Chapter 3

Exploratory compositional data analysis

Contents

3.1 Centre of a compositional data set
   3.1.1 Centre
   3.1.2 Centring a compositional data set

3.2 Covariance structure of a compositional data set
   3.2.1 Variation matrix
   3.2.2 Centred logratio covariance matrix
   3.2.3 Logratio covariance matrix
   3.2.4 Total variance
   3.2.5 Scaling a compositional data set

3.3 Reduced-dimensionality representation of a compositional data set
   3.3.1 Singular value decomposition
   3.3.2 Logcontrast principal components analysis
   3.3.3 Compositional biplot
   3.3.4 Interpretation of a compositional biplot
   3.3.5 Subcompositional analysis

Objectives

✓ To present the assumptions, principles, and techniques necessary to gain insight into compositional data via exploratory data analysis (EDA).
✓ To analyse the peculiarities of the reduced-dimensionality representation of a compositional data set.
3.1. Centre of a compositional data set

Standard descriptive statistics are not very informative in the case of compositional data. 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 compositional data. Therefore, it is necessary to introduce alternatives, which we find in the concepts of centre, variation matrix and total variance.

Let
\[
X = \{x_i = [x_{i1}, \ldots, x_{iD}] \in S^D : i = 1, \ldots, n\}
\]
be a compositional data set of size n. The n rows \(x_1, \ldots, x_n\) of the matrix \(X\) correspond to samples, and the D columns \(X_1, \ldots, X_D\) correspond to parts of compositional data.

3.1.1. Centre. A measure of central tendency for the compositional data set \(X\) is the closed geometric mean which is called the centre and is defined as
\[
g = C[g_1, \ldots, g_D], \quad \text{with} \quad g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}, \quad j = 1, \ldots, D.
\]
where \(C\) is the closure operator to constant \(\kappa\).

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,
\[
\text{clr } x = \left[\log \frac{x_1}{g(x)}, \ldots, \log \frac{x_D}{g(x)}\right],
\]
the geometric mean \(g(x) = \left(\prod_{j=1}^{D} x_j\right)^{1/D}\) is considered row-wise (i.e. by samples).

It is easy to prove that the centre \(g\) can be calculated from the arithmetic mean of the clr-transformed data set \(Z = \text{clr } X\). More precisely,
\[
g = \text{clr}^{-1}[\bar{Z}_1, \ldots, \bar{Z}_D] = C[\exp \bar{Z}_1, \ldots, \exp \bar{Z}_D],
\]
with
\[
\bar{Z}_j = \frac{1}{n} \sum_{i=1}^{n} \log \frac{x_{ij}}{g(x_i)}, \quad j = 1, \ldots, D.
\]
Similarly, \(g\) can be calculated from the arithmetic mean of the alr-transformed data set \(Y = \text{alr } X\), i.e.
\[
g = \text{alr}^{-1}[\bar{Y}_1, \ldots, \bar{Y}_{D-1}] = C[\exp \bar{Y}_1, \ldots, \exp \bar{Y}_{D-1}, 1],
\]
with
\[
\bar{Y}_j = \frac{1}{n} \sum_{i=1}^{n} \log \frac{x_{ij}}{x_D}, \quad j = 1, \ldots, D - 1.
\]

Notice that, although all samples \(x_i = [x_{i1}, \ldots, x_{iD}], i = 1, \ldots, n\) were closed to a scale constant \(\kappa\) (i.e. \(\sum_{j=1}^{D} x_{ij} = \kappa, i = 1, \ldots, n\)), the geometric mean vector

\[\text{This section is an adaptation of [DBB06, p. 161-163], [PET11, Sections 5.2 and 5.3, p. 38-40], [Ait86, Sections 4.1-4.9, p. 64-83]. Further information in [PET15, Sections 5.2-5.3, p. 66-69].}\]