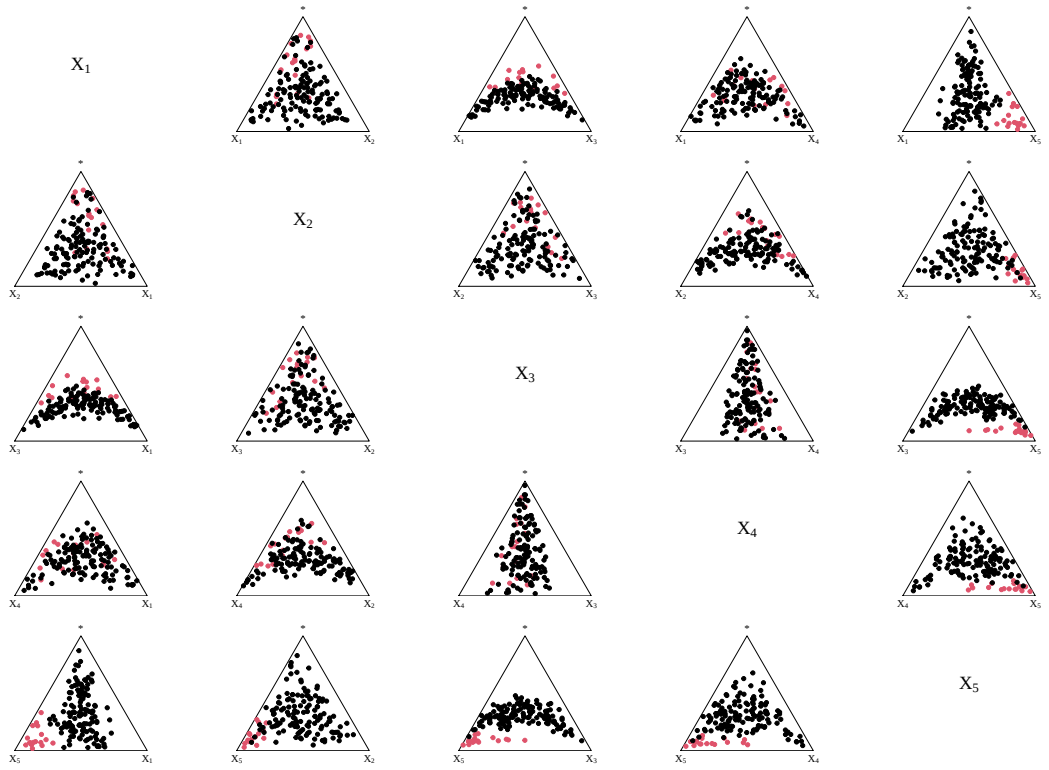


Figure 3.1: Data generating according to a Gaussian mixture model in $\mathscr{S}^5$ (*top*), and its Invariant Coordinates (*bottom*).

$$\begin{bmatrix} & \texttt{IC.1} & \texttt{IC.2} & \texttt{IC.3} & \texttt{IC.4} & \texttt{IC.5} \\ \texttt{X}_1 & 0.58 & -0.16 & -0.29 & 0.30 & -0.12 \\ \texttt{X}_2 & 0.39 & 0.39 & -0.48 & -0.34 & 0.57 \\ \texttt{X}_3 & -0.09 & 0.29 & 0.54 & -0.60 & 0.08 \\ \texttt{X}_4 & -0.15 & 0.19 & 0.62 & -0.22 & -0.35 \\ \texttt{X}_5 & -0.23 & -0.20 & -0.23 & 0.58 & 0.13 \end{bmatrix}$$

Figure 3.2: Covariance between log(**X**) and log(**Z**).

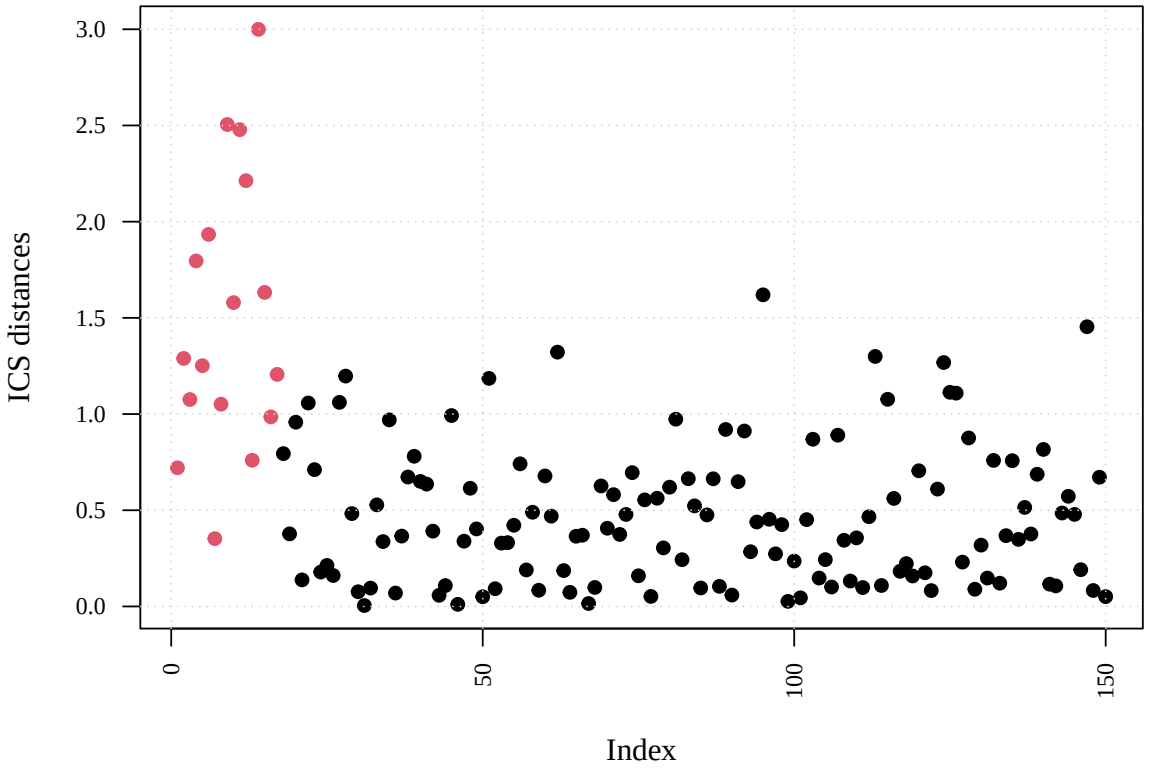


Figure 3.3: Outliers in selected Invariant Coordinates of the data set generated according to a Gaussian mixture model.

compositional Gaussian mixture model.

We first compute the matrix H and the Invariant Coordinates as we described at the top of this paragraph (see Figure 3.1).

Then, we select `IC.1` and regroup the other components, and we see that the outliers are best identified using this particular amalgamation.

Finally we plot the Aitchison norms of the selected components on Figure 3.3.

3.2.3 The effects of changing the contrast matrix

Let us suppose that we have two transformations ilr_1 and ilr_2, whose respective contrast matrices are V_1 and V_2.
Then for each $i \in \{1,2\}$, we define an orthogonal matrix Q_i^{*i} diagonalizing A^{*i} and note its associated clr-matrix $Q_i \in \mathscr{M}$.
The aim of the subsection is to connect Q_1 to Q_2. Let us write the equalities we know this far:

$$A = V_1 A^{*1} {V_1}^\top \qquad A = V_2 A^{*2} {V_2}^\top \tag{3.1}$$

$$Q_1 = V_1 Q_1^{*1} {V_1}^\top \qquad Q_1 = V_2 Q_1^{*2} {V_2}^\top \tag{3.2}$$

$$Q_2 = V_1 Q_2^{*1} {V_1}^\top \qquad Q_2 = V_2 Q_2^{*2} {V_2}^\top \tag{3.3}$$

$${Q_1^{*1}}^\top A^{*1} Q_1^{*1} = \Delta^* \qquad {Q_2^{*2}}^\top A^{*2} Q_2^{*2} = \Delta^* \tag{3.4}$$

The equations of line (3.4) use the fact that the positive semi-definite matrices A, A_1^{*1}, A_2^{*2} have the same non-zero eigenvalues, which justifies the notation Δ^* (without reference to ilr_1 or ilr_2) for the diagonal matrix contain them in decreasing order.

By equalizing the two different expressions of A on line (3.1), we get:

$$A^{*2} = V_{1,2} A^{*1} V_{1,2}^\top$$

where $V_{1,2} = {V_2}^\top V_1$ is a $(D-1) \times (D-1)$-orthogonal matrix, as stated by the remark after Proposition 2.2 on elliptical distributions on the simplex. We do the same with lines (3.2) and (3.3) and we obtain two similar equalities connecting Q_i^{*2} to Q_i^{*1} for $i \in \{1,2\}$.

Let us suppose that $Q_1 = Q_2$. Then:

$$Q_2^{*2} = V_{1,2} Q_1^{*1} V_{1,2}^\top$$

so when replacing Q_2^{*2} by its new expression in the right side equation of line (3.4):

$$V_{1,2} {Q_1^{*1}}^\top A^{*1} Q_1^{*1} V_{1,2}^\top$$

i.e. using this time the left side equation of line (3.4):

$$V_{1,2} \Delta^* V_{1,2}^\top = \Delta^*$$

Since this last equality is generally not true, we can deduce that in general:

$$Q_1 \neq Q_2$$

The intuition is that diagonalizing the matrix A^{*i} sums up to choosing an orthonormal basis of the simplex, just as picking a contrast matrix. So when deducing Q_i from Q_i^{*i} we should not use different contrast matrices, but the same one which we will denote V_0.
Now let us redefine Q_1 and Q_2 with this adjustment:

$$Q_1 = V_1 Q_1^{*1} {V_0}^\top \qquad Q_2 = V_2 Q_2^{*2} {V_0}^\top$$

so that:

$$Q_2^{*2} = V_{1,2} Q_1^{*1}$$

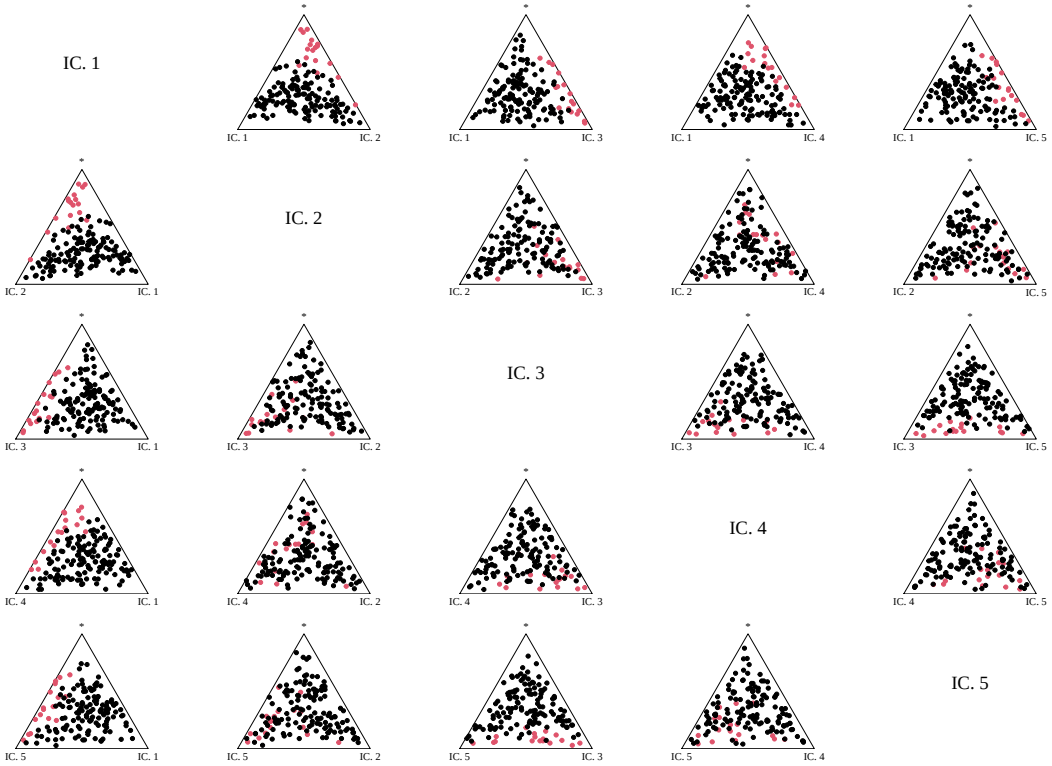
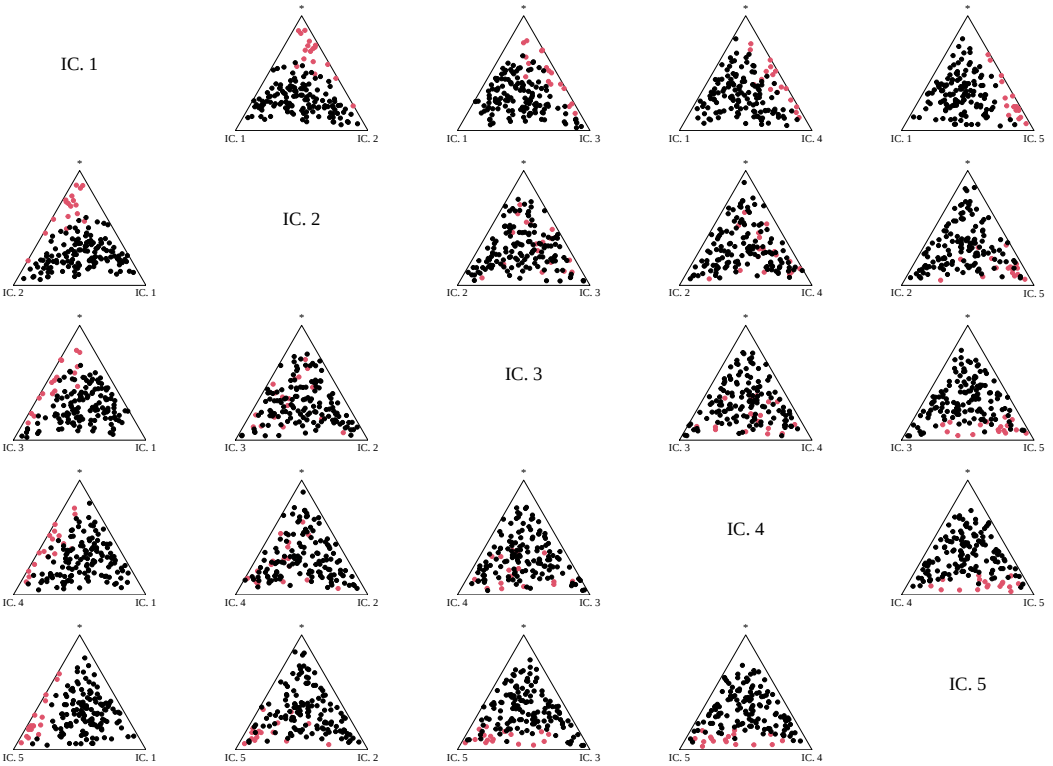


Figure 3.4: Invariant Coordinates of the data set generating according to a Gaussian mixture model in $\mathscr{S}^5$, with two different contrast matrices.



Since the eigenvectors are never unique, we still have in general:

$$Q_1 \neq Q_2$$

But let us assume that they are unique (which is realistic since under our continuous models the probability that two eigenvalues are equal is zero, so each eigenvector is unique up to its sign, which is still problematic but we could design a process that sets the signs). Then, by going up the previous reasoning we could deduce that $Q_1 = Q_2$.

Example. Figure 3.4 contains scatter plots of the Invariant Coordinates of the compositional mixture data set, with two different contrast matrices. We notice that because the signs were not set, only the plot of the amalgamation $(\texttt{IC.1}, \texttt{IC.2}, *)$ is identical between both subfigures. Because of the contrast matrices, one small change (for instance a vector turned into its inverse) on the Q^* matrix can induce more important changes on Q.

Globally, we conclude that this adaptation of the ICS method (using ilr) is dependent on the choice of ilr for the same reason that in ICS, the expression 'Invariant Coordinates' and their definition are slightly improper: because of the non-uniqueness of the eigenvectors.
A way to avoid this problem would be to choose a privileged contrast matrix, for instance picking an orthonormal basis that diagonalizes A in the clr-space $\mathscr{H}$.

3.3 Application to automotive market data & interpretations

To conclude this chapter, let us apply our suggestion of adapted ICS method for compositional data to the `BDDSegX` data set on the automotive market.

First, the Invariant Coordinates are computed, and the covariances between log-coordinates of the original and transformed data set can be found on Figure 3.6. It is more difficult to interpret the Invariant Coordinates since we did not define a good notion of correlation on the simplex, and that the covariances are less meaningful because the total variances of the coordinates differ from one another.

The Invariant Coordinates are plotted on Figure 3.5, and what is interesting is that, even without selection, we can see outliers appear, as well as groups (we remind that shades of blue are in chronological order, from darkest to lightest).

But if we try to select some coordinates, the outliers appear neither on the first nor on the last Invariant Coordinates, but rather on the fourth coordinate.
Figure 3.7 is a scatter plot of the ICS distances, i.e. the Aitchison norms of the amalgamation isolating `IC.4` from the other components. We clearly see that between 2008 and 2010 (which are the year of the financial crisis in Europe) is were we get the majority of outliers.

There is a noteworthy difference with the Figure 1.8 (which plots the ICS distances, but for the data set in volumes): after 2014, new outliers appear. This suggests a reinterpretation of the outliers as majors booms of European automotive market, which happened around 2008 and around 2015.

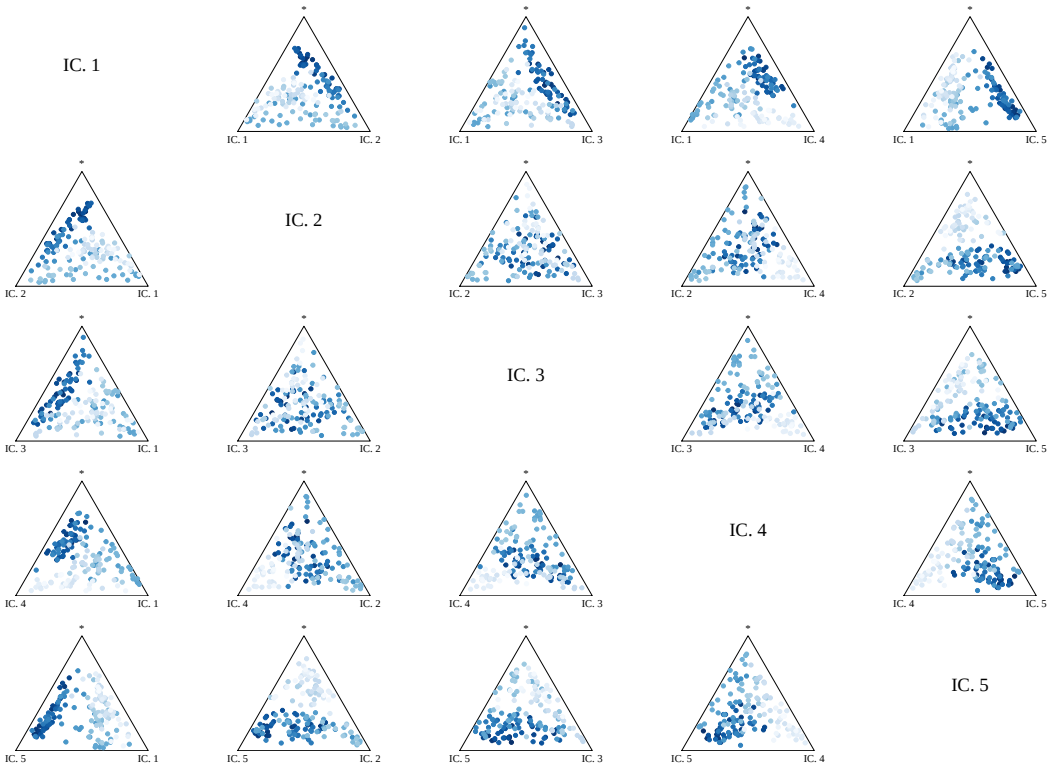


Figure 3.5: Invariant Coordinates of the `BDDSegX` data set (compositions).

$$
\begin{bmatrix}
 & \texttt{IC.1} & \texttt{IC.2} & \texttt{IC.3} & \texttt{IC.4} & \texttt{IC.5} \\
\texttt{S}_\texttt{A} & 0.14 & 0.11 & -0.12 & -0.17 & -0.08 \\
\texttt{S}_\texttt{B} & 0.02 & 0.01 & -0.05 & 0.02 & -0.02 \\
\texttt{S}_\texttt{C} & 0.01 & -0.03 & 0.04 & 0.04 & -0.04 \\
\texttt{S}_\texttt{D} & -0.11 & -0.03 & 0.09 & -0.06 & 0.17 \\
\texttt{S}_\texttt{E} & -0.32 & 0.05 & 0.17 & -0.03 & 0.30
\end{bmatrix}
$$

Figure 3.6: Covariance between log(**X**) and log(**Z**).

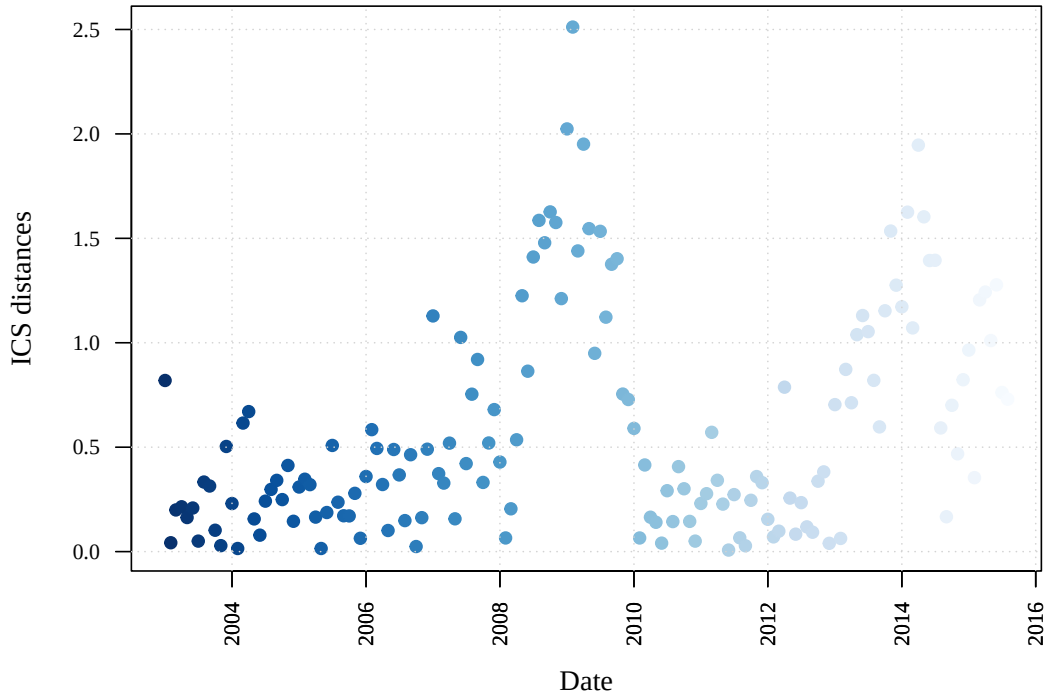


Figure 3.7: Distances of the selected Invariant Coordinates of the `BDDSegX` data set (compositions).

Conclusion

Adapting the Invariant Coordinate Selection method to compositional data can, as it is often the case when adapting general multivariate data analysis methods, be done through applying the classical ICS method after an ilr transformation.

Obviously, this does not provide a complete answer to the problem, since when transforming back through ilr^{-1} we lose most of the properties of ICS, namely that the directions of the outliers are the first coordinates.
Moreover, the ICS transformation matrix H depends on the choice of contrast matrix, even with an adjustment such as discussed at the end of Subsection 3.2.3.

A first way to avoid this dependence would be to choose a contrast matrix, for instance one that simultaneously diagonalizes the scatter pair, or one that transforms the orthogonal projection on the ilr-space into a subcomposition on the simplex, using balances (as in Section 4.8 of the book by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

In order to find a more intrinsic method, one could look for a link between the eigenvalues of the scatter-difference matrix $\hat{S}_1^{-1}\hat{S}_2$ and a notion of kurtosis according to the axes of the simplex (which are not orthogonal), instead of the approach that has been developed here which relies on the kurtosis according to the ilr coordinates.

Another approach would be to create a good notion of correlation between simplex coordinates (see the article by Quinn et al. 2017 which explains why correlation is not appropriate for relative data and suggests more pertinent methods).

Naturally, results like Theorem 3 in the article by Tyler et al. 2009 (proving that under a Gaussian mixture model, the direction of the outliers can be found on the first or last Invariant Coordinates) could be adapted to the ilr coordinates, but one could hope (if a more intrinsic method is found) to find similar results directly in simplex coordinates.

Appendix A

R scripts and plots

The scripts that lay the foundations for the development of an `ICSCoDa` package, and also enables the reproduction of the plots in this report can be found on the following GitHub:

`https://github.com/camillemndn/ICSCoDa`

R packages

[ANR18] Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. *ICSOutlier: Outlier Detection Using Invariant Coordinate Selection*. R package version 0.3-0. 2018. URL: `https://CRAN.R-project.org/package=ICSOutlier`.

[Ben+09] Tatiana Benaglia et al. "mixtools: An R Package for Analyzing Finite Mixture Models". In: *Journal of Statistical Software* 32.6 (2009), pp. 1–29. URL: `http://www.jstatsoft.org/v32/i06/`.

[Bor21] Hans W. Borchers. *pracma: Practical Numerical Math Functions*. R package version 2.3.3. 2021. URL: `https://CRAN.R-project.org/package=pracma`.

[Dah+19] David B. Dahl et al. *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4. 2019. URL: `http://xtable.r-forge.r-project.org/`.

[Fas16] Matteo Fasiolo. *An introduction to mvnfast. R package version 0.1.6*. 2016. URL: `https://CRAN.R-project.org/package=mvnfast`.

[Fas21] Matteo Fasiolo. *mvnfast: Fast Multivariate Normal and Student's t Methods*. R package version 0.2.7. 2021. URL: `https://github.com/mfasiolo/mvnfast/`.

[Gou+21] Vincent Goulet et al. *expm: Matrix Exponential, Log, etc*. R package version 0.999-6. 2021. URL: `http://R-Forge.R-project.org/projects/expm/`.

[MC20] Duncan Murdoch and E. D. Chow. *ellipse: Functions for Drawing Ellipses and Ellipse-Like Confidence Regions*. R package version 0.4.2. 2020. URL: `https://CRAN.R-project.org/package=ellipse`.

[Mes21] Stefano Meschiari. *latex2exp: Use LaTeX Expressions in Plots*. R package version 0.5.0. 2021. URL: `https://github.com/stefano-meschiari/latex2exp`.

[Neu14] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. 2014. URL: `https://CRAN.R-project.org/package=RColorBrewer`.

[NOT08] Klaus Nordhausen, Hannu Oja, and David E. Tyler. “Tools for Exploring Multivariate Data: The Package ICS”. In: *Journal of Statistical Software* 28.6 (2008), pp. 1–31. URL: http://www.jstatsoft.org/v28/i06/.

[NOT18] Klaus Nordhausen, Hannu Oja, and David E. Tyler. *ICS: Tools for Exploring Multivariate Data via ICS/ICA*. R package version 1.3-1. 2018. URL: https://CRAN.R-project.org/package=ICS.

[vTB21] K. Gerald van den Boogaart, Raimon Tolosana-Delgado, and Matevz Bren. *compositions: Compositional Data Analysis*. R package version 2.0-2. 2021. URL: http://www.stat.boogaart.de/compositions/.

[You+20] Derek Young et al. *mixtools: Tools for Analyzing Finite Mixture Models*. R package version 1.2.0. 2020. URL: https://CRAN.R-project.org/package=mixtools.

Bibliography

[ANR18] Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. "ICS for multivariate outlier detection with application to quality control". In: *Computational Statistics & Data Analysis* 128 (2018), pp. 184–199. URL: https://www.sciencedirect.com/science/article/pii/S0167947318301579.

[Car89] J.-F. Cardoso. "Source Separation Using Higher Order Moments". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1989, pp. 2109–2112.

[CJG16] Marc Comas-Cufí, Martín-Fernández Josep Antoni, and Mateu-Figueras Glòria. "Log-ratio methods in mixture models for compositional data sets". In: *SORT-Statistics and Operations Research Transactions* 1.2 (2016), pp. 349–374. URL: https://raco.cat/index.php/SORT/article/view/316149.

[Ego+11] Juan José Egozcue et al. "Elements of Simplicial Linear Algebra and Geometry". In: *Compositional Data Analysis*. John Wiley & Sons, Ltd, 2011. Chap. 11, pp. 139–157. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119976462.ch11.

[FHT18] Peter Filzmoser, Karel Hron, and Matthias Templ. *Applied Compositional Data Analysis. With Worked Examples in R*. Nov. 2018. URL: http://dx.doi.org/10.1007/978-3-319-96422-5.

[Mah36] Prasanta Chandra Mahalanobis. "On The Generalized Distance In Statistics". In: *Proceedings of the National Institute of Sciences (Calcutta)* 2 (1936), pp. 49–55.

[Mue+21] Christoph Muehlmann et al. "Independent Component Analysis for Compositional Data". In: *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*. Ed. by Abdelaati Daouia and Anne Ruiz-Gazen. Springer International Publishing, 2021, pp. 525–545. URL: https://doi.org/10.1007/978-3-030-73249-3_27.

[PET15] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modelling and Analysis of Compositional Data*. John Wiley & Sons, Ltd, 2015. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119003144.

[Qui+17] Thomas Quinn et al. "Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis". In: *Scientific Reports* 7 (Nov. 2017). URL: https://www.nature.com/articles/s41598-017-16520-0.

[Tyl+09] David E. Tyler et al. "Invariant co-ordinate selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3 (2009), pp. 549–592. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2009.00706.x.