Chapter 1

The geometric structure of the sample space

Contents

Objectives

- ✓ To show the nature of compositional data along with the inconsistency and difficulties involved in applying standard statistical analysis to this type of data.
- ✓ To present the simplex as the *natural* sample space of compositional data.
- ✓ To introduce the principles on which the statistical analysis of compositional data should be based according to their nature.
- ✓ To define the two basic operations on the simplex —perturbation and powering— on which the statistical analysis of compositional data is based.
- ✓ To learn how to structure the simplex $\mathcal{S}^D$ in a Euclidean space of dimension $D-1$.
- ✓ To introduce the concept of *logcontrast* on the simplex $\mathcal{S}^D$ with special emphasis on the additive and the centred logratio transformations.
- ✓ To show the procedure for calculating the coordinates of a composition with respect to an orthonormal basis of $\mathcal{S}^D$ introducing isometric logratio transformations.
- ✓ To show a procedure for selecting a suitable orthonormal basis that allows the coordinates of a composition to be easily interpreted.

1.1. The sample space of compositional data

1.1.1. The simplex $\mathcal{S}^D$ as sample space[†]. The concept of compositional data (CoDa) is the starting point for the development of all the geometric and algebraic results that are necessary for building up reliable probabilistic and statistical models for such data.

Following on from earlier developments in CoDa [**Ait86**], a compositional vector of D parts, $\mathbf{x} = [x_1, \ldots, x_D]$, is defined as a vector in which the only relevant information is contained in the ratios between its components. All components of the vector are assumed positive. Throughout the text, components are called *parts* and a compositional vector is called a *composition*. The notation of the vector with square brackets means that this vector is considered to be a row vector.

The assertion that all the relevant information is contained in the ratios implies that, if a is a real positive number, the vectors $[x_1, \ldots, x_D]$ and $[ax_1, \ldots, ax_D]$ essentially convey the same information and are thus indistinguishable. Therefore, a composition is a class of equivalent compositional vectors [**BMP03, BM16**]. That is, proportional vectors with positive parts are *compositionally equivalent*.

A way to simplify the use of compositions is to represent them in closed form, that is, as positive vectors, the parts of which add up to a positive constant, κ. Common values of the closure constant κ are 1 for parts per unit, 100 for percentages, or 10^6 for parts per million. A consequence of this is that a composition of D

[†]This section is an adaptation of [**EP06**, p. 145-147]. Further information in [**Ait86**, Sections 2.1-2.6, p. 24-38] and [**PET15**, Section 2.1, p. 8-12].