EVALUATION EXCERCISES

Compositional Analysis of Data with *CoDaPack*

Online Course

CoDa-Research Group

University of Girona
Spain

2021 (updated)

Contents

Chapter 1

Evaluation exercises: 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.

Chapter 1 - Evaluation exercises

A. Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be respectively the compositions $[0.1, 0.4, 0.5]$, $[0.7, 0.2, 0.1]$ and $[0.2, 0.8, 0.1]$.

1. Perturb the composition $\mathbf{x}$ by the composition $\mathbf{z}$. Link the ratios t_i/t_j of the perturbed composition $\mathbf{t} = \mathbf{z} \oplus \mathbf{x}$ with the ratios x_i/x_j and z_i/z_j.
2. Find the composition $\mathbf{p}$ such that $\mathbf{p} \oplus \mathbf{x} = \mathbf{y}$. Find the composition $\mathbf{p}^*$ such that $\mathbf{p}^* \oplus \mathbf{y} = \mathbf{x}$. Analyze the relationship between $\mathbf{p}$ and $\mathbf{p}^*$.
3. Find the composition $\mathbf{w} = 1.2 \odot \mathbf{x}$. Link the ratios w_i/w_j of the *powered* composition $\mathbf{w}$ with the ratios x_i/x_j of $\mathbf{x}$.

B. Vitamin C contains three elements —carbon (C), hydrogen (H), and oxygen (O)— according to the following *mass* percentages: 40.9%, 4.6% and 54.5%, respectively. To express the composition of vitamin C in *mole* percentages we need the atomic weight of their components C, H and O. They are 12.01, 1.01 and 16.00, respectively. To convert the mass of an element into moles it suffices to divide its weight (in grams) by its atomic weight.

1. Calculate the composition [C, H, O] of vitamin C expressed in *mole* percentages.
2. Interpret the conversion of *mass* percentages into *mole* percentages in terms of a perturbation operation in the simplex.

C. Let $\mathbf{x}_0$ and $\mathbf{v}$ be the compositions $[0.1, 0.4, 0.5]$ and $[0.2, 0.8, 0.1]$, respectively. Calculate the compositions $\mathbf{x}_0 \oplus (t \odot \mathbf{v})$, with t ranging from -3 to $+3$ (in steps of 0.25). Plot the resulting set of compositions in a ternary diagram. What do you observe?

D. The universal *law of radioactive decay* relates the current amount $N_A(0)$ of a radioisotope A and the amount $N_A(t)$ after a time t. The law establishes that

$$N_A(t) = N_A(0) \cdot \exp\left(-\frac{\ln 2}{T_{A,1/2}} \cdot t\right) ,$$

where $T_{A,1/2}$ is the *half-time* disintegration period of the radioisotope A, i.e. the time required for one half of the atoms in any starting sample of the radioisotope A to decay.

Suppose that a mineral assemblage contains radioactive isotopes. More specifically, the current composition (in *ppm*) of $[{}^{238}\mathrm{U}, {}^{232}\mathrm{Th}, {}^{40}\mathrm{K}]$ is $[110, 50, 150]$. The half-time disintegration periods of ^{238}U, ^{232}Th and ^{40}K are $4.468 \cdot 10^9$, $14.05 \cdot 10^9$ and $1.277 \cdot 10^9$ years, respectively.

1. What will be the composition of $[{}^{238}\mathrm{U}, {}^{232}\mathrm{Th}, {}^{40}\mathrm{K}]$ after 10^{10} years?
2. Demonstrate that the composition $\mathbf{x}_t$ of $[{}^{238}\mathrm{U}, {}^{232}\mathrm{Th}, {}^{40}\mathrm{K}]$ after t years can be expressed (using compositional operators) as

$$[110, 50, 150] \oplus \left(t \odot [2^{-\frac{1}{4.468\times 10^9}}, 2^{-\frac{1}{14.05\times 10^9}}, 2^{-\frac{1}{1.277\times 10^9}}]\right) .$$

3. Calculate $\mathbf{x}_t$ for $t = 0, 10^6, 10^7, 10^8, 10^9, 10^{10}, 2 \cdot 10^{10}, 3 \cdot 10^{10}, 4 \cdot 10^{10}, 5 \cdot 10^{10}$ and 10^{11}. Plot the resulting set of compositions on a ternary diagram. What do you observe?

E. Let $\mathbf{x} = [10, 20, 30, 40]$ and $\mathbf{y} = [50, 30, 10, 10]$ be compositions of $\mathcal{S}^4$.

1. Calculate the logratio vectors $\operatorname{alr}\mathbf{x}$, $\operatorname{alr}\mathbf{y}$, $\operatorname{clr}\mathbf{x}$ and $\operatorname{clr}\mathbf{y}$.
2. Calculate the composition $3 \odot \mathbf{x}$. Calculate the logratio vectors $\operatorname{alr}(3 \odot \mathbf{x})$ and $\operatorname{clr}(3 \odot \mathbf{x})$. Link them with vectors $3 \cdot \operatorname{alr}\mathbf{x}$ and $3 \cdot \operatorname{clr}\mathbf{x}$.
3. Calculate the composition $\mathbf{x} \oplus \mathbf{y}$. Calculate the logratio vectors $\operatorname{alr}(\mathbf{x} \oplus \mathbf{y})$ and $\operatorname{clr}(\mathbf{x} \oplus \mathbf{y})$. Link them with vectors $\operatorname{alr}\mathbf{x} + \operatorname{alr}\mathbf{y}$ and $\operatorname{clr}\mathbf{x} + \operatorname{clr}\mathbf{y}$.
4. Prove that $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is an orthonormal basis of $\mathcal{S}^4$, where
$$\begin{aligned} \mathbf{e}_1 &= \operatorname{clr}^{-1}[\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}] \\ \mathbf{e}_2 &= \operatorname{clr}^{-1}[\tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{2}}, 0, 0] \\ \mathbf{e}_3 &= \operatorname{clr}^{-1}[0, 0, \tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{2}}] \ , \end{aligned}$$
5. Calculate the *olr*-coordinates of the vectors $\mathbf{x}$ and $\mathbf{y}$ in the basis $\mathcal{B}$.
6. Calculate the *olr*-coordinates of the vectors $3 \odot \mathbf{x}$ and $\mathbf{x} \oplus \mathbf{y}$. Relate these coordinates with the *olr*-coordinates of $\mathbf{x}$ and $\mathbf{y}$.
7. Calculate the Aitchison norm of $\mathbf{x}$. Calculate the unitary composition $\mathbf{w} \in \mathcal{S}^4$ with the same direction and orientation as $\mathbf{x}$.
8. Calculate the Aitchison inner product $\langle \mathbf{x}, \mathbf{y} \rangle_a$. Calculate the norm of $\mathbf{y}$. Calculate the angle between $\mathbf{x}$ and $\mathbf{y}$.

F. Let $\mathbf{x}^*_{1,alr} = \ln 2 \cdot [3, 2, 1]$ be the *alr*-coordinates of a composition $\mathbf{x}_1 \in \mathcal{S}^4$. Let $\mathbf{x}^*_{2,clr} = \frac{\ln 2}{2} \cdot [3, 1, -1, -3]$ be the *clr*-coordinates of a composition $\mathbf{x}_2 \in \mathcal{S}^4$.

1. Calculate $\mathbf{x}_1$ and $\mathbf{x}_2$ to prove that $\mathbf{x}_1$ and $\mathbf{x}_2$ is the same composition.
2. Let $\mathbf{x}^*_{3,ilr} = \ln 2 \cdot [2, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ be the *olr*-coordinates of a composition $\mathbf{x}_3 \in \mathcal{S}^4$ with respect to the basis $\mathcal{B}$. Calculate $\mathbf{x}_3$ to prove that it is also the same composition.

G. Use the `householdbudget` CoDa set described in the Appendix. This data set records the expenditures on four commodity groups of forty single persons (twenty men and twenty women) living alone in rented accommodation. We restrict the analysis to the expenditures of the twenty men. Load the `householdbudgetmen.cdp` file from the CoDaPack instead of loading the complete data set.

We want to analyse the balance between `Food` and the three other expenditure parts. Moreover, we also wish to analyse the balance between `Hous` and the other non-food expenditure parts, and finally the balance between `Serv` and `Others`. We need to create the SBP according to these analyses.

1. Write the sign matrix $\mathbf{S}$ associated to the SBP.
2. Let $\mathcal{B}$ be the *olr*-basis associated to the SBP. Use CoDaPack (menu *Data* ▷ *Transformations* ▷ *raw-ILR*) to calculate the *olr*-coordinates of the twenty compositions [`Hous`, `Food`, `Serv`, `Others`] of the data file.
3. Take the *olr*-coordinates of one sample and interpret them in terms of the balances.
4. Plot the balance-dendrogram associated to the SBP (menu *Graphs* ▷ *Balance dendrogram*).

Chapter 2

Evaluation exercises: Exploratory analysis and distributions on the simplex

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.

Chapter 2 - Evaluation exercises

In this exercise we use the `householdbudget` CoDa set (see Appendix). Remember that this data set records the expenditures on four commodity groups of forty single persons (twenty men and twenty women) living alone in rented accommodation. The four groups of weekly expenditures are: a) housing, including fuel and light (`Hous`); b) foodstuffs, including alcohol and tobacco (`Food`); c) services, including transport and vehicles in daily hours (`Serv`); and d) other goods, including clothing, footwear and durable goods (`Others`). Restrict your analysis to the expenditures of the twenty men. Therefore, load the `householdbudgetmen.cdp` file in CoDaPack instead of loading the complete data set.

A. A first description of the data set: draw the quaternary diagram and calculate the basic compositional statistics. Describe the center of the data set and the log-ratio variances (pairwise and *clr*).

B. We want to analyse the balance between `Food` and the three other expenditure variables. Moreover, we also wish to analyse the balance between `Hous` and the other non-food expenditure variables, and finally the balance between `Serv` and `Others`.
   1. Write the sign matrix **S** associated to the above SBP.
   2. Use the (*Data* ▷ *Transformations* ▷ *Raw-ILR* menu) to calculate the *olr*-coordinates of the twenty compositions [`Hous`,`Food`,`Serv`,`Others`].
   3. Calculate the basic statistics of the *olr*-coordinates. Interpret the balances associated to each one of these three *olr*-coordinates.
   4. Plot the balance dendrogram associated to the SBP (*Graphs* ▷ *Balance dendrogram* menu). Indicate the most important features of the graph.
   5. Use CoDaPack (*Graphs* ▷ *Scatterplot 2D/3D* to plot the *olr*-coordinates in $\mathbb{R}^3$ and in the three 2-dimensional scatter plots ($\text{ilr}_1\times$ ilr_2, $\text{ilr}_1\times$ ilr_3 and $\text{ilr}_2\times$ ilr_3). Is there any particular pattern?
   6. Use the *Statistics* ▷ *Classical statistics summary* menu of CoDaPack to calculate the correlation between the three *olr*-coordinates. Do you detect linear relations between the *olr*-coordinates

   CoDaPack does not allow us to fit regression lines to bivariate data sets. Therefore, $\mathtt{ilr}_1$, $\mathtt{ilr}_2$ and $\mathtt{ilr}_3$ should be exported to a statistical software to estimate the coefficients of the regression lines.

   7. Export the CoDaPack data file from the *File* ▷ *Export* ▷ *Export Table to XLS...* menu (to obtain an Excel® file), or from *File* ▷ *Export* ▷ *Export Table to R...* menu to obtain an **R** file.
   8. Estimate the two coefficients of the simple regression line adjusted to the bivariate data ($\mathtt{ilr}_2$, $\mathtt{ilr}_3$).

   According to the equation of the regression line, it holds that

$$\mathtt{ilr}_3 \simeq -0.24 + 0.63 \cdot \mathtt{ilr}_2 \ .$$

Remember that the $\texttt{ilr}_2$ and $\texttt{ilr}_3$ coordinates correspond to specific balances between some parts of the full composition [Hous,Food,Serv,Others].

9. Demonstrate that the previous equation can be equivalently rewritten as

$$0.51 \cdot \ln(\texttt{Hous}) - 0.96 \cdot \ln(\texttt{Serv}) + 0.45 \cdot \ln(\texttt{Others}) \simeq 0.24 \ ,$$

where the first term of this equation is a logcontrast of parts of the composition [Hous,Food,Serv,Others].

C. Explore the covariance CoDa-biplot for the 4-part compositions [Hous,Food,Serv,Others]: redundant parts, linear association, independence, etc. Indicate an SBP suggested by the position of the rays of the *clr*-variables in the biplot.

D. Explore the form CoDa-biplot for the 4-part compositions [Hous,Food,Serv,Others]: identify some similar samples, extreme samples and the sample closest to the center of the data set.

E. Apply the constrained PCs algorithm to create an *olr*-basis.

F. Analyse which is the 3-part subcomposition with the largest total variance.

G. Analyse the normality of the CoDa-set. Consider the 3-part subcomposition retaining the largest proportion of total variance and plot the predictive regions. Are there samples very far from the center of the data set (*potential outliers*)?

Chapter 3

Evaluation exercises: data pre-processing: irregular data

Contents

Objectives

✓ To deal with the most common irregular data in CoDa: missing data, values below detection limit and zeros.

✓ To distinguish the type of zeros and accordingly decide the procedure for dealing with them.

✓ To know how to detect potential outliers in CoDa.

Chapter 3 - Evaluation exercises

The `houseexpend.cdp` file (see Appendix) records the composition of proportions of mean consumption expenditures of households on $D = 12$ domestic year costs in $n = 27$ states of the European Union in 2005 (`hexpcomplete`). In addition the data set contains the gross domestic product (`GDP05`) and (`GDP14`) in years 2005 and 2014, respectively.

A. Some samples of the CoDa set contain forced zeros. Note that the `houseexpend.cdp` file also has the data frames `hexp001` and `hexp000`. In the former the forced zeros are labelled "0[0.01]", whereas the second contains the original zero values.
   - Create and describe the zero pattern plot.
   - Go to the R-program and analyze the differences among the variation matrices by the groups of samples defined by the zero pattern.
   - Are these zeros essential zeros? Can the zeros be assumed to be rounded zeros?
   - Apply the multiplicative replacement to the CoDa set.
   - Use the modified log-ratio EM algorithm to impute the zeros.
   - Explore the data sets. That is, apply numerical and graphical descriptive statistics for the 12-part compositions for comparing the imputed CoDa sets from multiplicative, modified log-ratio EM algorithms and the original complete CoDa set. Which replacement method is more recommendable?
   - Explore the potential outliers in the three CoDa sets. Did the replacements create any artificial outlier?
   - Create a factor variable (binary variable) to indicate when a country takes a value below or above 100 in the `GDP05` variable. Explore the variation arrays and the *clr*-biplot for the groups defined by the factor variable.
   - In the light of the descriptive results, create an SBP and calculate the corresponding *olr*-coordinates with the aim to discriminate the two groups. Did you find any logratio that splits the samples of the two groups?

B. We focus on a subcomposition of $D = 4$ parts, which includes the expenditures on `foodstuff`, `housing`, `health`, and `communications`. The parts, `foodstuff` and `housing`, are basic costs, while `health` and `communications` are costs more related to economic status, that is, to `GDP`.
   - Apply numerical and graphical descriptive statistics for describing this subcomposition formed from the `hexpcomplete` data frame.
   - Analyse potential outliers as regards the 4-part subcomposition.

Chapter 4

Evaluation exercises: linear regression models (LRM)

Contents

Objectives

✓ To estimate and interpret an LRM when the response is compositional.
✓ To estimate and interpret an LRM when the predictor is compositional.
✓ To introduce some extensions for an LRM

Chapter 4 - Evaluation exercises

A. The `houseexpend.cdp` file (see Appendix) records the data set *household expenditures* proposed in [**Eg+12**] for creating an LRM. This data comes from Eurostat (the European Union's statistical information service) and is displayed in Table 5.2 of this article. The collected data in the file represents the composition of proportions of mean consumption expenditures of households on $D = 12$ domestic year costs in $n = 27$ states of the European Union in 2005. In addition the data set contains the gross domestic product (`GDP05`) and (`GDP14`) in years 2005 and 2014, respectively. Analogously to [**Eg+12**] we focus on a subcomposition of $D = 4$ parts, which includes the expenditures on `foodstuff`, `housing`, `health`, and `communications`. The parts, `foodstuff` and `housing`, are basic costs, while `health` and `communications` are costs more related to economic status, that is, to `GDP`. The file has several data frames of which we use `hexpcomplete`.

- Use a box plot to perform an outlier analysis in regard to the covariate `GDP05`.
- Apply log-transformation to `GDP05`.
- Could `Luxembourg` be considered as a potential outlier?
- Do an SBP and compute the *olr*-coordinates.
- Compare the correlation between the response variables (*olr*-coordinates) and the real covariates `GDP05` and ln(`GDP05`), respectively.
- Do the same with the variable `GDP05` where `Luxembourg` has been filtered out of the data frame.
- What is your recommendation?
- Exclude `Luxembourg` from further computations.
- Compositional LRM:
  - Build the LRM where the response is the 4-part subcomposition [`foodstuff`, `housing`, `health`, `communications`] and the real covariate is ln(`GDP05`).
  - Interpret the coordinate coefficients $\boldsymbol{\beta}^*$ and the compositional coefficients $\boldsymbol{\beta}$.
  - Analyse the quality of the LRM (R^2).
  - Evaluate the normality of the LRM (compositional residuals and significance of coefficients).
  - Analyse the potential outliers in the compositional residuals.

B. The `waste.cdp` CoDa set (see Appendix for a description) includes the variable `floating_population` measured in 215 municipalities of Catalonia [**CMF17**]. The variable is expressed as equivalent full-time residents. In consequence, floating population may be positive if the municipality is receiving more short term residents than it is sending elsewhere, or negative if the opposite holds (expressed as a percentage above -if positive- or below -if negative- the census count). The categorical variable `floating_population_cat` includes information on the floating population sign, coded as "+" or "–". The waste composition can be used to predict floating population:

x_1: non recyclable (grey waste container in Catalonia),
x_2: glass (bottles and jars of any colour: green waste container),
x_3: light containers (plastic packaging, cans and tetra packs: yellow container),
x_4: paper and cardboard (blue container), and
x_5: biodegradable waste (brown container).

We want to estimate the LRM where the waste composition is the predictor of the floating population:

- Analyze the classical statistics summary of `floating_population`. Perform a classical univariate normality test and draw a box plot (*Draw mean: none* option). Any suggestion regarding the normality and potential outliers of this variable?
- Analyze the compositional statistics of the waste composition [`x1_non_rec`, `x2_glass`, `x3_plastic`, `x4_paper`, `x5_bio`]. Which part dominates the geometric center? Which parts dominate the variance?
- Draw the biplot of the waste composition. As *Groups* use the variable `floating_population_cat`. Are there meaningful associations between the composition and floating population?
- Plot the balance dendrogram by `floating_population_cat` and calculate the corresponding *olr*-coordinates using the following SBP:

$\mathrm{olr_k}$	`x1_non_rec`	`x2_glass`	`x3_plastic`	`x4_paper`	`x5_bio`
x_1^*	$+1$	$+1$	-1	-1	-1
x_2^*	$+1$	-1	0	0	0
x_3^*	0	0	$+1$	-1	-1
x_4^*	0	0	0	$+1$	-1

- Compute the correlation coefficient of each *olr*-coordinate against the variable `floating_population`.
- Draw the scatterplot of the variable `floating_population` against each *olr*-coordinate.
- Estimate the LRM in coordinates. Set the same SBP partition as before and request residuals and fitted values. Are all the coefficients significant? Analyse the residuals of the LRM from the plots provided. What are your conclusions?
- Set an advanced filter to remove some potential outliers from the model (i.e., observations with `floating_population` > 1).
- Rerun the LRM and analyse the results. What are your conclusions? Which is the interpretation of the $\boldsymbol{\beta}^*$ coefficients?
- Compute the compositional gradient. Which waste components lead to an increase or decrease of the floating population in a municipality?

C. The variable `total` of the `waste.cdp` CoDa set (see Appendix) indicates the total waste generated in *tons per census inhabitant and year*. Take the LRM created in the previous excercise. With the R-prorgram add the logarithm of the variable `total` as another predictor into the LRM. Is the added predictor significant? Does it improve the predictive power of the LRM? Do the effects of the *olr*-coordinates change?

Specific references in this chapter

[CMF17] G. Coenders, J.A. Martín-Fernández and B. Ferrer-Rosell, *When relative and absolute information matter: compositional predictor with a total in generalized linear models*, Statistical Modelling **17**(6) (2017), 494–512.

[Eg+12] J.J. Egozcue, J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron and P. Filzmoser, *Simplicial regression. The normal model*, Journal of Applied Probability and Statistics **6** (2012), 87–108.

Chapter 5

Evaluation exercises: on the analysis of grouped data

Contents

Objectives

- ✓ To learn how to form groups when the data set is compositional.
- ✓ To introduce how to calculate the linear discriminant function as a log-contrast.
- ✓ To properly analyse the difference between the centres of several groups using the MANOVA test.

Chapter 5 - Evaluation exercises

A. A 4-group solution has been considered as the optimal grouping of the CoDa set `mammalsmilk` (see Appendix) when traditional statistical techniques are applied. Use the k-means algorithm to check, when using CoDa, if this 4-group solution is still appropriate.

B. The 45 samples of the `pottery` CoDa set (see Appendix) come from five different kiln sites.

   (a) After an exploratory analysis (including k-means), show that we can conclude that samples can be classified in only three groups.

   (b) Perform a LDA (3 groups) based on the chemical composition [Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, MnO, BaO].
   Evaluate the accuracy of the technique.

   (c) Classify the new sample $[11.5, 5.8, 6.1, 0.2, 0.2, 4.1, 0.5, 0.1, 0.02]$.

   (d) Make a MANOVA test (3 groups) so as to analyse and characterise the differences between groups.

C. The `serprot.txt` file (see Appendix) records 4-part compositions of serum proteins [albumin, pre-albumin, globulin A, globulin B] from blood samples of 36 patients. Thirty patients are known to have two different diseases (variable `disease` $= 1, 2$). Six patients are still unclassified (`disease` $= 0$).
Could you help us to classify the disease of these six patients?

D. The `milkcows` CoDa set (see Appendix) records the milk compositions for the 58 cows (before and after a strictly controlled dietary and hormonal regime over a period of eight weeks), showing the protein, milk fat, carbohydrate, calcium, sodium, potassium proportions by weight of total dietary content. The experiment with the cows consists of a strictly controlled dietary and hormonal regime over a period of eight weeks.
Could you help us to decide if the regime applied was useful? [*Note: the experimental design corresponds to a paired data design.*]

Appendix: Some typical compositional problems

Contents

We present the reader with a series of challenging problems in compositional data analysis, with typical data sets and the questions they pose. These come from a number of different disciplines and will be used to elicit the concepts and principles of compositional data analysis. For further detailed information see [**Ait86**, Sections 1.1-1.14, p. 1-20].

A.1. Chemical compositions of Romano-British pottery

In archaeology, the compositional analysis of raw materials (clays used to make pottery, lithic materials used to make stone tools, etc.) has become a key tool for examining trade and exchange in ancient economies. Different sources for such materials often have distinct chemical 'signatures' that can be identified in places far from their point of origin. One interpretative challenge is to take an often large, complex array of chemical assays and identify patterns that can be exploited in higher level interpretations.

The `pottery` data set consists of data pertaining to the chemical composition of 45 specimens of Romano-British pottery. The method used to generate these data is atomic absorption spectophotometry, and readings for nine oxides (Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, MnO, and BaO) are provided. These samples come from five different kiln sites, and one of the issues we want to consider is the degree to which compositional data help distinguish pottery from these various kilns.

☞ Back to Index

A.2. Arctic lake sediments at different depths

In sedimentology, specimens of sediments are traditionally separated into three mutually exclusive and exhaustive constituents —sand, silt and clay— and the proportions of these parts by weight are quoted as [sand, silt, clay] compositions. The `arctic_lake` data set records the [sand, silt, clay] compositions of 39 sediment samples at different water depths in an Arctic lake. Again we recognise substantial variability between compositions. Questions of obvious interest here are the following. Is sediment composition dependent on water depth? If so, how can we quantify the extent of the dependence? If we regard sedimentation as a process, do these data provide any information on the nature of the process? Even at this stage of investigation we can see that this may be a question of compositional regression.

☞ Back to Index

A.3. Household budget patterns

An important aspect in the study of consumer demand is the analysis of household budget surveys, in which attention often focuses on the expenditures of a sample of households on a number of mutually exclusive and exhaustive commodity groups and their relation to total expenditure, income, type of housing, household composition and so on. In the investigation of such data the pattern or composition of expenditures, the proportions of total expenditure allocated to the commodity groups, can be shown to play a central role in a form of budget share approach to the analysis. Assurances of confidentiality and limitations of space preclude the publication of individual budgets from an actual survey, but we can present a reduced version of the problem, which retains its key characteristics.

In a sample survey of single persons living alone in rented accommodation, twenty men and twenty women were randomly selected and asked to record over a period of one month their expenditures on the following four mutually exclusive and exhaustive commodity groups:

- `Hous:` Housing, including fuel and light.
- `Food:` Foodstuffs, including alcohol and tobacco.
- `Serv:` Services, including transport and vehicles.
- `Other:` Other goods, including clothing, footwear and durable goods.

The results are recorded in the `householdbudget` data set.

Interesting questions are readily formulated: to what extent does the pattern of the budget share of expenditures for men depend on the total amount spent? Are there any differences between men and women in their expenditure patterns? Are there any commodity groups which are given priority in the allocation of expenditure?

☞ Back to Index

A.4. Milk composition study

In an attempt to improve the quality of cow milk, milk from each of thirty cows was assessed by dietary composition before and after a strictly controlled dietary and hormonal regime over a period of eight weeks. Although seasonal variations in milk quality might have been regarded as negligible over this period, it was decided to have a control group of thirty cows kept under the same conditions but on a regular established regime. The sixty cows were of course allocated to control and treatment groups at random. The `milkcows` data set provides the complete set of before and after milk compositions for the sixty cows, showing the protein `(pr)`, milk fat `(mf)`, carbohydrate `(ch)`, calcium `(Ca)`, sodium `(Na)` and potassium `(K)` proportions by weight of total dietary content.

The purpose of the experiment was to determine whether the new regime had produced any significant change in the milk composition. It is, therefore, essential to have a clear idea of how change in compositional data is characterised by some meaningful operation. Thus, a key question here is how to formulate hypotheses of change of compositions, and indeed how we may investigate the full lattice of such hypotheses. Meanwhile we note that because of the before and after nature of the data within each experimental unit we have for compositional data the analogue of a paired comparison situation for real measurements where traditionally the differences in pairs of measurements are considered. Thus, we have to find the counterpart of difference for paired compositions.

☞ Back to Index

A.5. A statistician's time budget

Time budgets –how a day or a period of work is divided up into different activities– have become a popular source of data in psychology and sociology. To illustrate such problems we consider six daily activities undertaken by an academic statistician: teaching (`T`); consultation (`C`); administration (`A`); research (`R`); other wakeful activities (`O`); and sleep (`S`).

The **`statisticiantimebudget`** data set records the daily time (in hours) devoted to each activity, recorded on each of 20 days, selected randomly from working days in alternate weeks so as to avoid possible carry-over effects such as a short-sleep day being compensated by make-up sleep on the succeeding day. The six activities may be divided into two categories: 'work' comprising activities `T`, `C`, `A`, and `R`, and 'leisure', comprising activities `O` and `S`. Our analysis may then be directed towards the work pattern consisting of the relative times spent in the four work activities, the leisure pattern, and the division of the day into work time and leisure time. Two obvious questions are as follows. To what extent, if any, do the patterns of work and of leisure depend on the times allocated to these major divisions of the day? Is the ratio of sleep to other wakeful activities dependent on the times spent in the various work activities?

☞ Back to Index

A.6. The MN blood system

In humans the main blood group systems are the ABO system, the Rh system and the MN system. The *MN blood system* is a system of blood antigens also related to proteins of the red blood cell plasma membrane. The inheritance pattern of the MN blood system is autosomal with codominance, a type of lack of dominance in which the heterozygous manifests a phenotype totally distinct from the homozygous. The possible phenotypical forms are three blood types: type M blood, type N blood and type MN blood. The frequencies of M, N and MN blood types vary widely depending on the ethnic population. However, the Hardy-Weinberg principle states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. This implies that, in the long run, it holds that

$$\frac{x_{MM} \cdot x_{NN}}{(x_{MN})^2} = \frac{1}{4} \,, \tag{5.1}$$

where x_{MM} and x_{NN} are the genotype relative frequencies of MM and NN homozygotes, respectively, and x_{MN} is the genotype relative frequency of MN heterozygotes. This principle was named after G.H. Hardy and W. Weinberg demonstrated it mathematically.

We will use the `bloodMN` data set to analyse how the relative frequencies of MM, NN and MN blood types are distributed, and to verify the Hardy–Weinberg principle. This data set records the information on the absolute frequencies of M, N, and MN blood types observed in samples coming from different ethnic groups around the world. This data set comes from [**Boy50**].

☞ Back to Index

A.7. Mammal's milk

The mammalsmilk data set contains the percentages of five constituents (W: water, P: protein, F: fat, L: lactose, and A: ash) of the milk of 24 mammals. The data are taken from [**Har75**]. We will analyse whether there are large differences between the compositions of milks and classify them into groups according to the similarity of their constituents.

☞ Back to Index

A.8. Calc-alkaline and tholeiitic volcanic rocks

This petrafm data set is formed by 100 classified volcanic rock samples from Ontario (Canada). The three parts are:

$$[\mathtt{A} : Na_2O + K_2O; \mathtt{F} : FeO + 0.8998 \cdot Fe_2O_3; \mathtt{M} : MgO].$$

Rocks from the calc-alkaline magma series (25) can be well distinguished from samples from the tholeiitic magma series (75) on an AFM diagram. This data set is a typical example where a discriminant analysis based on the composition could be useful to classify new samples of volcanic rocks.

☞ Back to Index

A.9. Concentration of minor elements in carbon ashes

The montana data set consists of 229 samples of the concentration (in ppm) of minor elements [Cr, Cu, Hg, U, V] in carbon ashes from the Fort Union formation (Montana, USA), side of the Powder River Basin. The formation is mostly Palaeocene in age, and the coal is the result of deposition in conditions ranging from fluvial to lacustrine. All samples were taken from the same seam at different sites over an area of 430 km by 300 km, which implies that on average, the sampling spacing is 24 km. Using the spatial coordinates of the data, a semivariogram analysis was conducted for each chemical element in order to check for a potential spatial dependence structure in the data (not shown here). No spatial dependence patterns were observed for any component, which allowed us to assume an independence of the chemical samples at different locations.

The aforementioned chemical components actually represent a fully observed subcomposition of a much larger chemical composition. The five elements are not closed to a constant sum. Note that, as the samples are expressed in parts per million and all concentrations were originally measured, a residual element could be defined to fill up the gap to 10^6. We use this data set to evaluate the algorithms for missing data.

☞ Back to Index

A.10. Paleocological compositions

The foraminiferal data set (Aitchison, 1986) is a typical example of paleocological data. It contains compositions of 4 different fossils (Neogloboquadrina atlantica, Neogloboquadrina pachyderma, Globorotalia obesa, and Globigerinoides triloba) at

30 different depths. Due to the rounded zeros present in the data set we will apply some zero replacement techniques to impute these values in advance. After data preprocessing, the analysis that should be undertaken is the association between the composition and the depth.

☞ Back to Index

A.11. Pollen composition in fossils

The `pollen` data set is formed by 30 fossil pollen samples from three different locations (recorded in variable `group`) . The samples were analysed and the 3-part composition [`pinus`, `abies`, `quercus`] was measured. The aim was to determine whether the compositions differ significantly from one location to the other.

☞ Back to Index

A.12. Food consumption in European countries

The `alimentation` data set contains the percentages of consumption of several types of food in 25 European countries during the 80s. The categories are: red meat (pork, veal, beef), white meat (chicken), eggs, milk, fish, cereals, starch (potatoes), nuts, and fruits and vegetables. The file also contains a categorical variable that shows if the country is from the North or a Southern Mediterranean country. In addition, the countries are classified as Eastern European or as Western European. The aim is to analyse the similarities between countries as regards to their food consumption and to look for associations among the categorical variables.

☞ Back to Index

A.13. Household expenditures

From Eurostat (the European Union's statistical information service) the `houseexpend` data set records the composition on proportions of mean consumption expenditure of households expenditures on 12 domestic year costs in 27 states of the European Union. Some values in the data set are rounded zeros. In addition the data set contains the gross domestic product (`GDP05`) and (`GDP14`) in years 2005 and 2014, respectively. An interesting analysis is the potential association between expenditures compositions and GDP. Once a linear regression model is established, predictions can be provided.

☞ Back to Index

A.14. Serum proteins

The `serprot` data set records the percentages of the four serum proteins from the blood samples of 30 patients. Fourteen patients have one disease (1) and sixteen are known to have another different disease (2). The 4-compositions are formed by the proteins [albumin, pre-albumin, globulin A, globulin B]. The aim is to construct a diagnostic system based on these serum proteins so as to classify six new patients (0).

☞ Back to Index

A.15. Physical activity and body mass index

The `BMIPhisActi` data set records the proportion of daily time spent to sleep (`sleep`), sedentary behaviour (`sedent`), light physical activity (`Lpa`), moderate physical activity (`Mpa`) and vigorous physical activity (`Vpa`) measured on a small population of 393 children. Moreover the standardized body mass index (`zBMI`) of each child was also registered.

This data set was used in the example of the article [**Dum19**] to examine the expected differences in `zBMI` for reallocations of daily time between sleep, physical activity and sedentary behaviour. Because the original data is confidential, the data set `BMIPhisActi` includes simulated data that mimics the main features of the original data.

☞ Back to Index

A.16. Hotel posts in social media

The `weibo_hotels` data set aims at comparing the use of `Weibo` (`Facebook` equivalent in China) in hospitality `e`-marketing between small and medium accommodation establishments (private hostels, small hotels) and big and well-established business (such as international hotel chains or large hotels) in China. The 50 latest posts of the Weibo pages of each hotel ($n = 10$) are content-analyzed and coded regarding the count of posts featuring information on a 4-part composition [facilities, food, events, promotions]. Hotels were coded as large "L" or small "S" in the `hotel_size` categorical variable. As this small data set contains zeros we will use it to practice zero replacement methods for count zeros.

☞ Back to Index

A.17. The waste composition in Catalonia

The actual population residing in a municipality of Catalonia is composed by the census count and the so-called floating population (tourists, seasonal visitors, hostel students, short-time employees, and the like). Since actual population combines long and short term residents it is convenient to express it as equivalent full-time residents. Floating population may be positive if the municipality is receiving more short term residents than it is sending elsewhere, or negative if the opposite holds (expressed as a percentage above –if positive– or below –if negative– the census count). The `waste` data set includes this information in the variable `floating_population`. Floating population has a large impact on solid waste generation and thus waste can be used to predict floating population which is a hard to estimate demographic variable. This case study was presented in [**Coe17**].

Tourists and census population do not generate the same volume of waste and have different consumption and recycling patterns (waste composition). The Catalan Statistical Institute (IDESCAT) publishes official floating population data for all municipalities in Catalonia (Spain) above 5000 census habitants. The composition of urban solid waste is classified into $D = 5$ parts:

- x_1: non recyclable (grey waste container in Catalonia),
- x_2: glass (bottles and jars of any colour: green waste container),
- x_3: light containers (plastic packaging, cans and tetra packs: yellow container),
- x_4: paper and cardboard (blue container), and
- x_5: biodegradable waste (brown container).

☞ Back to Index

A.18. Employment distribution in EUROSTAT countries

According to the three–sector theory, as a country's economy develops, employment shifts from the primary sector (raw material extraction: farming, hunting, fishing, mining) to the secondary sector (industry, energy and construction) and finally to the tertiary sector (services). Thus, a country's employment distribution can be used as a predictor of economic wealth.

The `eurostat_employment_2008` data set contains EUROSTAT data on employment aggregated for both sexes, and all ages distributed by economic activity (classification 1983-2008, NACE Rev. 1.1) in 2008 for the 29 EUROSTAT member countries, thus reflecting reality just before the 2008 financial crisis. Country codes in alphabetical order according to the country name in its own language are: Belgium (BE), Cyprus (CY), Czechia (CZ), Denmark (DK), Deutchland–Germany (DE), Eesti–Estonia (EE), Eire–Ireland (IE), España–Spain (ES), France (FR), Hellas-Greece (GR), Hrvatska–Croatia (HR), Iceland (IS), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Macedonia (MK), Magyarország-Hungary (HU), Malta (MT), Netherlands (NL), Norway (NO), Österreich–Austria (AT), Portugal (PT), Romania (RO), Slovakia (SK), Suomi–Finland (FI), Switzerland (CH), Turkey (TR), United Kingdom (GB).

A key related variable is the logarithm of gross domestic product per person in EUR at current prices ("logGDP"). For the purposes of exploratory data analyses it has also been categorised as a binary variable indicating values higher or lower than the median ("Binary_GDP"). The employment composition ($D = 11$) is:

- Primary_sector (agriculture, hunting, forestry, fishing, mining, quarrying)
- Manufacturing
- Energy (electricity, gas and water supply)
- Construction
- Trade_repair_transport (wholesale and retail trade, repair, transport, storage, communications)
- Hotels_restaurants
- Financial_intermediation
- Real_estate (real estate, renting and business activities)
- Educ_admin_defense_soc_sec (education, public administration, defence, social security)
- Health_social_work

- Other_services (other community, social and personal service activities)

The aim is to construct a linear regression model to predict logGDP.

☞ Back to Index

Specific references in Appendix

[Ait86] J. Aitchison, *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability, Chapman & Hall Ltd., London (UK) [Reprinted in 2003 with additional material by The Blackburn Press], 1986, 416 p.

[Boy50] W.C. Boyd, *Genetics and the races of man: an introduction to modern physical anthropology*, Little, Brown & Co, 1950.

[Coe17] G. Coenders, J.A. Martín-Fernández and B. Ferrer-Rosell, *When relative and absolute information matter: compositional predictor with a total in generalized linear models*, Statistical Modelling **17(6)** (2017), 494–512.

[Dum19] D. Dumuid, Z. Pedisic, T.E. Stanford, J.A. Martín-Fernández, K. Hron, C. Maher, L.K. Lewis and T.S. Olds, *The Compositional Isotemporal Substitution Model: a Method for Estimating Changes in a Health Outcome for Reallocation of Time between Sleep, Sedentary Behaviour, and Physical Activity*, Statistical Methods in Medical Research **28(3)** (2019), 846–857, DOI:10.1177/0962280217737805.

[Har75] J.A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, 1975.