Chapter 2

Exploratory analysis

Contents

Objectives

- ✓ To present the assumptions, principles, and techniques necessary to gain insight into CoDa via exploratory data analysis (EDA).
- ✓ To analyse the peculiarities of the reduced-dimensionality representation of a CoDa set.
- ✓ To show a procedure for creating an SBP according the criterion of maximizing the proportion of total variability retained by the balances.
- ✓ To introduce the most important probability distributions models on the simplex.

2.1. Centre of a compositional data set [†]

Standard descriptive statistics are not very informative in the case of CoDa. In particular, the arithmetic mean and the variance or standard deviation of individual components do not fit in with the compositional geometry as measures of central tendency and dispersion. They were defined as such in the framework of Euclidean geometry in real space, which is not a sensible geometry for CoDa. Therefore, it is necessary to introduce alternatives, which we find in the concepts of *centre*, *variation matrix* and *total variance*.

Let

$$\mathbf{X} = \{\mathbf{x}_i = [x_{i1}, \ldots, x_{iD}] \in \mathcal{S}^D : \ i = 1, \ldots, n\} \tag{2.1}$$

be a CoDa set of size n. The n rows $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of the matrix $\mathbf{x}$ correspond to samples, and the D columns $X_1, \ldots, X_D$ correspond to parts of CoDa.

2.1.1. Centre. A measure of central tendency for the CoDa set $\mathbf{X}$ is the closed geometric mean which is called the *centre* and is defined as

$$\mathbf{g} = \mathcal{C}[g_1, \ldots, g_D], \quad \text{with} \quad g_j = (\prod_{i=1}^{n} x_{ij})^{1/n}, \ j = 1, \ldots, D. \tag{2.2}$$

where $\mathcal{C}$ is the closure operator to a constant κ.

Note that in the definition of the centre of a compositional data set the geometric mean is considered column-wise (i.e. by variables), whereas in the *clr* transformation,

$$\operatorname{clr} \mathbf{x} = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right] ,$$

the geometric mean $g(\mathbf{x}) = \left(\prod_{j=1}^{D} x_j \right)^{1/D}$ is considered row-wise (i.e. by samples).

It is easy to prove that the centre $\mathbf{g}$ can be calculated from the arithmetic mean of the *clr*-scores set $\mathbf{Z} = \operatorname{clr} \mathbf{x}$, where clr is applied row-wise. More precisely,

$$\mathbf{g} = \operatorname{clr}^{-1}[\bar{Z}_1, \ldots, \bar{Z}_D] = \mathcal{C}[\exp \bar{Z}_1, \ldots, \exp \bar{Z}_D], \tag{2.3}$$

[†] This section is an adaptation of [DBB06, p. 161–163], [Ait86, Sections 4.1-4.9, p. 64-83]. Further information in [PET15, Sections 5.2-5.3, p. 66-69].