LABS

Compositional Analysis of Data with *CoDaPack*

Online Course

CoDa-Research Group

University of Girona
Spain

2021 (updated)

Labs presentation

This material provides an introduction to the practical aspects of compositional analysis of data, as well as few additional advanced modeling topics. The Labs comprise exercises for each section of the course *Textbook*. When an exercise corresponds to an *advanced* concept then it is accordingly highlighted in blue. The *basic* exercises included in this material have been mostly developed using the package `CoDaPack`. However, some of the exercises are *manually* solved or some of the *advanced* exercises may need an `R` package.

We hope that these materials will help any scientist to master the basic techniques to undertake CoDa analysis.

Contents

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

Activities for Section 1.1

Open the CoDaPack program. We can see three different areas: the variables area where we will see the list of our variables, the data area where we will see the complete data set and the results area where the results of our analyses will appear. The plots will be opened in an independent graphical output. Now all the windows are empty but we can add manually some data:

- Go to the *Data* ▷ *Create new table*.
  - ○ Because we want to add a 3-part composition the we write 3 in the *Number of columns* window and 2 in the *Number of rows* window. We need a minimum of two rows because the first one will always contain the names of the parts.
  - ○ Click the *Ok* button.
  - ○ CoDaPack creates an empty table. We can write for example: the names (A, B, C) in the first row and $(0.2, 0.3, 0.5)$ in the second row.
  - ○ Click the *Accept* button.
  - ○ Finally we have to enter the name for the new table, write *example*.
  - ○ Click the *Ok* button.
- Observe that CoDaPack creates a table named *example*. The names of the parts are shown in the variables area.
- We can save the data frame using the standard menus *File* ▷ *Save Workspace...* or *File* ▷ *Save as...*. By default, a CoDaPack file has the file extension "∗.cdp".

In the menu *File* we have different options for deleting tables or for clearing the output area before saving a CoDaPack file.

We can delete variables through the menu *Data* ▷ *Delete variables*. We can also add some numeric variables using the menu *Data* ▷ *Add numeric variables*. In this case, we have only to choose the names of each variable we want to add in the window *Variables* and write the corresponding numeric values in the window *Data*. With this feature we can import a variable by a simple copy-paste action. If we want to add several columns we have to do this one variable at a time. Be aware that a valid variable name consists of letters, numbers and the dot or underline characters. The variable name must start with a letter or the dot not followed by a number.

- Use the menu *Data* ▷ *Add numeric variables* and the copy-paste action (one at a time) to add the following data set:

X1	X2	X3	as variable names, and
0.1	0.2	0.7	
0.3	0.1	0.6	as data.
0.25	0.25	0.5	

Now we will start working with the `statisticiantimebudget` CoDa set described in the Appendix. The option *File* ▷ *Open Workspace...* allows us to open the data set in a new table but this option automatically closes the previous one. Alternatively the menu *File* ▷ *Add Workspace...* will open the new data set without closing the previous one. Then, in the window *Tables:* we can choose which data set we want to active.

- As we don't want to use more the previous data set, go to the *File* ▷ *Open Workspace...* menu and select the `statisticiantimebudget.cdp` file.

In the data area we can see all the variables (in columns) and all the observations (in rows). The first column, `Day`, records the ordinal (from 1 to 20) of the controlled day while the other six columns record the number of daily hours (approximated to quarter hours) devoted by the statistician to teaching (`T`), consultation (`C`), administration (`A`), research (`R`), other wakeful activities (`O`) and sleep (`S`).

The columns highlighted in yellow color means categorical variables and the columns in white color means numerical variables. We can change, if necessary, the typology of our variables:

- From numerical to categorical: go to the *Data* ▷ *Manipulate* ▷ *Numeric to categorical* menu.
  - Select for example the variable `T` from the list.
  - Click the *Accept* button.
- From categorical to numerical: go to the *Data* ▷ *Manipulate* ▷ *Categorical to numeric* menu.
  - Select again the variable `T` from the list.
  - Click the *Accept* button.

CoDaPack can import/export files in formats csv/text, xls or RData. We recommend exploring the menus *File* ▷ *Import...* and *File* ▷ *Export...*.

In addition, the menu *File* includes a *Configuration* option that allows to customize the:

- decimal character,
- output format,
- table format, and
- exportations format
- ...

At this point we do not recommend to change the default values for the configuration.

Remember that the first column, `Day`, records the ordinal (from 1 to 20) of the controlled day while the other six columns record the number of daily hours (approximated to quarter hours) devoted by the statistician to teaching (`T`), consultation (`C`), administration (`A`), research (`R`), other wakeful activities (`O`) and sleep (`S`).

It should be noted that the data is already closed since the sum of the six parts is a constant equal to 24 (i.e, $\kappa = 24$).

To close data to a new constant (e.g. $\kappa = 100$):

- Go to the *Data* ▷ *Operations* ▷ *Subcomposition / Closure* menu.
  - ○ Move the variables `T`, `C`, `A`, `R`, `O` and `S` to the *Selected data* window.
  - ○ Write 100 in the window of the *Closure* of the *Options* menu.
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `clo_T`, `clo_C`, `clo_A`, `clo_R`, `clo_O` and `clo_S` containing the percentage of time devoted to each activity. Thus, for example, the associated composition of the time budget of the first day will be equal to [14.58, 9.38, 17.71, 10.42, 26.04, 21.88].

To summarize the total daily time spent by the statistician to working activities we will add (i.e., *amalgamate*) parts `T`, `C`, `A` and `R` into a single part that we will symbolize by `TCAR`:

- Go to the *Data* ▷ *Manipulate* ▷ *Calculate new Variable* menu.
  - ○ Move the variables `T`, `C`, `A` and `R` to the *Selected data* window.
  - ○ Click the *Accept* button.
  - ○ Write $x1 + x2 + x3 + x4$ in the *Enter expression* window (Note: internally CoDaPack associates $x1$, $x2$, $x3$ and $x4$ with the variables introduced in the *Selected data* window according to the order of entry).
  - ○ Write `TCAR` into the *Enter new variable name* window.
  - ○ Click the *Ok* button.

CoDaPack creates the new variable `TCAR` equal to the sum of variables `T`, `C`, `A` and `R`. Thus, e.g. `TCAR` is equal to 12.50 (=3.50+2.25+4.25+2.50) for the first day.

Just as the original data set, the amalgamated data set [`TCAR`,`O`,`S`] is also closed ($\kappa = 24$).

To close the amalgamated data set to constant $\kappa = 100$:

- *Data* ▷ *Operations* ▷ *Subcomposition / Closure* menu.

  - ◦ Move the variables `TCAR`, `O` and `S` to the *Selected data* window.
  - ◦ Write 100 in the window of the *Closure* of the *Options* menu.
  - ◦ Click the *Accept* button.

CoDaPack creates the new variables `clo_TCAR`, `clo_O.c` and `clo_S.c` containing the percentage of the daily time devoted by the statistician to working activities, other wakeful activities and sleep. Thus, for example, 52.08, 26.04 and 21.88 are the percentages of the first day.

The ternary diagram of `clo_TCAR`, `clo_O.c` and `clo_S.c` will enable us to visualise the percentages of time spent by the statistician doing these activities.

- Go to the *Graphs* ▷ *Ternary/Quaternary Plot* menu.
  - ◦ Move the variables `clo_TCAR`, `clo_O.c` and `clo_S.c` to the *Selected data* window.
  - ◦ Click the *Accept* button.

Each point in the ternary diagram corresponds to the time budget of one day. We can use the auxiliary variable, `Day`, to identify the points in the diagram:

- Go to the *Data* ▷ *Add observation names...* menu at the top left of the graphic window.
- Select the `Day` variable.
- Click the *Accept* button.

To display the `Day` label on the ternary diagram or to remove it from of a specific point:

- Double *click* on the point.

To display/remove the `Day` labels of all the points we have to:

- go to the *Data* menu at the top left of the graphic window to respectively activate/deactivate the *Show observation names* option.

To get a rough idea from the distance of one point to the sides of the ternary diagram:

- Activate the *Grid* option at the border of the graphic window.

The lines of the grid divide the sides of the triangle into ten parts. Thus, for example, on the fourth day (`Day`=4) the statistician spent slightly more than 50% of his/her time to working activities, between 10% and 20% of his/her time to other wakeful activities and a little more than 30% to sleep (Figure 1.1.a).

You can move the *zoom* slider (at the border of the graphic window) to increase or decrease the size of the diagram.

It is worth noting that to represent a ternary diagram the previous closure of the involved parts is not necessary.

Finally, to save the graph:

- Go to the menu *Files* ▷ *Save as...* at the top left of the graphic window and you can choose different formats.

If we are only interested in analysing how the statistician distributes his working time, we should focus our attention on the subcomposition [`T`,`C`,`A`,`R`]. We can represent this subcompositon in a quaternary plot. As it is not necessary to have closed data to represent it, it could be interesting apply here the closure operation in order to see the percentage of total working time spend to each working category. To close the parts of this subcomposition to constant $\kappa = 100$:

- Go to the *Data* ▷ *Operations* ▷ *Subcomposition / Closure* menu.
  - ○ Move the variables `T`,`C`,`A` and `R` to the *Selected data* window.
  - ○ Write 100 in the window of the *Closure* of the *Options* menu.
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `clo_T.c`, `clo_C.c`, `clo_A.c` and `clo_R.c`, containing the percentage of daily working time devoted by the statistician to teaching, consultation, administration and research. Thus, for example, 28.00%, 18.00%, 34.00% and 20.00% are the respective percentages corresponding to the first day.

The tetrahedral diagram of `clo_T.c`, `clo_C.c`, `clo_A.c` and `clo_R.c` will allow us to visualise the percentages of working time spent by the statistician on teaching, consultation, administration and research.

- Go to the *Graphs* ▷ *Ternary/Quaternary Plot* menu.
  - ○ Move the variables `clo_T.c`, `clo_C.c`, `clo_A.c` and `clo_R.c` to the *Selected data* window.
  - ○ Click the *Accept* button.

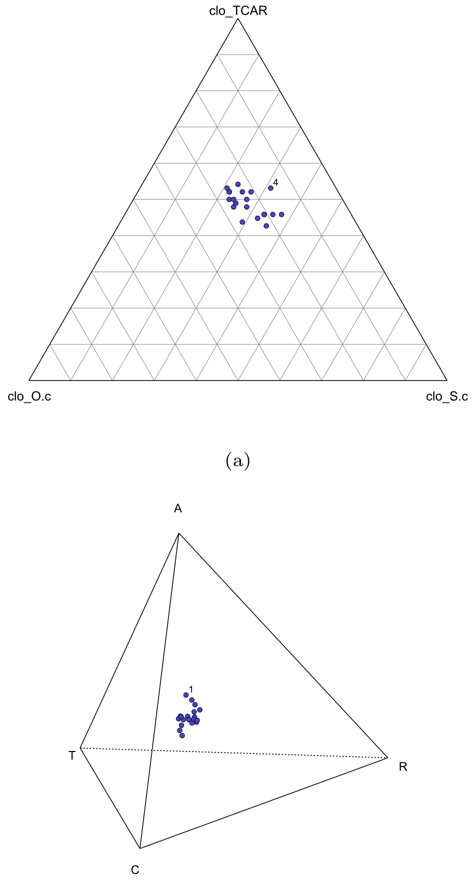
We would obtain the same graph from the *non closed* variables `T`, `C`, `A` and `R`. As before, we can use the `Day` variable to label the points. In addition, dragging the cursor over the diagram allows us to change the orientation of the tetrahedron (Figure 1.1b)

Let us calculate the covariance and correlation matrices of the `clo_TCAR`, `clo_O.c` and `clo_S.c` variables.

- Go to *Statistics* ▷ *Classical statistics summary*.
  - ○ Move the variables `clo_TCAR`, `clo_O.c` and `clo_S.c` to the *Selected data* window.
  - ○ Activate *Correlation Matrix* and *Covariance Matrix* in the *Options* menu (deactivate the remaining options).
  - ○ Click the *Accept* button.

We obtain the following two matrices:

```
Correlation:
            clo_TCAR   clo_O.c   clo_S.c
  clo_TCAR    1.0000    0.1253   -0.7527
   clo_O.c    0.1253    1.0000   -0.7475
   clo_S.c   -0.7527   -0.7475    1.0000
```



(a)

(b)

Figure 1.1. Statistician's time budget. (a) Ternary diagram of the daily time devoted to work (`TCAR`=`T`+`C`+`A`+`R`), to other wakeful activities (`O`) and to sleep (`S`) (e.g. on `Day`=4 the statistician spent 53.13%, 15.63% and 31.25% of his time to these activities, respectively). (b) Tetrahedral diagram of the daily working time devoted to teaching (`T`), consultation (`C`), administration (`A`) and research (`R`) activities e.g. on `Day`=1 the statistician spent 28%, 18%, 34% and 20% of his working time to `T`, `C`, `A` and `R`, respectively).

Covariance:

	clo_TCAR	clo_O.c	clo_S.c
clo_TCAR	11.4448	1.4214	-12.8662
clo_O.c	1.4214	11.2413	-12.6628
clo_S.c	-12.8662	-12.6628	25.5290

Note how the sum of the elements in the same row (or column) in the covariance matrix is effectively zero. Thus at least one entry in each row (or column) of the correlation matrix must be negative. For this reason the (crude) correlations in this matrix are considered to be spurious because they cannot be interpreted in the standard form. All this is due to the fact that the sum of the elements of variables `clo_TCAR`, `clo_O.c` and `clo_S.c` is constant ($\kappa = 100$) in each row.

When one wants to exit the CoDaPack program, we save the data set together with the new variables:

- Go to the *File* ▷ *Save as...* menu.
  - ○ Save the workspace with the name `statisticiantimebudget01.cdp`.
- Go to the *File* ▷ *Quit CoDaPack* menu and confirm the exit of the program.

1.2. Principles of CoDa analysis

Activities for Section 1.2

We continue with the analysis of the `statisticiantimebudget` CoDa set (Appendix) began in the previous sections.

- Open the `statisticiantimebudget01.cdp` file from the CoDaPack package going to the *File* ▷ *Open Workspace...* menu.

Let us calculate the correlation matrices of the closed ($\kappa = 100$) data set $\mathcal{C}[\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{R}, \mathtt{O}, \mathtt{S}]$ (represented by the variables `clo_T`, `clo_C`, `clo_A`, `clo_R`, `clo_O` and `clo_S`, and of the subcomposition $\mathcal{C}[\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{R}]$ (represented by the variables `clo_T.c`, `clo_C.c`, `clo_A.c` and `clo_R.c`).

- Go to *Statistics* ▷ *Classical statistics summary*.
  - ○ Move the variables `clo_T`, `clo_C`, `clo_A`, `clo_R`, `clo_O` and `clo_S` to the *Selected data* window.
  - ○ Activate *Correlation Matrix* in the *Options* menu (deactivate the remaining options).
  - ○ Click the *Accept* button.
- Repeat the same procedure with variables `clo_T.c`, `clo_C.c`, `clo_A.c` and `clo_R.c`.

We obtain the following two correlation matrices:

Correlation:

	`clo_T`	`clo_C`	`clo_A`	`clo_R`	`clo_O`	`clo_S`
`clo_T`	1.0000	-0.0581	-0.5501	0.4062	0.2582	-0.5657
`clo_C`	-0.0581	1.0000	-0.0538	**0.0260**	0.0487	-0.3085
`clo_A`	-0.5501	-0.0538	1.0000	-0.0833	-0.0230	-0.0369
`clo_R`	0.4062	**0.0260**	-0.0833	1.0000	-0.1571	-0.3812
`clo_O`	0.2582	0.0487	-0.0230	-0.1571	1.0000	-0.7475
`clo_S`	-0.5657	-0.3085	-0.0369	-0.3812	-0.7475	1.0000

Correlation:

	`clo_T.c`	`clo_C.c`	`clo_A.c`	`clo_R.c`
`clo_T.c`	1.0000	-0.4449	-0.7731	0.0190
`clo_C.c`	-0.4449	1.0000	-0.0054	**-0.4194**
`clo_A.c`	-0.7731	-0.0054	1.0000	-0.2800
`clo_R.c`	0.0190	**-0.4194**	-0.2800	1.0000

Note, for example, that whereas the correlation between `clo_C` and `clo_R` is positive ($r = 0.0260$) in the composition $\mathcal{C}[\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{R}, \mathtt{O}, \mathtt{S}]$ it is negative ($r = -0.4194$) in the subcomposition $\mathcal{C}[\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{R}]$. This fact highlights the subcompositional incoherence of the standard statistical analysis of closed data.

1.3. Perturbation and power operations in the simplex

Activities for Section 1.3

A. In a culture there are three types —A, B and C— of bacteria. The bacterium A increases by multiplying its amount by 2 each day. Similarly, the amounts of bacteria B and C are multiplied by 3 and 5 each day, respectively.
Today, the percentages of A, B and C in the culture are 50%, 30% and 20%, respectively. Therefore, tomorrow the percentages of $[A, B, C]$ will be equal to $\mathcal{C}[50 \times 2, 30 \times 3, 20 \times 5] = [34.5, 31.0, 34.5]$. Observe that

$$[34.5, 31.0, 34.5] = [2, 3, 5] \oplus [50, 30, 20] \ .$$

In general, if we represent by $\mathbf{x}_0$ the percentages of $[A, B, C]$ on time $t = 0$ (i.e. $\mathbf{x}_0 = [50, 30, 20]$) and by $\mathbf{p}$ the composition $\mathcal{C}[2, 3, 5]$, the percentages of $[A, B, C]$ on time t (days) —represented by $\mathbf{x}_t$— can be interpreted as

$$\mathbf{x}_t = \underbrace{\mathbf{p} \oplus \mathbf{p} \oplus \ldots \oplus \mathbf{p}}_{t} \oplus \mathbf{x}_0 = (t \odot \mathbf{p}) \oplus \mathbf{x}_0 \ .$$

In Table 1.1 we can see the evolution of percentages of $[A, B, C]$ from time $t = 0$ to $t = 10$.

Table 1.1. Evolution of percentages of bacteria $[A, B, C]$ in the culture, from time $t = 0$ to $t = 10$.

	time (days)										
%	0	1	2	3	4	5	6	7	8	9	10
A	50.00	34.48	20.62	10.78	5.09	2.24	0.95	0.39	0.16	0.06	0.03
B	30.00	31.03	27.84	21.83	15.45	10.21	6.48	4.01	2.45	1.49	0.90
C	20.00	34.48	51.55	67.39	79.47	87.55	92.57	95.59	97.39	98.45	99.08

Now we will go to CoDaPack to represent in a ternary diagram the evolution of percentages of bacteria $[A, B, C]$ in the culture, from time $t = 0$ to $t = 10$. The Excel® file `bacteria.xls` contains the data of Table 1.1. We will import this file from CoDaPack.

- Import the file `bacteria.xls` from the CoDaPack package going to the *File* ▷ *Import* ▷ *Import XLS Data...* menu. CoDaPack gives us a warning about the variables' names (click *Accept*). Select the excel file and CoDaPack shows us the spreadsheet options that allow us to change the data frame name, to indicate if our file has headings or the value/ prefix for the non-available/non-detected data (click *Ok*). Finally, in the window *Choose one sheet*, do a double click on `evolution`.

The first column `t` records the time (days), from 0 to 10, and the columns `A`, `B` and `C` the corresponding percentages of $[A, B, C]$.

We represent the 11 compositions $[A, B, C]$ in a ternary diagram to visualise their evolution over time.

- Go to the *Graphs* ▷ *Ternary/Quaterny Plot* menu.
  - Move the variables `A`, `B` and `C` to the *Selected data* window.
  - Click the *Accept* button.
  - Activate the *Grid* option on the border of the graphic window.
  - Go to the *Data* ▷ *Add observation names...* menu at the top left of the graphic window.
  - Select the variable `t`.
  - Click the *Accept* button.
  - Go again to the *Data* menu and activate the *Show observation names* option.

The ternary diagram highlights the *compositional* linear trend of the evolution of $[A, B, C]$ from $[50, 30, 20]$ (at $t = 0$) to $[0, 0, 100]$ (at $t = +\infty$).

B. We now use the `householdbudget` CoDa set described in the Appendix. This data set records the expenditures of four commodity groups of forty single persons (twenty men and twenty women) living alone in rented accommodation.

- Open the file `householdbudget.cdp` from the CoDaPack package going to the *File* ▷ *Open Workspace...* menu.

The first column `Id` records the identification number of the individuals (from 1 to 40) in the sample. The second column (`Sex`) records the gender of the individuals (m = man, w =woman). The other four columns record the total amount (in $US) of the weekly expenditures in: 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`).
It should be noted that the data is not closed.

- Use the *Data* ▷ *Operations* ▷ *Subcomposition/Closure* menu to close the variables `Hous`, `Food`, `Serv` and `Others` to the constant $\kappa = 100$.

The new variables give us the expenditure percentages on housing, foodstuffs, services and other goods of each individual. Thus, for example, the first person (`Id`=1) spends 32.44% on `Hous`, 38.58% on `Food`, 18.99% on `Serv` and 9.99% on `Others`.
The tetrahedral diagram of `Hous`, `Food`, `Serv` and `Others` will allow us to visualise the expenditure percentages (by gender) on the four commodity groups.

- Go to the *Graphs* ▷ *Ternary/Quaternary plot* menu.
    - ○ Move the variables `Hous`, `Food`, `Serv` and `Others` to the *Selected data* window (remember that it is not necessary to choose the closed composition).
    - ○ Select the variable `Sex` in the *Groups* window.
    - ○ Click the *Accept* button.

The tetrahedral diagram (Figure 1.2) highlights the differences in the expenditure patterns of both sexes.

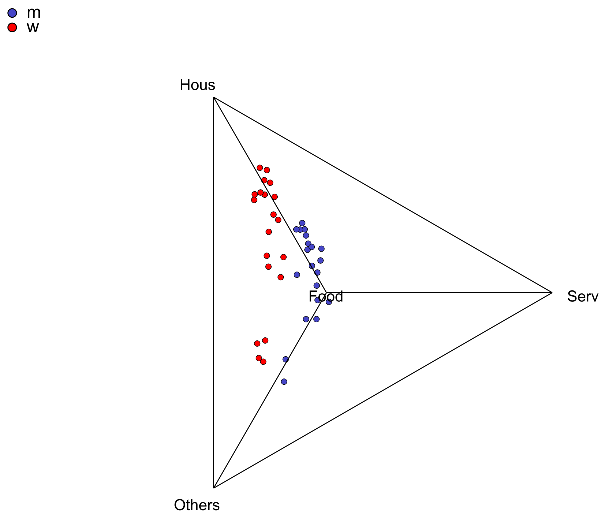


Figure 1.2. Quaternary diagram of the expenditures in each of the four commodity groups by gender (men in blue, women in red).

Suppose that over the last year, prices have increased 90% in housing, 50% in foodstuffs, 30% in services and 10% in other goods. If we assume that the spending habits of the people have not changed, the new expenditure percentages on the four commodity items can be obtained applying the perturbation $\mathbf{p} = \mathcal{C}[1.9, 1.5, 1.3, 1.1] = [0.328, 0.259, 0.224, 0.190]$ to each of the individual compositions.

- Go to the *Data* ▷ *Operations* ▷ *Perturbation* menu.
  - ○ Move the variables `Hous`, `Food`, `Serv` and `Others` to the *Selected data* window.
  - ○ Write 1.9 1.5 1.3 1.1 in the *Perturbation* window of the *Options* menu.
  - ○ Write 100 in the *Closure to* window of the *Options* menu.
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `Hous.x.1.9`, `Food.x.1.5`, `Serv.x.1.3` and `Others.x.1.1` containing the perturbed compositions. Thus, for example, the new expenditure percentages of the first person (`Id`=1) are 39.72% on `Hous`, 37.29% on `Food`, 15.91% on `Serv` and 7.08% on `Others`.

We can compare the tetrahedral diagram of compositions [`Hous`, `Food`, `Serv`, `Others`] and [`Hous.x.1.9`,`Food.x.1.5`,`Serv.x.1.3`,`Others.x.1.1`].

- Go to the *Graphs* ▷ *Ternary/Quaternary Plot* menu.
  - ○ Select and move the variables `Hous.x.1.9`, `Food.x.1.5`, `Serv.x.1.3` and `Others.x.1.1` to the *Selected data* window.
  - ○ Select the `Sex` variable in the *Groups* window.
  - ○ Click the *Accept* button.

The comparison of the two tetrahedral diagrams highlights that the two data sets (original and perturbed) have identical variation patterns. However, the points of the perturbed data set seem to be moved towards the `Hous.x.1.9` vertex. This displacement of cloud of points can be better viewed if we represent the ternary diagrams of any subcomposition involving three of the four components.

- Go to the *Graphs* ▷ *Ternary/Quaternary plot* menu.
  - ○ Move the variables `Hous`, `Serv` and `Others` to the *Selected data* window.
  - ○ Select the `Sex` variable in the *Groups* window.
  - ○ Click the *Accept* button.
- Repeat the same procedure with variables `Hous.x.1.9`, `Serv.x.1.3` and `Others.x.1.1`.

The comparison of the two ternary diagrams highlights that, in general, the perturbation increases the percentage of housing expenditures at the expense of the decreasing percentages of expenditures on services and other goods.

1.4. Original logratio analysis: alr and clr tranformations

Activities for Section 1.4

We will now work with the `bloodMN` CoDa set described in the Appendix. This data set records the information on the number of individuals with M, N and MN blood types observed in samples coming from different ethnic groups around the world.

- Open the `bloodMN.cdp` file from the CoDaPack package going to the *File* ▷ *Open Workspace...* menu.

The variables are:

- `population`: ethnic group origin of the sample.
- `popul_code`: numeric code of ethnic group.
- `total`: total sample size.
- `MN`: number of individuals in the sample with blood type MN (genotype MN, phenotype MN).
- `MM`: number of individuals in the sample with blood type M (genotype MM, phenotype M).
- `NN`: number of individuals in the sample with blood type N (genotype NN, phenotype N).

As the sample sizes are different, we start calculating the percentages of each blood type in each of the samples closing the data set to constant $\kappa = 100$.

- Go to the *Data* ▷ *Operations* ▷ *Subcomposition/Closure* menu and close ($\kappa = 100$) the variables `MN`, `MM` and `NN`.

The new variables `clo_MN`, `clo_MM` and `clo_NN` contain the blood type percentages for each sample [e.g. the percentages of the first sample (*ainu*, 1) are 50.20, 17.86 and 31.94, respectively].

Now, go to the *Graphs* ▷ *Ternary/Quaternary Plot* menu to plot the corresponding ternary diagram. Thus, we get a graphical overview of how the `MN`, `MM` and `NN` percentages are distributed. Mark the option *Grid* on the bottom of the window graphics to draw a grid inside the ternary diagram.

We can observe how the variable `MN` varies from 10 to 60%, while the `MM` and `NN` variables vary within the ranges 0-90% and 0-70%, approximately. Despite this variability, there is an evident pattern in this variation.

We want to identify the samples on the ternary diagram labelling the points of the graph:

- Go to the *Data* ▷ *Add observations names...* menu on the top of the graphic window.
  - Select the `popul_code` variable.
  - Click the *Accept* button.

Now, double click on a point of the graph to display its label. If you wish to label all the points, go to the *Data* ▷ *Show observations names* menu on the top of the graphic window. Alternatively you can select the `population` variable in the menu *Data* ▷ *Add observations names...* to label the points with names instead of numbers.

The three points closer to the `MN-MM` side of the triangle correspond to samples labelled 2 [*aleuts*], 21 [*eskimo*] and 4 [*amer_indians_navaho*], whereas the points closer to the `MN-NN` side correspond to samples labeled with 23 [*fijians*], 11 [*australian_aborigines*] and 39 [*papuans*].

Now we will apply the additive logratio transformation (*alr*) to our data. We consider the part/variable `MN` as the common denominator of the logratios in the *alr*-transformation.

- Go to the *Data* ▷ *Transformations* ▷ *ALR* menu.
  - Select the variables `MM`, `NN` and `MN` (in this order) from the *Available data* window and move them into the *Selected data* window [Please note that the last selected variable will be the common denominator of the logratios].
  - As we want to compute the additive logratio transformed vectors, we have to select the *Raw-ALR* option (note that the *ALR-Raw* option will compute the inverse transformation denoted al alr^{-1}).
  - Click the *Accept* button.

CoDaPack creates the new variables `alr.MM_MN` and `alr.NN_MN` corresponding to the logratios ln(`MM`/`MN`) and ln(`NN`/`MN`), respectively.

Now, we will represent a Cartesian diagram of our data from their *alr*-values using the variable `MN` as the common denominator of the logratios.

- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu.
  - Select the variables `alr.MM_MN` and `alr.NN_MN` from the *Available data* window and move them into the *Selected data* window.
  - Select *Set same scale* into *Options* window to use the same scale on the two axes of the Cartesian representation.
  - Click the *Accept* button.

The graph highlights the linear pattern of the *alr*-transformed data in correspondence with the curved pattern previously observed in the ternary plot. Almost all points have a negative ln(`NN`/`MN`) coordinate. This means that the proportion of individuals with genotype MN is greater than those with genotype NN in almost all samples.

Note that this graph is different from the ternary/quaternary graphs. This graph is generated from the R program and the options are not the same. For example, we cannot save this graph, for the moment we can only export it as a SVG format (see *File* ▷ *Export* ▷ *Export As SVG* at the top left menu of the graphic window). Also, it does not allow to identify the points with an auxiliary variable.

Nevertheless, using both, the graph and the numerical values contained in the data area part, we can see that sample 11 [*australian_aborigines* —the point at the upper left corner of the graph— is the one with the lowest proportion of individuals with genotype MM in comparison to individuals with genotype MN. In contrast, samples 21 [*eskimo*] and 4 [*amer_indians_navaho*,] —the two points located in the lower right corner of the graph— are those with a higher proportion of individuals with genotype MM relative to the proportion of individuals with genotype MN.

If we want to compare the proportions of individuals with MM and NN genotypes, we have to calculate the *alr*-values of the data set using one of these two variables as the common denominator of the logratios.

- Calculate the *alr*-scores of the data set using `NN` as the common denominator of the logratios
- Redraw the *alr*-plot.

It can be observed that almost all points (except for five points) have a positive ln(`MM`/`NN`) coordinate. This means that the proportion of individuals with the MM genotype is greater than those with the NN genotype in almost all samples. Samples 21 [*eskimo*] and 4 [*amer_indians_navaho*] —the two points in the upper right corner of the Cartesian plot `alr.MN_NN` vs `alr.MM_NN`— are those with the highest relative difference between both proportions.

Now we will apply the centred logratio transformation (*clr*) to our data.

- Go to the *Data* ▷ *Transformations* ▷ *CLR* menu.
  - ○ Select the variables `MN`, `MM` and `NN` from the *Available data* window and move them into the *Selected data* window.
  - ○ As we want to compute the centred logratio transformed vectors, we have to select the *Raw-CLR* option (note that the *CLR-Raw* option will compute the inverse transformation denoted al clr^{-1}).
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `clr.MN`, `clr.MM` and `clr.NN` corresponding to the centred logratios

$$\begin{aligned}\mathrm{clr}\,\mathtt{MN} &= \ln \frac{\mathtt{MN}}{(\mathtt{MN}\cdot\mathtt{MM}\cdot\mathtt{NN})^{1/3}}\ ,\\ \mathrm{clr}\,\mathtt{MM} &= \ln \frac{\mathtt{MM}}{(\mathtt{MN}\cdot\mathtt{MM}\cdot\mathtt{NN})^{1/3}}\ ,\\ \mathrm{clr}\,\mathtt{NN} &= \ln \frac{\mathtt{NN}}{(\mathtt{MN}\cdot\mathtt{MM}\cdot\mathtt{NN})^{1/3}}\ .\end{aligned}$$

Thus, for example, the *clr*-coordinates of the second sample (*aleuts*, 2) of the data set are

$$\mathrm{clr}\,\mathtt{MN} = 0.48 \quad , \quad \mathrm{clr}\,\mathtt{MM} = 1.31 \quad , \quad \mathrm{clr}\,\mathtt{NN} = -1.79\ .$$

The signs and absolute values of these coordinates reported that the proportion of MN and MM genotypes in this composition are higher than the (geometric) average of the three genotypic proportions, while the proportion of genotype NN is much smaller than this average. In fact, these proportions (percentages) are equal to $[\mathtt{MN}, \mathtt{MM}, \mathtt{NN}] = [29.55, 67.42, 3.03]$.

CoDaPack allows the Cartesian representation of our data from their *clr*-coordinates provided that compositions have at most three parts.

- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu.
  - ○ Select the variables `clr.MN`, `clr.MM` and `clr.NN` from the *Available data* window and move them into the *Selected data* window.
  - ○ Select the *Set same scale* into the `Options` window.
  - ○ Click the *Accept* button.

Once again, the *linearity* of the variation pattern of the data set is evident from the resulting *clr*-plot.

If you exit the CoDaPack program, save the workspace under the name `bloodMN01.cdp`.

1.5. The algebraic-geometric structure of the simplex

Activities for Section 1.5

We continue using the CoDaPack package to analyse the `bloodMN` CoDa set (see Appendix) that we began to analyse in the previous Activities Section.

- If necessary, retrieve the `bloodMN01.cdp` file from the CoDaPack package.

Now, we will apply the perturbation $\mathbf{p} = \mathcal{C}[3, 1, 1]$ to [`MN, MM, NN`] compositions to see graphically the changes caused by the perturbation on the original data set and also on the *alr* and *clr* transformed data sets.

- Go to the *Data* ▷ *Operations* ▷ *Perturbation* menu.
  - Select the variables `MN`, `MM` and `NN` from the *Available data* window and move them into the *Selected data* window.
  - Write 3 1 1 (in this order and without commas to separate the numbers) in the *Perturbation* window. These are the components of the perturbation vector $\mathbf{p}$. Note that it is not necessary to introduce the components 0.6, 0.2 and 0.2 of the closed vector $\mathcal{C}[3, 1, 1] = [0.6, 0.2, 0.2]$.
  - Write 100 in the *Closure to* window.
  - Click the *Accept* button.

CoDaPack creates the new variables `MN.×.3.0`, `MM.×.1.0` and `NN.×.1.0`, corresponding to the value of parts of the perturbed data set. Observe that the ratios x_1/x_2 and x_1/x_3 in the perturbed data set are triple that of the original ones because $p_1/p_2 = p_1/p_3 = 3/1 = 3$, whereas the ratio x_2/x_3 does not change because $p_2/p_3 = 1/1 = 1$.

- To visualise the effect of the perturbation, plot the ternary diagram of the new variables `MN.×.3.0`, `MM.×.1.0` and `NN.×.1.0`. Compare it with the ternary diagram of the original variables `MN, MM` and `NN`.

The comparison of the two ternary diagrams highlights that the two data sets (original and perturbed) have identical variation patterns. However, the points of the perturbed data set have moved towards the `MN.×.3.0` vertex.

Let's see now how the perturbations in the simplex correspond to the translations in the real space when the compositions are expressed in *alr*-values.

- Go to the *Data* ▷ *Transformations* ▷ *ALR* menu to calculate the *alr*-transformation of the `MM.×.1.0`, `NN.×.1.0` and `MN.×.3.0` [Select the variables in this order].
- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu to represent the *alr*-plot of the new variables `alr.MM.×.1.0_MN.×.3.0` and `alr.NN.×.1.0_MN.×.3.0` [Select *Set same scale* in the *Options* window]. Compare it with the scatterplot of the initial *alr* variables `alr.MM_MN` and `alr.NN_MN`.

Comparing the Cartesian representations of the *alr*-values $(\ln(\texttt{MM/MN}), \ln(\texttt{NN/MN}))$ and $(\ln\{(1*\texttt{MM})/(3*\texttt{MN})\}, \ln\{(1*\texttt{NN})/(3*\texttt{MN})\})$ we see two identical 'clouds' of points. The only difference are the values on the axis. Observe that we can move from one to the other by a translation. It is easy to prove that $(\ln\{(1\cdot\texttt{MM})/(3\cdot\texttt{MN})\}, \ln\{(1\cdot\texttt{NN})/(3\cdot\texttt{MN})\}) = (\ln\texttt{(MM/MN)}, \ln\texttt{(NN/MN))} - (\ln 3, \ln 3)$.

The same applies when the compositions are expressed in *clr*-coordinates.

- Go to the *Data* ▷ *Transformations* ▷ *CLR* menu to calculate the *clr*-coordinates of the variables `MN.×.3.0`, `MM.×.1.0` and `NN.×.1.0`.
- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu to represent the *clr*-plot of the new variables `clr.MN.×.3.0`, `clr.MM.×.1.0` and `clr.NN.×.1.0` [Select *Set same scale* in the *Options* window].
- Compare the resulting representation with the *clr*-plot of the original variables `MN, MM` and `NN`.

Again the shape of the clouds of points is identical. Using the 3D scatterplot is more difficult to appreciate the movement. We suggest to use the corresponding three 2D scatterplots to see more clearly the movement.

We suggest deleting the newly created variables:

- Go to the *Data* ▷ *Delete variables* menu.
  - ○ Highlight the variables to remove.
  - ○ Click the *Accept* button.

Now, we apply the power operator (with $a = 2$) to the `bloodMN` data set.

- Go to the *Data* ▷ *Operations* ▷ *Power transformation* menu.
  - ○ Select the variables `MN`, `MM` and `NN` from the *Available data* window and move them into the *Selected data* window.
  - ○ Write 2 into the *Power* box in the *Options* window.
  - ○ Write 100 into the *Closure to* box in the *Options* window.
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `MN.pt.2.0`, `MM.pt.2.0` and `NN.pt.2.0`, corresponding to the parts of the powered compositions $2 \odot \mathbf{x}$ (closed to 100). Note that the ratios x_i/x_j in the powered data set are equal to the square of the corresponding ratios in the original data set (because $a = 2$).

- To visualise the effect of the "$2\,\odot$" operator, plot the ternary diagram of the new variables `MN.pt.2.0`, `MM.pt.2.0` and `NN.pt.2.0`. Compare it with the ternary diagram of the original variables `MN, MM` and `NN`.

The comparison of the two ternary diagrams highlights that the two data sets (original and *powered*) have similar variation patterns. However, the points of the powered data set have moved towards the `MNˆ2.0` vertex and, moreover, their variability has been increased.

- Go to the *Data* ▷ *Transformations* ▷ *ALR* menu to apply the *alr*-transformation to the `MM.pt.2.0`, `NN.pt.2.0` and `MN.pt.2.0` [Select the variables in this order].
- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu to represent the *alr*-plot of the new variables `alr.MM.pt.2.0_MN.pt.2.0`, `alr.NN.pt.2.0_MN.pt.2.0` [Select *Set same scale* in the *Options* window].

- Compare the resulting plot with the *alr*-plot (ln `(MM/MN)` vs ln `(NN/MN)`).

- Go to the *Data* ▷ *Transformations* ▷ *CLR* menu to calculate the *clr*-coordinates of the `MN.pt.2.0`, `MM.pt.2.0` and `NN.pt.2.0`.
- Go to the *Graphs* ▷ *Scatterplot 2D/3D* menu to represent the *clr*-plot of the new variables `clr.MN.pt.2.0`, `clr.MM.pt.2.0` and `clr.NN.pt.2.0` [Select *Set same scale* in the *Options* window].
- Compare the resulting plot with the *clr*-plot of the original data (`MN`, `MM`, `NN`).

Apparently, the *alr* and *clr* plots of the powered data set look identical to the corresponding *alr* and *clr* plots of the original set. In fact, all coordinates have been multiplied by two since it holds that $\mathrm{alr}\,(2\odot\mathbf{x}) = 2\cdot(\mathrm{alr}\,\mathbf{x})$ and $\mathrm{clr}\,(2\odot\mathbf{x}) = 2\cdot(\mathrm{clr}\,\mathbf{x})$.

- Repeat the process by applying the "0.5⊙" operator to the original data set.

Now we compare the (MN,MM,NN) blood composition of the ethnic groups 11 (*australian_aborigines*) and 39 (*papuans*). We need to calculate the perturbation difference between the $\mathbf{x}_{11}$ and $\mathbf{x}_{39}$ blood compositions of the two ethnic groups, that is,

$$\begin{aligned}\mathbf{p}_{11-39} = \mathbf{x}_{11} \ominus \mathbf{x}_{39} = &\ [30.50, 2.25, 67.25] \ominus [24.00, 7.00, 69.00] \\ = &\ \mathcal{C}\left[\tfrac{30.50}{24.00}, \tfrac{2.25}{7.00}, \tfrac{67.25}{69.00}\right] = [49.51, 12.52, 37.97]\ .\end{aligned}$$

Therefore, we can move from $\mathbf{x}_{39}$ to $\mathbf{x}_{11}$ by applying the perturbation $\mathbf{p} = [p_{MN}, p_{MM}, p_{NN}] = [49.51, 12.52, 37.97]$.

As the quotient p_{MN}/p_{MM} is equal to $49.51/12.52 = 3.95$, the ratio $(\mathtt{MN/MM})_{11} = 3.95 \times (\mathtt{MN/MM})_{39}$, that is, the ratio between the frequencies of genotypes MN and MM in the ethnic group 11 (*australian_aborigines*) is 3.95 times the same ratio in the ethnic group 39 (*papaus*). Similarly, $(\mathtt{MN/NN})_{11} = 1.30 \times (\mathtt{MN/NN})_{39}$ and $(\mathtt{MM/NN})_{11} = 0.33 \times (\mathtt{MM/NN})_{39}$, because $49.51/37.97 = 1.30$ and $12.52/37.97 = 0.33$.

Likewise, the perturbation difference between the ethnic groups 23 (*fijians*) and 39 (*papuans*) is equal to:

$$\begin{aligned}\mathbf{p}_{23-39} = \mathbf{x}_{23} \ominus \mathbf{x}_{39} = &\ [44.50, 11.00, 44.50] \ominus [24.00, 7.00, 69.00] \\ = &\ \mathcal{C}\left[\tfrac{44.50}{24.00}, \tfrac{11.00}{7.00}, \tfrac{44.50}{69.00}\right] = [49.55, 38.61, 15.84]\ .\end{aligned}$$

Therefore, $(\mathtt{MN/MM})_{23} = 1.18\times(\mathtt{MN/MM})_{39}$, $(\mathtt{MN/NN})_{23} = 2.87\times(\mathtt{MN/NN})_{39}$ and $(\mathtt{MM/NN})_{23} = 2.44 \times (\mathtt{MM/NN})_{39}$.

We can use the Aitchison distance to measure the difference between two ethnic groups with regard to their (MN, MM, NN) blood composition. It is dangerous to estimate the differences from the ternary diagram, since the distance does not correspond to the Euclidean distance on this graph. However, the estimates can be visualised with the 3D scatterplot (or better the 2D scatterplots) from the *clr*-vector. Note that distances cannot be visualised using the 2D scatterplot of the *alr*-vector because the *alr* transformation does not preserve the distances, it is not an isometric transformation.

CoDaPack does not have a menu to compute the Aitchison distance between compositions. In practice, the distance between two compositions $\mathbf{x}$ and $\mathbf{y}$ can be calculated from the Euclidean distance between the corresponding $\mathrm{clr}\,\mathbf{x}$ and $\mathrm{clr}\,\mathbf{y}$ transformed vectors. Let us calculate the distance between samples 11 and 39 from the variables `clr.MN`, `clr.MM` and `clr.NN` in the `Data Set` window of CoDaPack.

- ⋆ Coordinates of compositions 11 (*australian_aborigines*), 23 (*fijians*) and 39 (*papuans*):

$$\begin{aligned}
\mathbf{x}_{11} &= [30.50, 2.25, 67.25] &\rightarrow\quad \mathrm{clr}\,\mathbf{x}_{11} &= [0.61, -2.00, 1.40]\\
\mathbf{x}_{23} &= [44.50, 11.00, 44.50] &\rightarrow\quad \mathrm{clr}\,\mathbf{x}_{23} &= [0.47, -0.93, 0.47]\\
\mathbf{x}_{39} &= [24.00, 7.00, 69.00] &\rightarrow\quad \mathrm{clr}\,\mathbf{x}_{39} &= [0.06, -1.17, 1.11]
\end{aligned}$$

- ⋆ Aitchison distances:

$$\mathrm{d_a}(\mathbf{x}_{11}, \mathbf{x}_{39}) = \left((0.61 - 0.06)^2 + (-2.00 + 1.17)^2 + (1.40 - 1.11)^2\right)^{1/2} = 1.03$$
$$\mathrm{d_a}(\mathbf{x}_{23}, \mathbf{x}_{39}) = \left((0.47 - 0.06)^2 + (-0.93 + 1.17)^2 + (0.47 - 1.11)^2\right)^{1/2} = 0.80\ .$$

Remember that the distance between two compositions $\mathbf{x}$ and $\mathbf{y}$ could also be calculated from the Aitchison norm of the composition $\mathbf{x} \ominus \mathbf{y}$. Therefore, in our case, $\mathrm{d_a}(\mathbf{x}_{11}, \mathbf{x}_{39}) = \|\mathbf{p}_{11-39}\|_\mathrm{a}$ and $\mathrm{d_a}(\mathbf{x}_{23}, \mathbf{x}_{39}) = \|\mathbf{p}_{23-39}\|_\mathrm{a}$. Moreover, the norms $\|\mathbf{p}_{11-39}\|_\mathrm{a}$ and $\|\mathbf{p}_{23-39}\|_\mathrm{a}$ could be calculated from the Euclidean norm of the *clr* transformed vectors $\mathrm{clr}\,\mathbf{p}_{11-39}$ and $\mathrm{clr}\,\mathbf{p}_{23-39}$.

Therefore, from a compositional point of view and considering the compositional nature of the data, sample 39 (*papuans*) differs from sample 11 (*australian_aborigines*) a little more than it differs from sample 23 (*fijians*) with regard to their (MN, MM, NN) blood composition. In contrast, the Euclidean distance $\mathrm{d}(\mathbf{x}_{11}, \mathbf{x}_{39})$ (equal to 8.24) is much lower than the Euclidean distance $\mathrm{d}(\mathbf{x}_{23}, \mathbf{x}_{39})$ (equal to 32.19).

If you exit the CoDaPack program, save the workspace under the name `bloodMN02.cdp`.

1.6. Compositional-linear dependence, basis and coordinates

Activities for Section 1.6

In this section we use again the `householdbudget` data set described in the Appendix. This data set records the expenditures of four commodity groups of forty single persons (twenty men and twenty women) living alone in rented accommodation.

- Open the file `householdbudget.cdp` from the CoDaPack package going to the *File* ▷ *Open Workspace...* menu.

A *clr-biplot* is a reduced-dimensionality representation of a CoDa set. That is, given a CoDa set where the samples have more than three parts ($D > 3$), the *clr*-scores are calculated and we create an *approximate* representation of the *clr*-scores set in a scatterplot. The idea consists of creating a new set of $D-1$ *clr*-variables (*principal* *components*) where few of them account for a large proportion of the variance of the CoDa set. Next Chapter, which is devoted to exploratory techniques, includes a detailed introduction of the *clr*-biplot representation.

To create the *clr*-biplot of the `householdbudget` data set:

- Go to the *Graphs* ▷ *CLR biplot* menu.
  - Select the variables `Hous`, `Food`, `Serv`, and `Other` (in this order) from the window *Available data* and move them into the window *Selected data*.
  - Activate the *Add coordinates* option.
  - Click the *Accept* button.

CoDaPack creates three new columns —`UD1`, `UD2`, and `UD3`— with the PC-scores of the compositions of the *centred* CoDa set. These PC-scores are coordinates with respect to an orthonormal log-ratio basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ associated to the three principal components PC1, PC2, and PC3, respectively. In addition, CoDaPack writes in the *output* window the *clr*-coordinates of the elements of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ (Table 1.2).

Table 1.2. *clr*-coordinates of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ associated to PCs of `householdbudget`.

PCs	clr.Hous	clr.Food	clr.Serv	clr.Others
$\mathrm{clr}\,\mathbf{e}_1$	0.0311	0.7440	-0.1182	-0.6569
$\mathrm{clr}\,\mathbf{e}_2$	-0.8592	0.3506	0.1857	0.3230
$\mathrm{clr}\,\mathbf{e}_3$	0.1036	0.2712	-0.8376	0.4627

Vectors in Table 1.2 are in the 3-dimensional subspace $U = \{\mathbf{z} \in \mathbb{R}^4 : z_1 + \ldots + z_4 = 0\}$. For example, the *clr*-scores of $\mathbf{e}_1$ are $(0.0311, 0.7440, -0.1182, -0.6569)$, where 0.0311+0.7440-0.1182-0.6569= 0.

clr-biplot consists of a classical biplot applied to the *centred clr*-scores. We explore now the relationship between the *clr*-scores of a sample and the *olr*-coordinates in columns `UD1`, `UD2`, and `UD3`. First of all, to obtain the *clr*-scores of the CoDa set:

- Go to the menu *Data* ▷ *Transformations* ▷ *CLR*.
- Select the variables `Hous, Food, Serv, Others` from the window *Available data* and move them into the window *Selected data*.
- By default the option $Raw - CLR$ in the window *Options* is already activated.
- Click the button *Accept*.
- Go to menu *File* ▷ *Configuration* to define four decimals in the table: *Table format 0.0000*.

CoDaPack creates four columns —`clr.Hous`, `clr.Food`, `clr.Serv` and `clr.Others`— with the *clr*-scores of the compositions. For example, in the first row we have $\mathrm{clr}\,\mathbf{x}_1 = (0.3850, 0.5583, -0.1502, -0.7931)$.

We aim to explore the relationship between these values in $\mathrm{clr}\,\mathbf{x}_1$ and the values in the first row of columns `UD1`, `UD2`, and `UD3`. Because the *clr*-biplot is created using the *centred clr*-scores, we have to calculate the center of the *clr*-scores set. Using the menu *Statistics* ▷ *Classical statistics summary* we calculate the center of the *clr*-transformed CoDa set: $\mathbf{c} = (0.7294, -0.3546, -0.2612, -0.1136)$. This center is the arithmetic mean of the *clr*-variables. We introduce the concepts associated to the center in the next chapter, devoted to exploratory analysis. Let $\mathrm{clr}_c \mathbf{x}_1$ the centred vector of $\mathrm{clr}\, \mathbf{x}_1$, that is,

$$\begin{aligned} \mathrm{clr}_c \mathbf{x}_1 = \mathrm{clr}\, \mathbf{x}_1 - \mathbf{c} = (0.3850, 0.5583, -0.1502, -0.7931)- \\ -(0.7294, -0.3546, -0.2612, -0.1136) = (-0.3444, 0.9129, 0.1110, -0.6795) \end{aligned}$$

In the first row of columns `UD1`, `UD2`, and `UD3` we have the coordinates of the centred vector $\mathrm{clr}_c \mathbf{x}_1$ in the new basis: $(1.1018, 0.4170, -0.1955)$. That is, it holds

$$\begin{aligned} \mathrm{clr}_c \mathbf{x}_1 &= 1.1018 \cdot \mathrm{clr}\, \mathbf{e}_1 + 0.4170 \cdot \mathrm{clr}\, \mathbf{e}_2 - 0.1955 \cdot \mathrm{clr}\, \mathbf{e}_3 \\ \mathrm{clr}\, \mathbf{x}_1 - \mathbf{c} &= 1.1018 \cdot \mathrm{clr}\, \mathbf{e}_1 + 0.4170 \cdot \mathrm{clr}\, \mathbf{e}_2 - 0.1955 \cdot \mathrm{clr}\, \mathbf{e}_3 \\ \mathrm{clr}\, \mathbf{x}_1 &= \mathbf{c} + 1.1018 \cdot \mathrm{clr}\, \mathbf{e}_1 + 0.4170 \cdot \mathrm{clr}\, \mathbf{e}_2 - 0.1955 \cdot \mathrm{clr}\, \mathbf{e}_3 \ . \end{aligned}$$

To calculate the expression in percentages of $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ we must apply the inverse of the *clr*-transformation. To do this with CoDaPack we have to create a new table where the *clr*-scores of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are introduced:

- Go to the *Data* ▷ *Create new table*.
  - ○ To add three *clr*-scores vectors of 4-part composition: write 4 in the *Number of columns* window and 4 in the *Number of rows* window. We need four rows because the first one has to contain the names of each *clr*-score.
  - ○ Click the *Accept* button.
  - ○ CoDaPack creates an empty table. In the first row we write the names $clr.Hous, clr.Food, clr.Serv, clr.Others$. In the rows second, third and fourth we introduce the numerical values of Table 1.2.
  - ○ Click the button *Accept*.
  - ○ Finally we have to enter the name for the new table, write *clrPCs*.
  - ○ Click the button *Ok*.
- Observe that CoDaPack creates a table named *clrPCs*. We can save it using the standard menus *File* ▷ *Save Workspace...* or *File* ▷ *Save as...*.

To apply the *clr*-inverse transformation:

- Go to the menu *Data* ▷ *Transformations* ▷ *CLR*.
- Select the variables $clr.Hous, clr.Food, clr.Serv, clr.Others$ from the window *Available data* and move them into the window *Selected data*.
- Activate the option $CLR - Raw$ in the window *Options*.
- Click the button *Accept*.

CoDaPack creates four columns —`inv.clr.Hous`, `inv.clr.Food`, `inv.clr.Serv`, and `inv.clr.Others`— with the values (in proportions) of the elements $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. Table 1.3 shows that, in percentages, $\mathbf{e}_1 = (22.71, 46.32, 19.56, 11.41)$, where the *relevant* coefficients are respectively "46.32%" and "11.41%". Note that these coefficients are the values furthest away from $\mathbf{u} = (25\%, 25\%, 25\%, 25\%)$, the neutral composition. This fact can be illustrated with some samples in the CoDa set.

For example, the sample in the row number 13 has a *clr*-score $\mathbf{e}_{13,1} = 2.5687$, the largest positive value among the values in column `UD1`. The sample, in original units, is $\mathbf{x}_{13} = (32.6446, 54.4507, 6.1934, 1.9355)$ where the relative large and small values correspond to `Food` and `Others`, respectively. In contrast, the sample $\mathbf{x}_{32} = (100.7724, 6.0644, 58.3216, 245.9312)$, where the relative small and large values correspond to `Food` and `Others`, has a *clr*-score $\mathbf{e}_{32,1} = -2.4767$, the largest value among the negatives ones.

Table 1.3. Values (in %) of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ associated to PCs of `householdbudget`.

composition	Hous	Food	Serv	Others
$\mathbf{e}_1$	22.71	46.32	19.56	11.41
$\mathbf{e}_2$	9.56	32.06	27.19	31.19
$\mathbf{e}_3$	24.97	29.53	9.74	35.76

1.7. Scale invariant logratios or logcontrasts

Activities for Section 1.7

In this section we transform a linear combination of logratios to a logcontrast (log-linear combination of original parts, where the coefficients sum zero).

A *clr-biplot* is a reduced-dimensionality representation of a CoDa set. That is, given a CoDa set where the samples have more than three parts ($D > 3$), the *clr*-scores are calculated and we create an *approximate* representation of the *clr*-scores set in a scatterplot. The idea consists of creating a new set of $D-1$ *clr*-variables (*principal components*: PC) creating logcontrasts of the original variables, where few of the PC account for a large proportion of the variance of the CoDa set. Next Chapter, which is devoted to exploratory techniques, includes a detailed introduction of the *clr*-biplot representation.

We focus on the first PC $\mathbf{e}_1$ (see Table 1.2) of the *clr*-biplot of the `householdbudget` CoDa set (see Appendix). Table 1.2 shows that *clr*-scores of $\mathbf{e}_1$ are $\operatorname{clr}\mathbf{e}_1 = (0.0311, 0.7440, -0.1182, -0.6569)$, where 0.0311+0.7440-0.1182-0.6569= 0. Therefore, for $\operatorname{clr}\mathbf{e}_1$, the linear combination is

$$\begin{aligned}\operatorname{clr}\mathbf{e}_1 &= 0.0311 \cdot \operatorname{clr}(Hous) + 0.7440 \cdot \operatorname{clr}(Food) - 0.1182 \cdot \operatorname{clr}(Serv) \\ &\quad - 0.6569 \cdot \operatorname{clr}(Others) \ .\end{aligned}$$

Expressing each *clr*-score as a logratio it holds

$$\operatorname{clr}\mathbf{e}_1 = 0.0311 \cdot \ln\frac{Hous}{g} + 0.7440 \cdot \ln\frac{Food}{g} - 0.1182 \cdot \ln\frac{Serv}{g} - 0.6569 \cdot \ln\frac{Others}{g} \ ,$$

where $g = (Hous \cdot Food \cdot Serv \cdot Others \cdot)^{1/4}$. For the property of the function logarithm $\ln \frac{a}{b} = \ln a - \ln b$, we can write

$$\begin{aligned}\mathrm{clr}\,\mathbf{e}_1 &= 0.0311 \cdot \ln(Hous) + 0.7440 \cdot \ln(Food) - 0.1182 \cdot \ln(Serv) - 0.6569 \cdot \ln(Others) \\ &\quad - (0.0311 + 0.7440 - 0.1182 - 0.6569) \cdot \ln g .\end{aligned}$$

Because 0.0311+0.7440-0.1182-0.6569$=0$, we obtain the logcontrast for the composition $\mathbf{e}_1$:

$$\mathrm{clr}\,\mathbf{e}_1 = 0.0311 \cdot \ln(Hous) + 0.7440 \cdot \ln(Food) - 0.1182 \cdot \ln(Serv) - 0.6569 \cdot \ln(Others),$$

where the vector of coefficients is $\mathbf{a} = (0.0311, 0.7440, -0.1182, -0.6569)$. In this vector the more *relevant* coefficients are "+0.7440" and "−0.6569", for the positive and negative case, respectively. In percentages, the first PC is $\mathbf{e}_1 = (22.71, 46.32, 19.56, 11.41)$ (see Table 1.3), where the *relevant* coefficients are respectively "46.32%" and "11.41%". Note that these coefficients are the values furthest away from the neutral composition $\mathbf{u} = (25\%, 25\%, 25\%, 25\%)$.

1.8. Representation of compositions by orthonormal coordinates

Activities for Section 1.8

In this section, we focus on the proof that the *principal components* of the *clr*-biplot of the `householdbudget` CoDa set (see Appendix) form an *olr*-basis. Because orthogonal compositions are linear independent, a set of $D-1$ orthogonal compositions constitutes a basis of $\mathcal{S}^D$. To prove that $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ (see Table 1.3) is an *olr*-basis we have to check if $||\mathbf{e}_k||_a = 1$, for $k = 1,2,3$ and $< \mathbf{e}_k, \mathbf{e}_j >_a = 0$, for $k \neq j = 1,2,3$. This calculation is equivalent to $||\mathrm{clr}\,\mathbf{e}_k|| = 1$, for $k = 1,2,3$ and $< \mathrm{clr}\,\mathbf{e}_k, \mathrm{clr}\,\mathbf{e}_j >= 0$, for $k \neq j = 1,2,3$ for the Euclidean norm and inner product.

Vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are unitary because

$$\begin{aligned}||\mathrm{clr}\,\mathbf{e}_1||^2 &= (0.0311)^2 + (0.7440)^2 + (-0.1182)^2 + (-0.6569)^2 &= 1 \text{ ,} \\ ||\mathrm{clr}\,\mathbf{e}_2||^2 &= (-0.8592)^2 + (0.3506)^2 + (0.1857)^2 + (0.3230)^2 &= 1 \text{ , and} \\ ||\mathrm{clr}\,\mathbf{e}_3||^2 &= (0.1036)^2 + (0.2712)^2 + (-0.8376)^2 + (0.4627)^2 &= 1 \text{ .}\end{aligned}$$

To do these calculations with CoDaPack:

- Use the menu *Data* ▷ *Create new Table* to create a table with the following data set:

`clre1`	`clre2`	`clre3`	as variable names, and
0.0311	−0.8592	0.1036	
0.7440	0.3506	0.2712	as data.
−0.1182	0.1857	−0.8376	
−0.6569	0.3230	0.4627	

- Go to the *Data* ▷ *Manipulate* ▷ *Calculate new Variable* menu.
  - ○ Move the variable `clre1` to the *Selected data* window.

  - Click the *Accept* button.
  - Write $x1 * x1$ in the *Enter expression* window (Note: internally CoDaPack associates $x1$ with the variable introduced in the *Selected data* window).
  - Write `clre1.2` into the *Enter new variable name* window.
  - Click the *Ok* button.

- Go to the *Statistics* ▷ *Classical statistics summary* menu.
  - Move the variable `clre1.2` to the *Selected data* window.
  - Select the option *Mean*.
  - Click the *Accept* button.

Observe that the arithmetic mean of vector $((\mathrm{clr}\,\mathbf{e}_{11})^2, (\mathrm{clr}\,\mathbf{e}_{12})^2, (\mathrm{clr}\,\mathbf{e}_{13})^2, (\mathrm{clr}\,\mathbf{e}_{14})^2)$ (*Output* window) takes the value 0.25. This means that $\sum_{k=1}^{4}(\mathrm{clr}\,\mathbf{e}_{1k})^2 = 1$. Therefore, the vector $\mathbf{e}_1$ is unitary ($||\mathbf{e}_1||_\mathrm{a} = ||\mathrm{clr}\,\mathbf{e}_1|| = 1$). We can analogously repeat the calculations above to prove that vectors $\mathbf{e}_2$ and $\mathbf{e}_3$ are unitary as well.

In addition, vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are orthogonal because

$$\begin{aligned}
< \mathrm{clr}\,\mathbf{e}_1, \mathrm{clr}\,\mathbf{e}_2 > &= (0.0311)\cdot(-0.8592) + (0.7440)\cdot(0.3506) + (-0.1182)\cdot(0.1857) + \\
&+ (-0.6569)\cdot(0.3230) = 0\ , \\
< \mathrm{clr}\,\mathbf{e}_1, \mathrm{clr}\,\mathbf{e}_3 > &= (0.0311)\cdot(0.1036) + (0.7440)\cdot(0.2712) + (-0.1182)\cdot(-0.8376) + \\
&+ (-0.6569)\cdot(0.4627) = 0\ , \text{and} \\
< \mathrm{clr}\,\mathbf{e}_2, \mathrm{clr}\,\mathbf{e}_3 > &= (-0.8592)\cdot(0.1036) + (0.3506)\cdot(0.2712) + (0.1857)\cdot(-0.8376) + \\
&+ (0.3230)\cdot(0.4627) = 0\ .
\end{aligned}$$

For checking these calculations with CoDaPack we can use the same data we introduced above:

- Go to the *Data* ▷ *Manipulate* ▷ *Calculate new Variable* menu.
  - Move the variables `clre1` and `clre2` to the *Selected data* window.
  - Click the *Accept* button.
  - Write $x1 * x2$ in the *Enter expression* window (Note: internally CoDaPack associates $x1$ and $x2$ with the variables `clre1` and `clre2` introduced in the *Selected data* window according to the order of entry).
  - Write `clre1.clre2` into the *Enter new variable name* window.
  - Click the *Ok* button.
- Go to the *Statistics* ▷ *Classical statistics summary* menu.
  - Move the variable `clre1.clre2` to the *Selected data* window.
  - Select the option *Mean*.
  - Click the *Accept* button.

In the *Output* window we observe that the arithmetic mean of vector $(\mathrm{clr}\,\mathbf{e}_{11}\cdot\mathrm{clr}\,\mathbf{e}_{21}, \mathrm{clr}\,\mathbf{e}_{12}\cdot\mathrm{clr}\,\mathbf{e}_{22}, \mathrm{clr}\,\mathbf{e}_{13}\cdot\mathrm{clr}\,\mathbf{e}_{23}, \mathrm{clr}\,\mathbf{e}_{14}\cdot\mathrm{clr}\,\mathbf{e}_{24})$ is 0. This means that $\sum_{k=1}^{4}\mathrm{clr}\,\mathbf{e}_{1k}\cdot\mathrm{clr}\,\mathbf{e}_{2k} = 0$. Therefore, vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are orthogonal ($< \mathbf{e}_1, \mathbf{e}_2 >_\mathrm{a} = < \mathrm{clr}\,\mathbf{e}_1, \mathrm{clr}\,\mathbf{e}_2 > = 0$). We can analogously repeat the calculations above to prove that vectors $\mathbf{e}_1$ and $\mathbf{e}_3$ are orthogonal, as well as vectors $\mathbf{e}_2$ and $\mathbf{e}_3$.

An analogous calculation can be done using a matrix expression. Let $\mathbf{\Phi}$ be the $(D-1)\times D$-matrix whose i-th row is the vector $\mathrm{clr}\,\mathbf{e}_i$, for $i = 1, \ldots, D-1$, where

$D = 4$:

$$\mathbf{\Phi} = \begin{pmatrix} 0.0311 & 0.7440 & -0.1182 & -0.6569 \\ -0.8592 & 0.3506 & 0.1857 & 0.3230 \\ 0.1036 & 0.2712 & -0.8376 & 0.4627 \end{pmatrix}$$

To prove that $\{\mathbf{e}_1, \mathbf{e}_2,\ \mathbf{e}_3\}$ is an *olr*-basis is equivalent to check if the matrix $\mathbf{\Phi}$ satisfies $\mathbf{\Phi} \cdot \mathbf{\Phi}^t = \mathbf{I}_{D-1}$, where $\mathbf{I}_{D-1}$ is the identity matrix of dimension $D - 1$:

$$\mathbf{\Phi}\cdot\mathbf{\Phi}^t = \begin{pmatrix} 0.0311 & 0.7440 & -0.1182 & -0.6569 \\ -0.8592 & 0.3506 & 0.1857 & 0.3230 \\ 0.1036 & 0.2712 & -0.8376 & 0.4627 \end{pmatrix} \begin{pmatrix} 0.0311 & -0.8592 & 0.1036 \\ 0.7440 & 0.3506 & 0.2712 \\ -0.1182 & 0.1857 & -0.8376 \\ -0.6569 & 0.3230 & 0.4627 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The basis $\{\mathbf{e}_1, \mathbf{e}_2,\ \mathbf{e}_3\}$ is an *olr*-basis because the three elements are unitary and orthogonal by pairs. Using a *clr*-biplot we obtain an *olr*-basis $\{\mathbf{e}_1, \mathbf{e}_2,\ \mathbf{e}_3\}$ which *clr*-scores and also the coordinates of the samples in this basis (columns `UD1`, `UD2`, and `UD3`) are *data-driven* because both depend on the particular CoDa set. Consequently, when removing just one sample of the CoDa set, one obtains a different *olr*-basis. On the other hand, an *olr*-basis associated to an SBP (see Section 1.9) is *expert-driven*, the *clr*-scores of the basis are the same regardless the CoDa set analyzed.

1.9. olr -basis associated to a sequential binary partition

Activities for Section 1.9

A. We continue using the CoDaPack package to analyse the `bloodMN` CoDa set (see Appendix) that we began to analyse in previous Activities sections.

- If necessary, retrieve the `bloodMN02.cdp` file from the CoDaPack package.

Now, we will apply an isometric logratio transformation (*ilr*) to our data.

Recall that to define an *ilr*-transformation, it is necessary to define previously the *olr*-basis of $\mathcal{S}^D$ on which it is defined. Therefore we define a hierarchical partition of the three parts MN, MM and NN.

- Go to the *Data* ▷ *Transformations* ▷ *Raw-ILR* menu.
  - ○ Select the variables `MN`, `MM` and `NN` (in this order) from the *Available data* window and move them into the *Selected data* window.

    CoDaPack asks us to choose the vectors of the *olr*-basis using a sequential binary partition (SBP) of parts.

The program offers a default hierarchical partition. Click on the *Default partition* button. The sign matrix **S** linked to this partition is displayed in the *Defined partition* window:

i	MN	MM	NN
1	$+1$	$+1$	-1
2	$+1$	-1	0

We prefer defining the hierarchical partition to our convenience linked to the following sign matrix:

i	MN	MM	NN
1	-1	$+1$	$+1$
2	0	$+1$	-1

According to the first row of the sign matrix, the first *olr*-coordinate is a *normalised* logratio interpreted as a *balance* between the geometric mean of the frequencies of genotypes MM and NN (with $+1$ in the sign matrix), and the frequency of genotype MN (with -1 in the sign matrix). The second *olr*-coordinate (corresponding to the second row of the sign matrix) is a *normalised* logratio of the frequencies of genotypes MM and NN.

- ∘ Click on the *Define Manually* button.
  The *Binary Partition Menu* appears to introduce our sign matrix:
  - ⋆ Click on the (MM,A)-cell to change the default *minus* sign for the *plus* sign.
  - ⋆ Click on the (NN,A)-cell to change the default *minus* sign for the *plus* sign.
  - ⋆ Click on the *Next* button.
  - ⋆ Click on the (MM,B)-cell to change the default *minus* sign for the *plus* sign.
  - ⋆ Click on the *Next* button.
  - ⋆ Click on the *Accept* button.

  The new sign matrix is displayed in the *Defined partition* window.
- ∘ Click on the *Accept* button.

Our sign matrix is displayed in the *Results* window of CoDaPack. Moreover two new variables `ilr.1` and `ilr.2` are added to the *Variables* window.

According to equations

$$
\begin{aligned}
\phi_{ij} &= 0, && \text{if } s_{ij} = 0;\\
\phi_{ij} &= +\tfrac{1}{n_i}\sqrt{\tfrac{n_i \cdot d_i}{n_i + d_i}}, && \text{if } s_{ij} > 0;\\
\phi_{ij} &= -\tfrac{1}{d_i}\sqrt{\tfrac{n_i \cdot d_i}{n_i + d_i}}, && \text{if } s_{ij} < 0,
\end{aligned}
\tag{1.1}
$$

these variables provide the *olr*-coordinates of compositions with respect to the *olr*-basis of $\mathcal{S}^D$

$$\begin{aligned} \mathbf{e}_1 = & \ \mathrm{clr}^{-1}\left[-\sqrt{\tfrac{2}{3}}, \tfrac{1}{2}\sqrt{\tfrac{2}{3}}, \tfrac{1}{2}\sqrt{\tfrac{2}{3}}\right] \ , \\ \mathbf{e}_2 = & \ \mathrm{clr}^{-1}\left[0, \tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{2}}\right] \ . \end{aligned}$$

The *olr*-coordinates x_1^* and x_2^* of a composition [MN,MM,NN] are equal to

$$\begin{aligned} x_1^* = & \ -\sqrt{\tfrac{2}{3}} \ln \mathtt{MN} + \tfrac{1}{2}\sqrt{\tfrac{2}{3}} \ln \mathtt{MM} + \tfrac{1}{2}\sqrt{\tfrac{2}{3}} \ln \mathtt{NN} = \sqrt{\tfrac{2}{3}} \ln \tfrac{(\mathtt{MM}\cdot\mathtt{NN})^{1/2}}{\mathtt{MN}} \ , \\ x_2^* = & \ \tfrac{1}{\sqrt{2}} \ln \mathtt{MM} - \tfrac{1}{\sqrt{2}} \ln \mathtt{NN} = \tfrac{1}{\sqrt{2}} \ln \tfrac{\mathtt{MM}}{\mathtt{NN}} \ . \end{aligned}$$

The first *olr*-coordinate x_1^* provides the normalised balance of the geometric mean of MM and NN with respect to MN. The second *olr*-coordinate x_2^* coincides with the logratio $\ln(\mathtt{MM}/\mathtt{NN})$ affected by a normalisation factor.

CoDaPack allows us to draw the Cartesian representation of our data from their *olr*-coordinates provided that compositions have at most four parts.

- Go to the *Graphs* ▷ *ILR/CLR plot* menu.
  - ○ Select the variables `MN`, `MM` and `NN` (in this order) from the *Available data* window and move them into the *Selected data* window.

    As in the previous menu, CoDaPack asks us to choose the vectors of the *olr*-basis using a sequential binary partition (SBP) of parts. We use the same SBP as before. To do this we use a *copy&paste*. That is, we select the sign matrix displayed in the *Results* window of CoDaPack to *copy* and afterwards *paste* it into the *Defined base:* window. Otherwise, we have to reintroduce the sign matrix as before.
  - ○ Click the *Accept* button.

Observe that all compositions have the first *olr*-coordinate negative. It means that MN is always greater than the geometric mean of MM and NN in all samples. The graph also highlights the low variability of the first *olr*-coordinate compared to the high variability of the second *olr*-coordinate. This fact emphasises again the *linear* pattern of the data set. Since this pattern in the plot is parallel to the vertical axis corresponding to the second *olr*-coordinate, it would not be unreasonable to summarise the pattern of variation in our data by saying that *the first olr-coordinate remains approximately constant*, which is the same as saying that

$$\sqrt{\frac{2}{3}} \ln \frac{(\mathtt{MM}\cdot\mathtt{NN})^{1/2}}{\mathtt{MN}} \simeq k_1 \ \rightarrow \ \frac{\mathtt{MM}\cdot\mathtt{NN}}{\mathtt{MN}^2} \simeq k_2 .$$

If we take the value of the constant k_1 as the average of the first *olr*-coordinates of the samples ($k_1 = -0.54$), the second member of the above equation is rewritten as

$$\frac{\mathtt{MM}\cdot\mathtt{NN}}{\mathtt{MN}^2} \simeq 0.27 \ .$$

This equality is simply a numerical confirmation of the Hardy-Weinberg principle of genetic equilibrium (see *BloodMN* description in the Appendix). Remember that this principle states that

$$\frac{\texttt{MM}\cdot\texttt{NN}}{\texttt{MN}^2}=0.25\ .$$

This *law* can be rewritten in logcontrast format as

$$\ln\texttt{MM}+\ln\texttt{NN}-2\ln\texttt{MN}=-\ln 4\ .$$

In any case, our data confirm the Hardy-Weinberg principle.

To save the data set together with the new variables, go to the *File* ▷ *Save Workspace...* menu and save the workspace with the name `bloodMN03.cdp`.

B. The representation of compositions by their *olr*-coordinates facilitates the analysis of CoDa if we choose a suitable *olr*-basis. Now, we continue the analysis of the `statisticiantimebudget` CoDa set described in the Appendix.
First, we analyse the daily time devoted by the statistician to non-working activities (other wakeful activities (`O`) and sleep (`S`)) compared to time devoted to working activities (teaching (`T`), consultation (`C`), administration (`A`) and research (`R`)). With this aim, the first step of the SBP consists of dividing the composition into two groups of parts: `O` and `S` (with $+1$) in one group, and `T`, `C`, `A` and `R` (with -1) in the other group. In the second step of the SBP we compare the daily time devoted to other wakeful activities (`O`, with $+1$) with the sleeping time (`S`, with -1). In the third step of the SBP we compare the teaching and research activities (`T` and `R`, with $+1$) with the other working activities (`C` and `A`, with -1). In the fourth step of the SBP we compare `T` (with $+1$) with `R` (with -1) Finally, in the last step we compare `C` (with $+1$) with `A` (with -1). Therefore, the sign matrix **S** associated to this SBP is equal to

i	T	C	A	R	O	S
1	-1	-1	-1	-1	$+1$	$+1$
2	0	0	0	0	$+1$	-1
3	$+1$	-1	-1	$+1$	0	0
4	$+1$	0	0	-1	0	0
5	0	$+1$	-1	0	0	0

- Go to the *Data* ▷ *Transformations* ▷ *Raw-ILR* menu.
  - Move the variables `T`, `C`, `A`, `R`, `O` and `S` (in this order) to the *Selected data* window.
  - Click the the *Define Manually* button to define the SBP.
  - Using the *Binary Partition Menu* introduce the signs of the SBP.
  - Click on the *Accept* button.

The new variable `ilr.1`, that is, the first *olr*-coordinate, gives the balance between the geometric mean of subcompositions sub($\mathbf{x}$; `O`,`S`) and sub($\mathbf{x}$; `T`,`C`,`A`,`R`) of $\mathbf{x}$. Note that this first *olr*-coordinate is positive in all 20 samples (days). Therefore the geometric mean of the daily time devoted by the statistician to `O` and S is systematically greater than the geometric mean of the daily time spent in working activities.

We can use CoDaPack to perform a brief statistical analysis of all *olr*-coordinates.

- Go to the *Statistics* ▷ *Classical statistics summary* menu.
  - Move the variables `ilr.1`, `ilr.2`, `ilr.3`, `ilr.4` and `ilr.5` to the *Selected data* window.
  - Select *Mean*, *Percentile 0 100* and *Standard Deviation* in the *Options* window.
  - Click on the *Accept* button.

We confirm from the *Statistics* table in the *Results* window of CoDaPack that the first *olr*-coordinate varies from 0.6043 to 1.1192, with a mean equal to 0.8531 and a moderate standard deviation equal to 0.1528. The mean of the second *olr*-coordinate —the balance between O (with $+1$) and sleep S (with -1)— is negative (-0.1313), and its standard deviation (0.2199) is very high in relation to the mean. Therefore, although the statistician generally devoted more time to sleep than to other wakeful activities, he does not have a clear daily pattern. With regard to working activities, he usually devotes more time to T and R than to C and A (mean(`ilr.3`)$= 0.1224 > 0$), although the daily variability (std. dev.(`ilr.3`)$= 0.1933$) is considerable. The statistician regularly devotes more time to teaching than to research (mean(`ilr.4`)$= 0.1891 > 0$), and more time to administration than to consulting (mean(`ilr.5`)$= -0.1203 < 0$).

CoDaPack provides a graphical facility to represent a system of coordinates based on a SBP: a *balance-dendrogram*

- Go to the *Graphs* ▷ *Balance dendrogram* menu.
  - Move the variables T, C, A, R, O and S (in this order) to the *Selected data* window.
  - Select and *copy* the sign matrix displayed into the *Results* window and *paste* it in the *Defined partition* window. Otherwise, we have to reintroduce the sign matrix as before.
  - Deactivate the *Add statistics* option.
  - Click the *Accept* button.

In this plot one can see the hierarchical partition of the parts created by an SBP. The structure of the SBP is different when one uses pivot coordinates (Table 1.4):

Table 1.4. Pivot coordinates for the `statisticiantimebudget` data set

i	T	C	A	R	O	S
1	$+1$	-1	-1	-1	-1	-1
2	0	$+1$	-1	-1	-1	-1
3	0	0	$+1$	-1	-1	-1
4	0	0	0	$+1$	-1	-1
5	0	0	0	0	$+1$	-1

To create the balance-dendrogram for the pivot coordinates (Fig. 1.4)

- Go to the *Graphs* ▷ *Balance dendrogram* menu.

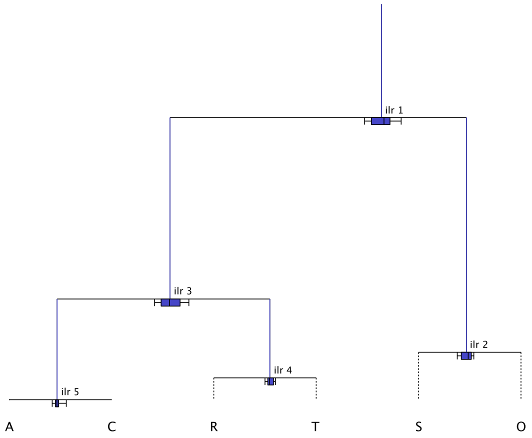


Figure 1.3. A balance dendrogram for the `statisticiantimebudget` data set)

- ○ Move the variables `T`, `C`, `A`, `R`, `O` and `S` (in this order) to the *Selected data* window.
- ○ Click the the *Define Manually* button to define the SBP.
- ○ Using the *Binary Partition Menu* introduce the signs of the SBP according Table 1.4.
- ○ Deactivate the *Add statistics* option.
- ○ Click the *Accept* button.

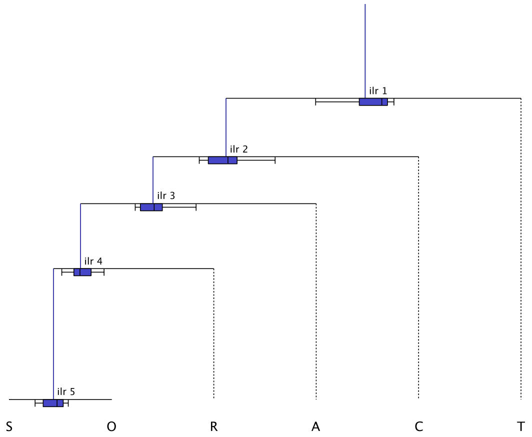


Figure 1.4. The balance-dendrogram of the pivot coordinates for the `statisticiantimebudget` data set

A balance-dendrogram plot is useful for the exploratory analysis of CoDa, a working topic for the next Chapter.

The chapter's key concepts

✓ CoDa provide only relative information between their components.

✓ Any meaningful function of a composition can be expressed in terms of ratios of their components, that is the function must be scale-invariant.

✓ Any analysis involving CoDa must be subcompositionally coherent.

✓ The simplex is the sample space of CoDa.

✓ Perturbation and powering are the basic operations in the simplex.

✓ The analysis of CoDa is based on the logratios of their components.

✓ The simplex $\mathcal{S}^D$ can be viewed as a real vector space of dimension $D-1$, perturbation being the internal operation and powering the external one.

✓ A logcontrast is any real scale invariant log-ratio function of the components of a composition.

✓ The *alr* and the *clr* transformations are linear maps from the vector space $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ and $\mathbb{R}^D$, respectively. These transformations and their inverses let us operate easily with compositions.

✓ The simplex $\mathcal{S}^D$ is also a Euclidean space because the *clr* transformation allows the Euclidian metric of the real space to be exported to $\mathcal{S}^D$.

✓ Given a composition $\mathbf{x} \in \mathcal{S}^D$, the components in $\mathbb{R}^{D-1}$ of the real vector $\mathrm{alr}\,\mathbf{x}$ can be interpreted as the coordinates of $\mathbf{x}$ in a (not orthonormal) basis of $\mathcal{S}^D$.

✓ Given an *olr*-basis $\mathcal{B}$ of $\mathcal{S}^D$, the isometric logratio transformation *ilr* (associated to $\mathcal{B}$) send each composition $\mathbf{x}$ to vector $\mathbf{x}^*$ in $\mathbb{R}^{D-1}$ whose components are the coordinates of $\mathbf{x}$ in relation to the basis $\mathcal{B}$.

✓ The SBP is a procedure to easily obtain the *olr*-basis of $\mathcal{S}^D$. In this case, the coordinates of a composition (relative to one of these bases) can be easily interpreted as *balances* between the geometric means of two subsets of parts.

Chapter 2

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.

2.1. Centre of a compositional data set

Activities for Section 2.1

We want to explore the `bloodMN` CoDa set (see Appendix). This set records the information on the absolute frequencies of M, N and MN blood types observed in samples of people coming from different ethnic groups around the world.

- Load the `bloodMN02.cdp` file in the CoDaPack package.

To calculate the center and percentiles of the set `bloodMN`:

- Go to the menu *Statistics* ▷ *Compositional statistics summary*.
  - Move the variables `MN`, `MM` and `NN` to the *Selected data* window.
  - Click *Percentile* in the *Options* window.
  - Uncheck the options *Variation Array* and *Total Variance*.
  - Click the *Accept* button.

In the *Output* window CoDaPack provides information about the sample size ($n = 49$) and the centre $\mathbf{g} = [0.4776, 0.3487, 0.1737]$ of the CoDa set. It also provides the minimum (column with label 0), the maximum (column with label 100) and the three quartiles Q_1, Q_2 and Q_3 (columns with labels 25, 50 and 75, respectively) for each part of the closed 3-part compositions ($\kappa = 1$).

We will perform the centring of the CoDa set.

- Go to the *Data* ▷ *Operations* ▷ *Centering* menu.
  - Move the variables `MN`, `MM` and `NN` to the *Selected data* window.
  - Click the *Accept* button.

CoDaPack creates the new columns `c_MN`, `c_MM` and `c_NN` which are the parts of the centred set. These parts are closed to $\kappa = 1$. The centre of the centred set is the barycentre $\mathbf{u} = (1/3, 1/3, 1/3)$ of $\mathcal{S}^3$. We can easily check it with CoDaPack:

- Go to the menu *Statistics* ▷ *Compositional statistics summary*.
  - Move the variables `c_MN`, `c_MM` and `c_NN` to the *Selected data* window.
  - Uncheck the options *Variation Array* and *Total Variance*.
  - Click the *Accept* button.
- Go to the *Graphs* ▷ *Ternary/Quaternary plot* menu.
  - Move the variables `MN`, `MM` and `NN` to the *Selected data* window.
  - Click the *Accept* button.
  - Activate the *Grid* option at the border of the graphic window.
  - Activate the *Show Center* option.
  - Activate the *Centered* option. For illustration purposes, CoDaPack moves the data set but the original center remains at the same place.
- Similarly, plot the ternary diagram of `c_MN`, `c_MM` and `c_NN` variables.

The center of the centred set is located on the barycenter $(1/3, 1/3, 1/3)$ of the triangle. The comparison of the ternary plots highlights the equality of the patterns of variation of the original and the centred sets. Keep in mind that the centred set is obtained by applying the perturbation $\ominus \mathbf{g}$ to the original data set, $\mathbf{g}$ being its (compositional) centre. We use the facilities of the CoDaPack to visualise how the centring operation in the simplex corresponds to a movement to the origin in the log-ratio space.

- Go to the *Graphs* ▷ *ILR/CLR plot* menu.
  - ○ Move the variables `MM, MN, NN` to the *Selected data* window.
  - ○ Click the *Default Partition* button.
  - ○ Click the *Accept* button.
  - ○ Repeat the graph using the variables `c_MM, c_MN, c_NN` as the *Selected data*.

Observe that the *olr*-scatter plot (`ilr.1 vs ilr.2`) for the centred set can be obtained by translation to the origin of the original point cloud. The same effect can be observed when one represents the corresponding *clr*-scores and the *alr*-coordinates using the *Graphs* ▷ *Scatterplot 2D/3D* menu.

We save the file for the next activities:

- Go to the *File* ▷ *Save Workspace...* menu.

2.2. Covariance structure of a compositional data set

Activities for Section 2.2

- Load the `bloodMN02.cdp` file saved in Section 2.1.

We calculate the most common compositional statistics of the CoDa set to explore the covariance structure: variation array and total variance.

- Go to the *Statistics* ▷ *Compositional statistics summary* menu.
  - ○ Move the variables `MN`, `MM` and `NN` to the *Selected data* window.
  - ○ Select the options *Variation Array* and *Total Variance*.
  - ○ Click the *Accept* button.

In the *Output* window CoDaPack provides information about the *variation array*. Below the diagonal of this array, we find the arithmetic mean of the pairwise logratios:

$$\begin{aligned} \text{mean}(\ln(MM/MN)) &= -0.3145 \\ \text{mean}(\ln(NN/MN)) &= -1.0115 \\ \text{mean}(\ln(NN/MM)) &= -0.6970 \ . \end{aligned}$$

Above the diagonal, we find the variance of the pairwise logratios:

$$\begin{aligned} \text{var}(\ln(MN/MM)) &= 0.4476 \\ \text{var}(\ln(MN/NN)) &= 0.4195 \\ \text{var}(\ln(MM/NN)) &= 1.6522 \ . \end{aligned}$$

On the assumption that the logarithm of the variances of logratios follow a *t*-Student distribution, CoDaPack highlights in dark blue those variances below the 5^{th} percentile, in light blue those from the 5^{th} to 25^{th} percentile, in light red those from the 75^{th} to 95^{th} percentiles, and in dark red those above the 95^{th} percentile. In our case, the light red colour on the 1.6522 cell informs us that the variance of $\ln(MM/NN)$ is moderately high.

The added column to the right of the variation array contains the variances in the centred logratio covariance matrix $\mathbf{\Gamma}$, i.e. its diagonal. Thus, for example,

$$0.0091 = \mathrm{var}\left(\ln \frac{MN}{(MN \cdot MM \cdot NN)^{1/3}}\right) .$$

In the lower right-hand corner we find the total variance (= 0.8398) of the CoDa set, that is, the trace of covariance matrix $\mathbf{\Gamma}$.

The variances of all pairwise logratios of parts of the centred set remain unchanged, that is, the pattern of variation of the centred set is the same as that of the original data set.

- Go to the *Statistics* ▷ *Compositional statistics summary* menu.
  - ○ Move the variables `c_MN`, `c_MM` and `c_NN` to the *Selected data* window.
  - ○ Click the *Accept* button.

Notice how the variation array of the centred set is the same as that of the original data set as regards the pairwise log-ratio variances, whereas the estimation of the expected pairwise logratios are zero.

Now we will scale (to unity) the total variance of the CoDa set. It can be achieved (rounding error excepted) powering the centred set by $(\mathrm{totvar}(\mathbf{X}))^{-1/2}$, that is, powering by $0.8398^{-1/2} = 1.0912$.

- Go to the *Data* ▷ *Operations* ▷ *Power transformation* menu.
  - ○ Move the variables `c_MN`, `c_MM` and `c_NN` to the *Selected data* window.
  - ○ Write 1.0912 in the *Power* box in the *Options* window.
  - ○ Activate *Closure result* box in the *Options* window.
  - ○ Write 1 in the *Closure to* box in the *Options* window.
  - ○ Click the *Accept* button.

CoDaPack creates the new variables `c_MN.pt.1.0912`, `c_MM.pt.1.0912` and `c_NN.pt.1.0912` which are the new scaled parts.

Obviously, the total variance of the scaled CoDa set must be equal to 1. Graphically, the scaled set can be obtained by *shrinking* (if $\mathrm{totvar}(\mathbf{X}) > 1$) or *dilating* (if $\mathrm{totvar}(\mathbf{X}) < 1$) the centred set.

- Go to the *Statistics* ▷ *Compositional statistics summary* menu.
  - ○ Move the variables `c_MN.pt.1.0912`, `c_MM.pt.1.0912` and `c_NN.pt.1.0912` to the *Selected data* window.
  - ○ Click the *Accept* button.

Observe how the total variance of the scaled CoDa set is equal to 0.9999 (≈ 1). The original total variance 0.8398 changed to 1 (rounding errors excepted) by the factor 1.0912^2 ($0.8398 \cdot 1.0912^2 \approx 1$). The same factor explains the change between the pairwise logratios. For example, the logratio variance $\mathrm{var}\left(\ln MM/\ln MN\right)$ of the original data set (0.4476) changes to 0.5330 for the scaled set ($0.4476 \cdot 1.0912^2 \approx 0.5330$)

Because the original CoDa set has a total variance close to one (1), no large differences are expected between the original and scaled data. Indeed, for example, the first sample (first row in the data frame) of the centred set is $(0.3089, 0.1505, 0.5406)$ whereas the first sample of the scaled CoDa set is $(0.3032, 0.1384, 0.5584)$. The ternary diagrams of both data sets supports this idea:

- Go to the *Graphs* ▷ *Ternary/Quaternary plot* menu.
  - ○ Move the variables `c_MM.pt.1.0912`, `c_MN.pt.1.0912` and `c_NN.pt.1.0912` (in this order) to the *Selected data* window.
  - ○ Click the *Accept* button.
- Repeat the same graph using the parts `c_MM`, `c_MN` and `c_NN`.

As expected, the ternary diagrams are very similar. The same effect occurs with the *alr*-, *clr*- and the *olr*-scatter plots.

2.3. CoDa-dendrogram

Activities for Section 2.3

We work with the `mammalsmilk` CoDa set described in the Appendix. This CoDa set contains the percentages of five constituents —water, protein, fat, lactose and ash— in the milk of 24 mammals.

- Load the `mammalsmilk.cdp` file in the CoDaPack package: *File* ▷ *Open Workspace...*.

The variables are:

- `mammal`: type of mammal.
- `code`: numeric code of the mammal.
- `W`: percentage of water in the milk of the mammal.
- `P`: percentage of protein in the milk of the mammal.
- `F`: percentage of fat in the milk of the mammal.
- `L`: percentage of lactose in the milk of the mammal.
- `A`: percentage of ash in the milk of the mammal.

Observe that the 5-part compositions [`W`, `P`, `F`, `L`, `A`] are closed to $\kappa = 100$.

Assume that an analyst is interested in split the role of parts `W` and `A` from the rest of parts `P`, `F`, and `L`. Among these parts, the interest is in the part `P`. Following this requirements, the SBP proposed is

Sign matrix **S** of the SBP

i	W	P	F	L	A	p	n
1	$+1$	-1	-1	-1	$+1$	2	3
2	$+1$	0	0	0	-1	1	1
3	0	$+1$	-1	-1	0	1	2
4	0	0	$+1$	-1	0	1	1

To create the CoDa-dendrogram:

- Go to the *Graphs* ▷ *Balance dendrogram* menu.
  - Select the variables `W`, `P`, `F`, `L` and `A` (in this order) from the *Available data* window and move them into the *Selected data* window.
  - Activate the *Add balances* and the *Add statistics* option.
  - Click the *Define Manually* button and create the SBP.
  - Click the *Accept* button.

Figure 2.1 shows the CoDa-dendrogram derived from the SBP:

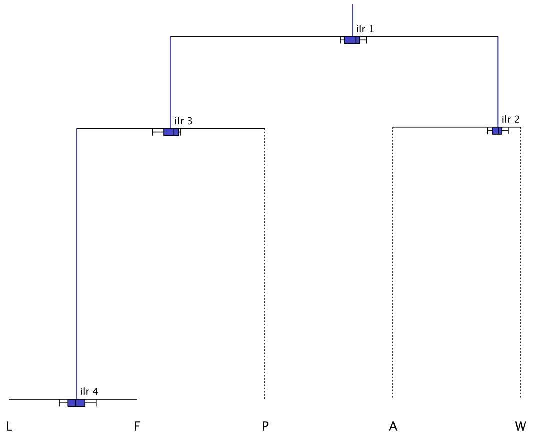


Figure 2.1. A balance dendrogram of the `mammalsmilk` CoDa set.

The fourth balance has the largest variance (0.9138) whereas the shortest vertical link (0.1100) corresponds to the first balance. The logratio $\ln F/L$ retains 55.68% of the total variance ($1.6411 = 0.1100 + 0.3064 + 0.3109 + 0.9138$) of the CoDa set, certainly a large amount for a simple logratio. That is, the ratio F/L varies across the samples, being one of the factors that characterizes the mammals milk. In contrast, the ratio between the average of parts W, A and the average of parts P, F, L has slight variation. The *olr*-coordinate $\frac{1}{\sqrt{2}}\ln(W/A)$ has the largest mean (3.2189) indicating the proportion of water is larger than the ash content: in average, the proportion of ash is 1% the water content ($1\% \approx e^{-\sqrt{2}\cdot 3.2189}$).

2.4. Reduced-dimensionality representation of a compositional data set

Activities for Section 2.4

- Load the `mammalsmilk.cdp` file (see Appendix) in the CoDaPack package: *File* ▹ *Open Workspace...*.

Before making the CoDa-biplot of this set, we start by performing a brief descriptive statistics analysis.

- Go to the *Statistics* ▹ *Compositional statistics summary* menu for a short overview of the compositional statistics of the data set.

In the *Output* window we can observe that:

⋆ The centre of the CoDa set is equal to $\mathbf{g}$ =[`W`, `P`, `F`, `L`, `A`]= [0.8351,0.0506,0.0641, 0.0414, 0.0088].

⋆ The two pairwise logratios with the largest variability are ln(`F`/`L`) (=1.8276) and ln(`P`/`L`) (=1.3620).

⋆ The two pairwise logratios with the smallest variability are ln(`W`/`L`) (=0.1307) and ln(`P`/`A`) (=0.2851).

⋆ The *clr*-variable with largest variability is $\ln(\mathtt{L}/(\mathtt{W}\cdot\mathtt{P}\cdot\mathtt{F}\cdot\mathtt{L}\cdot\mathtt{A})^{1/5})$ (=0.5254).

⋆ The *clr*-variable with smallest variability is $\ln(\mathtt{A}/(\mathtt{W}\cdot\mathtt{P}\cdot\mathtt{F}\cdot\mathtt{L}\cdot\mathtt{A})^{1/5})$ (=0.1692).

⋆ The total variance is equal to 1.6411.

- Go to the *Data* ▹ *Transformations* ▹ *CLR* menu for calculating the *clr*-scores of the compositions.

Now we will draw the CoDa-biplot of the set:

- Go to the *Graphs* ▹ *CLR biplot* menu.
  - ○ Select the variables `W`, `P`, `F`, `L` and `A` (in this order) from the `Available data` window and move them into the `Selected data` window.
  - ○ Activate the `Add coordinates` option.
  - ○ Click the `Accept` button.

CoDaPack creates four new variables —`UD1`, `UD2`, `UD3` and `UD4`— with the coordinates of the compositions of the *centred* set with respect to the *olr*-basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$ associated to the four principal components PC1, PC2, PC3 and PC4, respectively. Moreover CoDaPack writes in the output window the *clr*-scores of $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ and $\mathbf{e}_4$. For example, the *clr*-scores of $\mathbf{e}_1$ (which corresponds to PC1) is equal to $[0.3804, -0.3446, -0.5384, 0.6515, -0.1489]$. In the column labelled `Cum. Prop. Exp.` the cumulative total variance retained by the PCs it is specified: the PC1 retains 74.28% of the total variance, the PC1 + PC2 retains 91.35% and the PC1+PC2+PC3, 99.22%. Therefore, the two-dimensional biplots PC1×PC2 and PC1×PC3 provide good approximations (91.35% and 82.15%, respectively) of the four-dimensional scatterplot of *olr*-coordinates set.

We move to the *Graphic* window where CoDaPack shows the *clr*-biplot. From the small window (situated in the right-hand corner of the *Graphic* window), CoDaPack allows us to modify (from 0 to 1) the parameter α used in the rank-2 approximation of the *clr*-scores set $\mathbf{Z}$ of the centred set $\mathbf{X}$ ($\mathbf{Z} = \operatorname{clr}\mathbf{X}$). For $\alpha = 0$, the resulting (covariance) biplot gives a good representation of the *clr*-variables, whereas for $\alpha = 1$ the (form) biplot provides a good representation of *clr*-scores vectors.

Be sure that the display corresponds to a covariate biplot, that is, the number 0 must be written in the small window in the corner of the graphic window. Figure 2.2 shows the covariance CoDa-biplot for the data set.

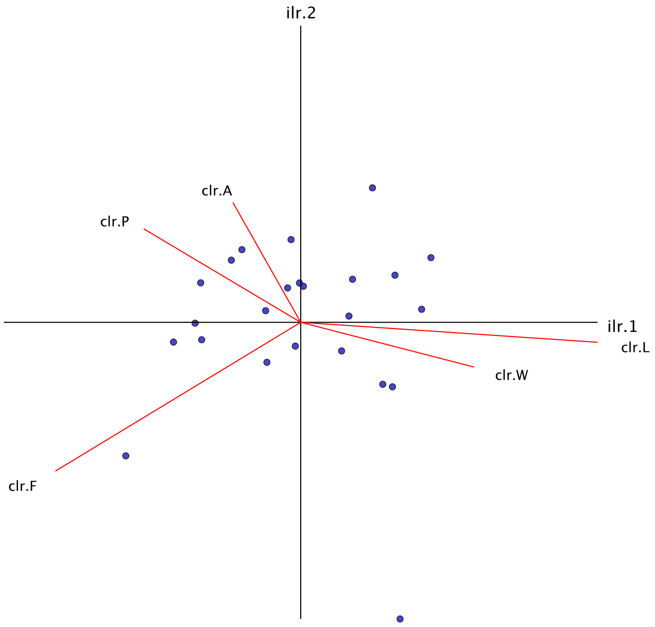


Figure 2.2. Covariance *clr*-biplot of the `mammalsmilk` CoDa set.

- Click alternatively on the *XY* and *XZ* buttons (at the bottom of the graphic window) to see the two 2-dimensional projections of the scatterplot on the PC1×PC2 and on PC1×PC3 planes, respectively.

The plots highlight:

- ⋆ PC1 (the horizontal axis) differentiates the compositions of mammal's milk in the opposition between the `clr.L` and `clr.W` variables (positive part of the axis) in front the other ones. That is, mammal's milk with relative high `L` and `W` proportions have relative low content on the other parts and vice-versa.
- ⋆ PC2 (the vertical axis) differentiates the compositions of mammal's milk in the opposition between `clr.A` and `clr.P` in front `clr.F`.
- ⋆ The largest ray is the `clr.L` ray. This suggests that `clr.L` is the *clr*-variable with the highest variability.
- ⋆ The shortest ray is the `clr.A` ray. This suggests that `clr.A` is the *clr*-variable with the smallest variability.
- ⋆ The largest link is $\overline{\texttt{clr.F}\,\texttt{clr.L}}$. This suggests that ln(`F`/`L`) is the pairwise logratio with higher variability.

⋆ The shortest link is $\overline{\texttt{clr.W}\,\texttt{clr.L}}$. This suggests that ln(`W/L`) is the pairwise logratio with smallest variability.

⋆ The ends of rays `clr.F`, `clr.W` and `clr.L` are approximately collinear. This suggests that the subcomposition sub{F,W,L} has an approximate one-dimensional compositional pattern.
  To confirm this perception
  - Go to *Graphs* ▷ *Ternary/Quaternary Principal Components*.
    - ○ Move the variables `W`, `F`, and `L` to the *Selected data* window.
    - ○ Click the *Accept* button.

  CoDaPack plots the ternary diagram of sub{W,F,L} and also the PC1 and PC2 directions of the subcomposition.
  - Activate the *Centered* and *Grid* options (at the bottom of the graphic window) to centre the data and get a better view of the pattern.

  The graph confirms the compositional linear pattern of the subcomposition sub{W,F,L}. Moreover, in the *Output* window, we see how PC1 retains a high percentage (97.89%) of the total variance of this subcomposition. In the *Output* window, CoDaPack writes the compositions — $[0.3707, 0.1299, 0.4993]$ and $[0.1321, 0.3414, 0.5265]$— that give the directions of the PC1 and PC2 axes of sub{W,F,L}, respectively. These two compositions are unitary (its norm is equal to one) and orthogonal (the inner product is equal to zero).

Come back to the PC1×PC2 (covariance) biplot of the full data set.

⋆ Graphically, the links $\overline{\texttt{clr.F}\,\texttt{clr.P}}$ and $\overline{\texttt{clr.L}\,\texttt{clr.P}}$ seem to be approximately orthogonal. This suggests that the correlation between the logratios, ln(`F/P`) and ln(`L/P`), is approximately equal to 0.
  - ○ Go to *Data* ▷ *Transformations* ▷ *ALR*.
  - ○ Move the variables `F`, `L`, and `P` (in this order) to the *Selected data* window.
  - ○ Click the *Accept* button.
  - ○ Go to *Statistics* ▷ *Classical statistics summary*.
  - ○ Move the variables `alr.F_P` and `alr.L_P` to the *Selected data* window.
  - ○ Uncheck all the options but the *Correlation Matrix* option.

  The correlation coefficient (Pearson) between these two logratios is equal to 0.0116 (*R* program: *p*-value = 0.957). Therefore, the null hypothesis of independence (i.e. null correlation) between ln(`F/P`) and ln(`L/P`) cannot be rejected.

⋆ Similarly, the links $\overline{\texttt{clr.F}\,\texttt{clr.P}}$ and $\overline{\texttt{clr.W}\,\texttt{clr.P}}$ are approximately orthogonal. Thus, the correlation coefficient between the logratios, ln(`F/P`) and ln(`W/P`), is equal to 0.1190 (*R* program: *p*-value = 0.580). Therefore, the null hypothesis of independence between the two logratios again cannot be rejected.

Come back to the biplot graph.

- Change the value in the small window on the corner of the *Graphic* window to 1.0 to obtain the PC1×PC2 (form) biplot of the full data set.
  This option allows us to obtain a good representation of observations (Fig. 2.3).
- Go to the *Data* ▷ *Add observation names...* menu at the top left of the graphic window.
  - ○ Select the variable `mammal`.
  - ○ Click the button *Accept*.

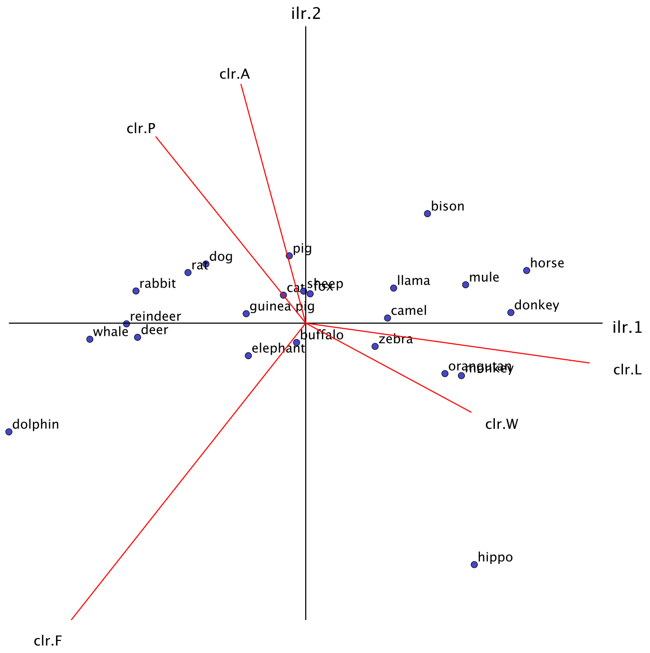


Figure 2.3. Form *clr*-biplot of the `mammalsmilk` CoDa set.

- Click the *Data* ▷ *Show observation names...* option at the top left of the graphic window.
- Click alternatively on the XY and XZ buttons (on the bottom of the graphic window) to see the two two-dimensional projections of the scatterplot on the PC1×PC2 and on PC1×PC3 planes, respectively.
- Drag the cursor over the graph to obtain a three-dimensional representation of the projections of observations on the PC1×PC2×PC3 subspace.

The plots can be used for:

⋆ Visualizing the similarity between some compositions of mammal's milk.
For example, horse's and donkey's milks seem to have a similar composition [`W`, `P`, `F`, `L`, `A`] because they are very close to each other. Both are in the positive part of the first PC axis (*ilr*.1) suggesting relative large values in parts `W` and `L` and small in the rest. In fact, the respective compositions are $[89.25, 2.58, 0.99, 6.84, 0.35]$ and $[90.30, 1.70, 1.40, 6.20, 0.40]$.

⋆ Detecting possible outlier compositions of mammal's milk.
For example, the composition of dolphin's milk $[48.89, 11.54, 38.01, 0.98, 0.58]$ is a potential outlier with respect to other compositions of mammal's milk.

⋆ Identifying the compositions more *representative* of the whole population, that is, those nearest to the centre of the biplot.
Thus, the buffalo's milk $[80.98, 5.82, 7.79, 4.64, 0.77]$ composition is quite *similar* to the centre $\mathbf{g} = [0.8351, 0.0506, 0.0641, 0.0414, 0.0088]$ of the CoDa set.

Finally, suppose that we want to investigate whether there is any subcomposition (with the minimum number of parts) capable of retaining a 90% or more variability of the full CoDa set.

To do this, we must calculate the total variance of the successive 4-part subcompositions [W, P, F, L, A] of composition.

- Go to the *Statistics* ▷ *Compositional statistics summary* menu.
  - Move the four variables W, P, F and L to the *Selected data* window.
  - Enable the *Percentile* and *Variation Array* options and keep active the *Centre* and *Total Variance* options.
  - Click on the *Accept* button.
- Repeat the same procedure with respect to the subcompositions [W, P, F, A], [W, P, L, A] and [P, F, L, A].

Remember that the total variance of the full CoDa set is equal to 1.6411.

The subcomposition [W, P, F, L] is the one that retains highest variability. Its total variance is equal to 1.4296, lower than 90% of the total variance of the full compositions.

2.5. Principal balances

Activities for Section 2.5

We are interested in creating PBs for the 5-part compositions [W, P, F, L, A] recorded in the mammalsmilk.cdp file (Appendix).

- Load the mammalsmilk.cdp file in the CoDaPack package: *File* ▷ *Open Workspace...*.

The *clr*-biplot (Fig. 2.2) suggests that the candidate for the first PB is associated to the partition $(+1, -1, -1, +1, -1)$. That is, we split the *clr*-variables into the positive (W, L) and negative (P, F, A) part of the first PC axis (*ilr*.1). According the second PC axis, the set (P, F, A) can be split into (P,A) and F. Following this procedure, the SBP proposed for the PBs is:

Sign matrix **S** of the SBP

i	W	P	F	L	A	p	n
1	+1	−1	−1	+1	−1	2	3
2	+1	0	0	−1	0	1	1
3	0	+1	+1	0	−1	2	1
4	0	+1	−1	0	0	1	1

To analyze the decomposition of the variance:

- Go to the *Data* ▷ *Transformations* ▷ *Raw-ILR* menu.
  - Move the five variables W, P, F, L and A (in this order) to the *Selected data* window.

- ○ Click on the *Define Manually* button.
- ○ Introduce the SBP proposed.
- ○ Click on the *Accept* button.

- Go to the *Statistics* ▷ *Classical statistics summary* menu.
  - ○ Calculate the covariance matrix of the *olr*-variables $(ilr.1, ilr.2, ilr.3, ilr.4)$ created above.

The variance of the four *olr*-variables are $(1.1054, 0.0653, 0.2281, 0.2422)$ which respectively retain $(67.36\%, 3.98\%, 13.90\%, 14.76\%)$ of total variance. Consequently, the PBs are $(PB1, PB2, PB3, PB4) = (ilr.1, ilr.4, ilr.3, ilr.2)$.

A different approach to create a PB set consists of imitating the constrained PCs algorithm. The following table shows the PC loadings and % of variance retained by the PCs in the *clr*-biplot (Section 2.4):

						Cum. Prop.
PC	clr.W	clr.P	clr.F	clr.L	clr.A	Ret. (%)
PC1	0.3804	−0.3446	−0.5384	0.6515	−0.1489	74.28
PC2	−0.2048	0.4285	−0.6815	−0.0914	0.5492	91.35
PC3	0.2487	0.6762	−0.1749	−0.0842	−0.6658	99.22
PC4	−0.7426	0.2011	0.1229	0.6000	−0.1814	100.00

For example, the first vector (PC1= $ilr.1{=}\gamma_1$) of the *olr*-basis is the logcontrast

$$\gamma_1 = 0.3804 \ln W - 0.3446 \ln P - 0.5384 \ln F + 0.6515 \ln L - 0.1489 \ln A \ ,$$

which retains 74.28% of the total variance. The possibilities for approximating $\boldsymbol{\gamma}_1$ by a balance $\boldsymbol{\alpha}_1$ are

	Sign vector $\mathbf{s}_1$ for $\boldsymbol{\alpha}_1$						
i	W	P	F	L	A	p	n
1	0	0	−1	+1	0	1	1
2	+1	0	−1	+1	0	2	1
3	+1	−1	−1	+1	0	2	2
4	+1	−1	−1	+1	−1	2	3

The following table shows the possible PBs vectors $\boldsymbol{\alpha}_1$ derived from each possible sign vector $\mathbf{s}_1$. The variance and total variance retained by each one is also provided.

							Cum. Prop.
$PB1.i$	$\ln W$	$\ln P$	$\ln F$	$\ln L$	$\ln A$	Variance	Ret. (%)
PB1.1	0	0	$-\frac{\sqrt{2}}{2}$	$+\frac{\sqrt{2}}{2}$	0	0.9138	55.68
PB1.2	$+\frac{1}{2}\sqrt{\frac{2}{3}}$	0	$-\sqrt{\frac{2}{3}}$	$+\frac{1}{2}\sqrt{\frac{2}{3}}$	0	0.9626	58.66
PB1.3	$+\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$+\frac{1}{2}$	0	1.1220	68.37
PB1.4	$+\frac{1}{2}\sqrt{\frac{6}{5}}$	$-\frac{1}{3}\sqrt{\frac{6}{5}}$	$-\frac{1}{3}\sqrt{\frac{6}{5}}$	$+\frac{1}{2}\sqrt{\frac{6}{5}}$	$-\frac{1}{3}\sqrt{\frac{6}{5}}$	1.1054	67.36

For example, consider the vector $\boldsymbol{\alpha}_1 = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, 0)$ to illustrate the calculation of variance retained (1.1220).

- Go to the *Data* ▷ *Manipulate* ▷ *Calculate new Variable* menu.
  - ○ Move the four variables `W`, `P`, `F`, and `L` (in this order) to the *Selected data* window.
  - ○ Click on the *Accept* button. A new window *Create new Variable* is opened.
  - ○ Introduce the expression $1/2 * log(x1) - 1/2 * log(x2) - 1/2 * log(x3) + 1/2 * log(x4)$ into the *Enter expression* cell.
  - ○ Introduce the name *PB1.3* into the *Enter new variable name* cell.
  - ○ Click on the *Ok* button.

  CoDaPack creates the column *PB1.3*.
- Go to the *Statistics* ▷ *Classical statistics summary*.
  - ○ Calculate the covariance matrix of the column *PB1.3* created above.

  The variance of the variable (1.1220) is shown in the *Output* window.

We should do a similar procedure for calculating the variance of the other possible vectors $\boldsymbol{\alpha}_1$. If we select the vector $PB1 = \boldsymbol{\alpha}_1 = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, 0)$ then there is only one possible *parent* (change of sign excepted) and there is only one possibility for each *child* (change of sign excepted) to complete the SBP. The following table shows the sign matrix of the SBP and the variance retained by each *olr*-vector of the basis. The PBs set are these *olr*-vectors ordered by the variance retained.

	Sign matrix for PBs								Cum. Prop.
i	W	P	F	L	A	p	n	Variance	Ret. (%)
PB3	−1	−1	−1	−1	+1	1	4	0.2115	12.89
PB1	+1	−1	−1	+1	0	2	2	1.1220	68.37
PB4	+1	0	0	−1	0	1	1	0.0653	3.98
PB2	0	+1	−1	0	0	1	1	0.2422	14.76

The first PB retains 68% of total variance (1.6411), a reasonable approximation of the variance retained by the first PC (74.28%).

2.6. Distributions on the simplex

Activities for Section 2.6

CoDaPack includes the possibility of plotting the predictive and confidence regions in the ternary. Also, for the general case, one can test if the normal distribution on the simplex fits well a sample distribution.

- In CoDaPack, load the `alimentation.cdp` file (see Appendix).

The `alimentation` CoDa set contains the percentages of consumption of several kinds of foodstuffs in European countries during the 80s. The variables are:

- `RM`: red meat (pork, veal, beef)
- `WM`: white meat (chicken)

- `E`: eggs
- `M`: milk
- `F`: fish
- `C`: cereals
- `S`: starch (potatoes)
- `N`: nuts
- `FV`: fruits and vegetables

The file also contains more information:

- `Country`: name of the country
- `CountryID`: two letters identifying the country
- `North/Med`: `north` = Northern European country; `med` = Mediterranean or Southern European country
- `East/West`: `East` = Eastern (pro-soviet) country; `West`: Western (pro-USA) country

We want to analyse whether a normal distribution on the simplex $\mathcal{S}^3$ fits the subcomposition [`WM`, `F`, `N`].

To calculate the estimates of the parameters $\boldsymbol{\mu}^*, \boldsymbol{\Omega}$ of [`WM`, `F`, `N`]:

- Go to the *Data* ▷ *Transformations* ▷ *Raw-ILR* menu and transform the variables `WM`, `F` and `N` (select the variables in the above order and use the default partition).
- Go to the *Statistics* ▷ *Classical statistics summary* menu and calculate the *mean* and the *covariance matrix* of the logratio variables `ilr.1` and `ilr.2`.

The mean vector and the covariance matrix are

$$\widehat{\boldsymbol{\mu}^*} = [0.4840, 0.6148] \qquad \widehat{\boldsymbol{\Omega}} = \begin{bmatrix} 0.9022 & -0.2388 \\ -0.2388 & 0.6316 \end{bmatrix}, \tag{2.1}$$

which are the estimates of the expectation and the covariance of a normal distribution that we can fit to the [`WM`, `F`, `N`] observations.

Assuming that [`WM`, `F`, `N`] follows a normal distribution on the simplex, we can plot the predictive curves (for a given predictive levels) on the ternary diagram:

- Go to the *Graphs* ▷ *Predictive Region* menu;
  - Select the variables `WM`, `F` and `N`;
  - Write `0.10 0.25 0.75 0.90` in the *Predictive level* box.
  - Click the *Accept* button.

It is easy to interpret the predictive curves plotted in the ternary (Figure 2.4). For example, the most internal curve corresponds to the predictive level $1 - \alpha = 0.10$. It means that a sample from a normal distribution on $\mathcal{S}^3$ (with parameters equal to those of Equation 2.1) will be in the region bounded by this curve with a probability equal to 0.1. That is, one expects 10% of samples in this region.

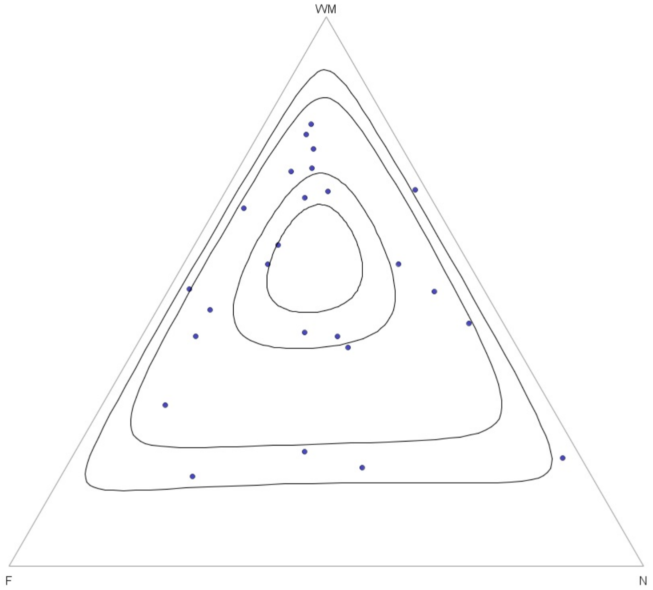


Figure 2.4. 10%, 25%, 75% and 90% predictive regions for the [`WM`, `F`, `N`] subcomposition in the `alimentation` CoDa set.

We can also plot *centre confidence curves* (for a given confidence levels) on the ternary diagram:

- Go to the *Graphs* ▷ *Center Confidence Region* menu;
  - ○ Select the variables `WM`, `F` and `N`;
  - ○ Write `0.95` in the *Confidence level* box.
  - ○ Click the *Accept* button.

The plotted curve is the boundary of a 95% confidence region for the mean of the normal distribution on $\mathcal{S}^3$ which fits the [`WM`, `F`, `N`] samples. That is, we have a 95% confidence that the true mean of the normal random composition falls into the region.

With CoDaPack we can test if a normal distribution on $\mathcal{S}^D$ fits a CoDa set well:

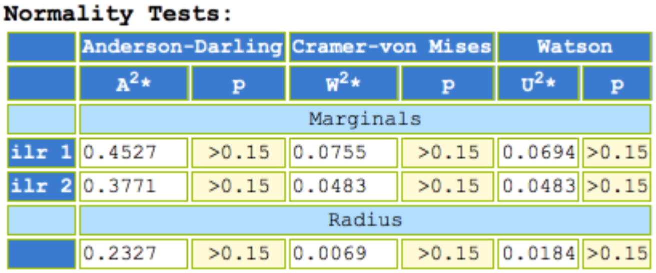
- Go to the *Statistics* ▷ *Log-Ratio Normality Test* menu.
  - ○ Select the variables `WM`, `F` and `N`.
  - ○ Select the *Default Partition* in the *Options* window.
    This selection implies that the normality tests will be applied to the *olr*-coordinates associated to the *Default Partition*, that is, to the *olr*-variables

$$\mathrm{ilr}_1 = \sqrt{\frac{2}{3}} \ln \frac{(\mathtt{WM} \cdot \mathtt{F})^{1/2}}{\mathtt{N}} \qquad \mathrm{ilr}_2 = \sqrt{\frac{1}{2}} \ln \frac{\mathtt{WM}}{\mathtt{F}}$$

  - ○ Click the *Accept* button.

CoDaPack calculates the Anderson-Darling (A^2*), Cramér-von Mises (W^2*) and Watson (U^2*) statistics for: 1) the marginal univariate distributions ilr_1 and ilr_2;

and 2) the multivariate radius test. It also gives the approximate p-values associated to the statistics (Fig. 2.5).



Normality Tests:

	Anderson-Darling		Cramer-von Mises		Watson	
	A2*	p	W2*	p	U2*	p
	Marginals					
ilr 1	0.4527	>0.15	0.0755	>0.15	0.0694	>0.15
ilr 2	0.3771	>0.15	0.0483	>0.15	0.0483	>0.15
	Radius					
	0.2327	>0.15	0.0069	>0.15	0.0184	>0.15

Figure 2.5. Log-ratio normality tests for the [`WM`, `F`, `N`] subcomposition in the `alimentation` CoDa set.

Observe that the three p-values of the radius test are above 0.15 suggesting failure to reject the null hypothesis H_0*: The logratio normal model fits well the CoDa set.* The same holds for all p-values of the marginal univariate distributions ilr 1 and ilr 2.

The results of marginal normality tests depend on the *olr*-basis on which the coordinates of compositions are expressed.

- Repeat the tests of logratio normality for the same subcomposition [WM, F, N] using the *ilr*-variables associated to the *Partition*: (+1, −1, +1) and (+1, 0, −1).

Observe that the statistics and p-values for the radius test are invariant. Once more, all p-values for the univariate tests are greater than 0.15 but the test statistic values differ. This example suggests that in the situations in which we reject the multivariate normality and want to investigate the marginals, we have to analyse several possibilities as regards the SBP in the *olr*-basis.

The normality on the simplex is invariant by the group of perturbations, that is, if a CoDa set $\mathbf{X}$ is well fitted by a normal distribution $\mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}^*, \boldsymbol{\Omega})$, then the perturbed data set $\mathbf{p} \oplus \mathbf{X}$ will also be well fitted by the normal distribution $\mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}^* + \operatorname{olr}\mathbf{p}, \boldsymbol{\Omega})$ (we assume that the parameters of the normal distribution are estimated from the *olr*-coordinates). Let's check this property.

- Go to the *Data* ▷ *Operations* ▷ *Perturbation* menu. Apply the perturbation $\mathbf{p} = [1, 1, 10]$ to the subcompositions [`WM`, `F`, `N`].
- Go to the *Data* ▷ *Transformations* ▷ *Raw-ILR* menu. Calculate the *olr*-coordinates of the perturbed data set [`WM.x.1.0`, `F.x.1.0`, `N.x.10.0`] with the default partition.
- Go to the *Statistics* ▷ *Classical statistics summary* menu. Calculate the mean and the covariance matrix of the *olr*-coordinates `ilr.1.c` and `ilr.2.c`. Compare both estimated parameters with the corresponding parameters estimated from the original subcompositions.

As expected, covariance matrices are equal whereas the *olr*-coordinates of the center $\boldsymbol{\mu}^* = (0.4840, 0.6148)$ moves to $\boldsymbol{\mu}^* + \operatorname{olr} \mathbf{p} = (-1.3960, 0.6148)$. That is,

$$\operatorname{olr} \mathbf{p} = (-1.3960, 0.6148) - (0.4840, 0.6148) = (-1.88, 0.00).$$

Note that this shift is the same for each sample in the CoDa set. Indeed, the *olr*-coordinates of the first sample are $(-1.91, 1.38)$ that moves to $(-3.79, 1.38)$ by $(-1.88, 0.00)$.

- Go to the *Graphs* ▷ *Predictive Region* menu. Draw the predictive curves for the perturbed subcompositions at levels 0.10, 0.25, 0.75 and 0.90. Compare the plot with the predictive curves adjusted to the original subcompositions.
  We observe that the CoDa set and the predictive regions move to the vertex N. This is because the perturbation $\mathbf{p} = [1, 1, 10]$ increases the ratios N/WM and N/F by a factor 10/1.
- Go to the *Graphs* ▷ *Center Confidence Region* menu. Draw the centre confidence region of the mean (at level 95%) for the perturbed subcompositions. Compare the plot with the 95% center confidence region calculated from the original subcompositions.
  We observe an analogous effect than for the predictive regions.
- Go to the *Statistics* ▷ *Log-Ratio Normality Test* menu. Compute the normality tests for the perturbed subcompositions. Compare the normality tests calculated from the original subcompositions.
  Importantly, both tests, univariate marginals and radius test, are invariant under perturbation of the CoDa set.

The chapter's key concepts

✓ The closed geometric mean is the measure of central tendency for a CoDa set.

✓ The covariance structure of a CoDa set can be described in different ways from the variation matrix or from the covariance matrices of the *clr*, *alr* and *ilr* transformed data set.

✓ The total variance is an overall measure of the total relative variability of a CoDa set.

✓ CoDa-dendrogram summarizes the coordinates created by an SBP.

✓ The biplot of the *clr*-transformed data set allows us to perform a first exploratory analysis of a CoDa set so as to discover significant statistical relationships between logratios of the parts and potential clusters of ‘similar’ compositions.

- ✓ Principal balances algorithms provide a particular type of SBP created by a data-driven procedure.
- ✓ Normal distribution on the simplex is a probability model that is consistent with simplicial geometry.

Chapter 3

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.

3.1. Missing data

Activities for Section 3.1

With the goal of evaluating the performance of the log-ratio EM algorithm to missing data, we will use the Montana CoDa set (see Appendix). This is a real data set originally without missing data. The data consists of $N = 229$ samples of the concentration (in ppm) of $D = 5$ minor elements [Cr, Cu, Hg, U, V] in carbon ashes from the Fort Union formation (Montana, USA). Actually, this vector of elements represents a subcomposition of a much larger composition available, and the data are not closed to a constant sum.

We shall import the Excel file Montana.xls containing the data. This file contains two sheets: Montana and MontanaNA. Note that the variables are the chemical elements [Cr, Cu, Hg, U, V] and a factor type indicating whether the original sample will have missing values or not. In the Montana sheet, the chemical variables have their original values, that is, the chemical data set is complete. After we randomly forced missing values, in the parts [Cr, Cu, Hg, U] of those samples where the factor type is equal to *Missing*, the resulting data set is recorded in the MontanaNA sheet. The label *NA* (Not Available) was used in Excel to indicate that the original value was forced to missing.

- In the menu *File* ▷ *Import* ▷ *Import XLS Data*, we select the Excel file Montana.xls
  - Write *Montana* under *Data Frame Name* in the menu *Import*.
  - Click the button *OK* in the menu *Import*.
  - Select the sheet Montana (double click) in the window *Choose one sheet*.
- Go again to the menu *File* ▷ *Import* ▷ *Import XLS Data* to select the Excel file Montana.xls
  - Write *MontanaNA* under *Data Frame Name* in the menu *Import*.
  - Click the button *OK* in the menu *Import*.
  - Select the sheet MontanaNA (double click) in the window *Choose one sheet*. Note the default “NA” label in the option *Non available data*.

We suggest to save a copy as a CoDaPack file (use the menu *File* ▷ *Save workspace...*).

- Select the complete Montana data set in box *Tables* (at the upper left of CoDaPack window).
- Draw the covariance *clr*-biplot using the variable type as a group (*full obs* and *missing*; Fig. 3.1).
  - How is the pattern of the two groups of samples ($missing/fullobs$) in the biplot? Are they separated? Mixed?
  - Does it suggest that missingness is completely at random (MCAR)?

  The two PC axes of the biplot account 75.95% of the variance, suggesting a quite high quality of the plot. Blue and red circles are mixed in Fig. 3.1 suggesting no dependence of the missingness as regards the observed values (MAR) or the missing values (NMAR). In addition, we do not observe a relevant number of samples far from the center, suggesting one can assume there are not potential outliers.

- Calculate the variation array grouping the results by `type`.
  - ○ How are the variation arrays?
  - ○ Do they suggest that missingness is completely at random (MCAR)?

  The two variation arrays have a similar structure. Indeed, the highest pairwise log-ratio variances are in $\ln(Cr/Hg)$, $\ln(U/Hg)$, and $\ln(V/Hg)$. In addition, $\ln(Cr/V)$, and $\ln(Cu/V)$ have the smallest variances in both groups. Moreover, expected values of the pairwise log-ratio are very similar in both variation arrays. These results suggest that the forced missing values are MCAR.

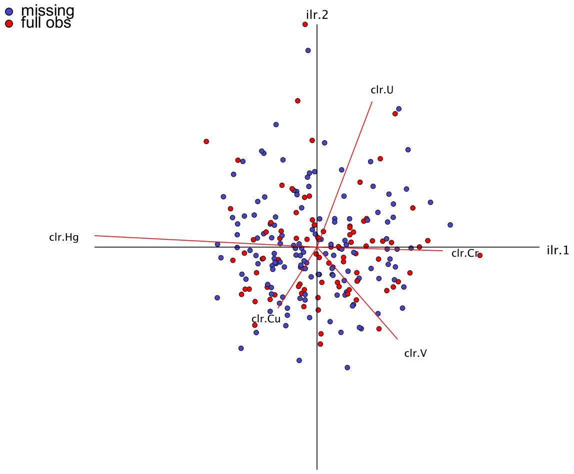


Figure 3.1. Covariance *clr*-biplot. Original `Montana` CoDa set. Blue circles are samples used to force missing values.

When missingness is at random or completely at random the missing data can be imputed using the log-ratio EM algorithm.

- Take the data frame `MontanaNA`.
- Go to the menu *Irregular data* ▷ *ZPatterns plot* (Fig. 3.2). Select all the parts. Click at the options *Show means* and *Show percentages*. Introduce NA as the label into the box *Label*. Click the *Accept* button.
  - ○ How are the percentages of NA values by parts? (bars at top)
  - ○ Are there relevant relative differences among the geometric means across the different patterns?
  - ○ Do they suggest that missingness is completely at random (MCAR)?

  Numerical results shown by CoDaPack indicate that the overall percentage of cells in data matrix with NA values is quite high: 15.37%. Top bars show that the percentage of randomly forced NA values in the four parts [Cr, Cu, Hg, U] are very similar, suggesting that the probability of a forced NA value is the same. Right side bars show that quite a high percentage of samples (55.46%) have at least one NA value. The percentages of patterns with only one NA value is similar ($9\% - 10\%$ approx.) and similar behavior is observed respectively for patterns with two and three NA values. In addition, these percentages (one NA, two NA, and three NA) approximately decrease in a geometric series suggesting a random process as origin of the NA values.

On the other side, there are no relevant relative differences between the geometric means. For example, the ratios Cu/Cr take respectively the values 38.66/3.24, 40.74/2.84, 40.05/3.06, and 40.12/2.75 across the patterns where both parts are observed parts. For other pairwise ratios the conclusion is similar. We have not found evidences pointing that the missingness is MAR or NMAR.

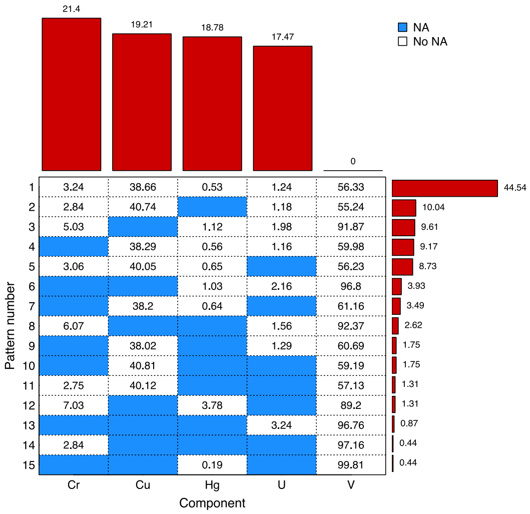


Figure 3.2. NA patterns plot for forced missing values in the `Montana` CoDa set. Blue cells are NA values.

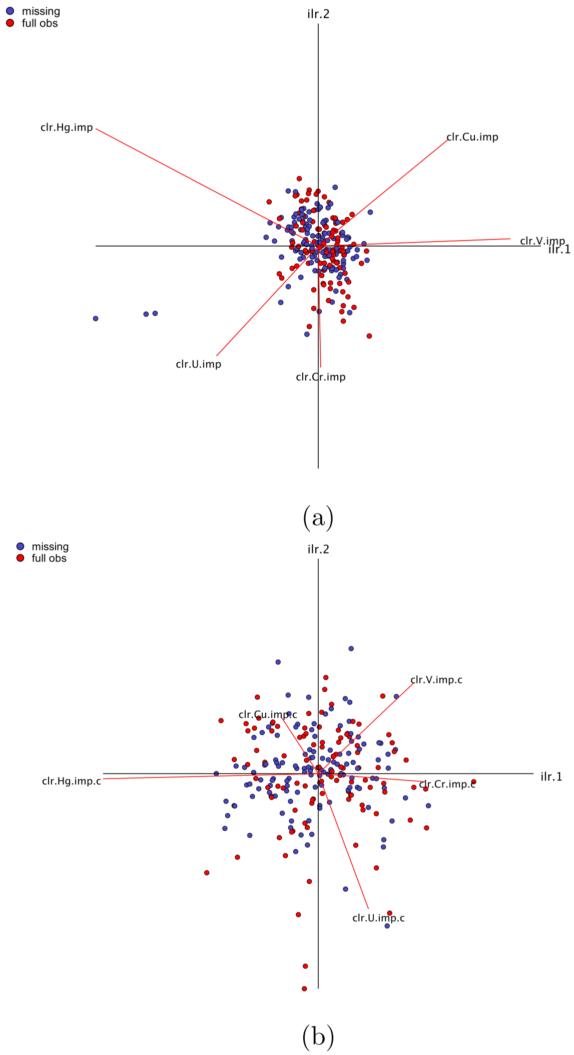
We proceed to impute NA values using CoDaPack:

- Go to the menu *Irregular data* ▷ *Logratio-EM missing replacement*.
  - ○ Move the variables `Cr`, `Cu`, `Hg`, `U` and `V` to the *Selected data* window.
  - ○ Keep default options (that is, do not select the *Rob* robust option).
  - ○ Click on the button *Accept*.

CoDaPack requires each sample to have at least two observed parts. Additionally, at least one part has to have complete data for all samples (unless the robust option is selected). In any case, the `MontanaNA` CoDa set fulfills both conditions. CoDaPack created the new variables `Cr.imp`, `Cu.imp`, `Hg.imp`, `U.imp` and `V.imp` containing the imputed values obtained from the application of the log-ratio EM-replacement algorithm.

- Draw a covariance biplot with the imputed variables `Cr.imp`, `Cu.imp`, `Hg.imp`, `U.imp` and `V.imp` using the variable `type` as a group (Fig. 3.3a).
- Compare the biplot of the original and the imputed data sets.

- Repeat the imputation and the biplot by selecting the robust option in the menu *Irregular data* ▷ *Logratio-EM missing replacement* (Fig. 3.3b).
- Compare the biplot of the original and the imputed data sets (the new imputed variable names are `Cr.imp.c`, `Cu.imp.c`, `Hg.imp.c`, `U.imp.c` and `V.imp.c`).
- Did the robust option prevent new potential outliers among the imputed samples?



(a)

(b)

Figure 3.3. Covariance biplot of the imputed `MontanaNA` CoDa set: (a) classical log-ratio EM algorithm; (b) robust log-ratio EM algorithm.

The biplot in Fig. 3.3a retains 85.58% of the variance, suggesting a quite high quality of the plot. The biplot shows three samples that may be considered as potential outliers because they are far from the center of the data set. The presence of these samples affects the scale of the axes. Therefore, the cloud of the data set seems to be very compact. The biplot in Fig. 3.3b retains 79.08% of the variance, suggesting as well a quite high quality of the projection of the data set into the PCs. When comparing the rays of the variables in the biplots it is observed that the robust option preserves better the covariance structure, preventing the effect of new potential outliers. Consequently, we recommend to make the planned statistical analysis –DA, cluster, or MANOVA, among other– using the complete data set provided by the robust log-ratio EM algorithm.

3.2. Essential zeros

Activities for Section 3.2

We shall import the `Excel` file `foraminiferal.xls` (see Appendix) containing compositional variables (closed to one) and a factor variable `zeros`, indicating if the sample has or not a zero value. For this activity we assume that the zeros in the 30 samples are *absolute/essential* zeros.

- In the menu *File* ▷ *Import* ▷ *Import XLS Data...*, select the *Excel* file *foraminiferal.xls*. Click the button *Open*.
- Keep the default options and click the button *Ok* in the *Import Menu*.

We suggest to save a copy of the file as a CoDaPack file (use the menu *File* ▷ *Save as...*) because the data set will be used in other activities.

The data set has four parts `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`, and the zeros are recorded as 0. Observe that there are 5 zeros in parts `glob_obesa` and `glob_triloba` (samples 7, 17, 21, 25 and 30). Because variable `code` is not useful we decide to remove it. In the menu *Data* ▷ *Delete variables*, select the variable *code*. Click the button *Accept*. Save a copy of the file as a CoDaPack file (use the menu *File* ▷ *Save Workspace...*).

To explore the data set:

- Plot the quaternary diagram of the variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`.

○ What happens to the observations with zeros?

CoDaPack shows a message where we are informed that samples with zeros are not included in the plot. Figure 3.4 shows the compositions without zeros. We observe that the cloud is close to the vertex `neogl_atl` suggesting the samples take the highest values in this part.

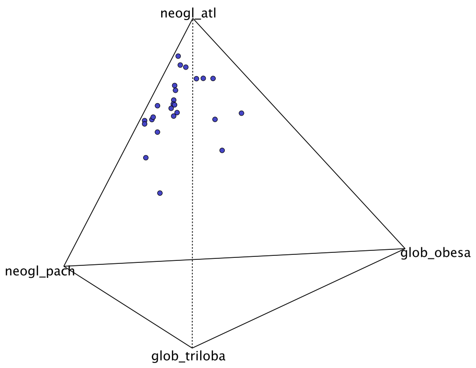


Figure 3.4. Compositions without zeros of the `foraminiferal` CoDa set in the quaternary plot.

Once we have *detected* that the data set has zeros then we proceed to analyze the pattern of zeros:

- Go to menu *Irregular data* ▷ *ZPatterns plot*.
  - ○ Select the variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`.
  - ○ Select *Show Percentages* and *Show means* in the window *Options*.
  - ○ Click the button *Accept*.
- Interpret the zero-patterns plot (Fig. 3.5).

Figure 3.5 indicates that the data set has only three patterns of zeros, where 83.33% of samples are full observed. Three samples (i.e., 10%) have a zero in the part `glob_obesa` and two samples (i.e., 6.67%) have a zero in the part `glob_triloba`. When one analyzes the ratios between parts of the geometric means across the patterns then we observe some differences. For example, the ratio $\frac{\texttt{neogl\_atl}}{\texttt{neogl\_pach}}$ takes respectively the values 69.39/23.13, 59.9/35.87, and 54.06/24.74. Or, the ratio $\frac{\texttt{neogl\_pach}}{\texttt{glob\_obesa}}$ takes the value 23.13/4.26 in the first pattern and the value 24.74/21.2 in the third pattern. Although, the sample sizes are very small, it is worth to analyze if there are relevant differences among the patterns.

To analyze and test the differences among patterns we need to export the data to the program R.

- Go to menu *File* ▷ *Export* ▷ *Export Data to R Data...*.
- Select the four variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`.
- Click the button *Accept*.
- Insert a name for your R-data set (e.g., `foraminiferal.RData`), select your appropriate folder and click the button *Save*.

In R or RStudio,

- Load your data (e.g., `foraminiferal.RData`) and check you have the dataframe *data*.

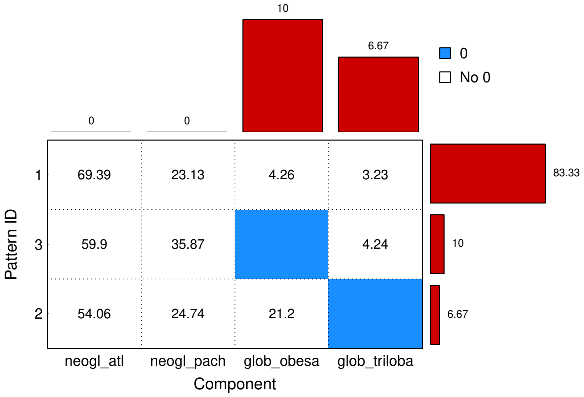


Figure 3.5. Zero patterns in the `foraminiferal` data set.

- Load the package *zCompositions*.
- To get the variation array by pattern, execute the instruction *zVarArray(data)* (Table 3.1).
- Compare the variation arrays of the three patterns and the overall variation array.

As we described above, the expected value of the logratios suggests some differences. For example the logratio between the parts `neogl_pach` and `glob_obesa` takes an expected value 1.69 for the first pattern whereas in the second pattern it takes the value 0.15. In Table 3.1 one can detect some differences for the values of the log-ratio variance among the different patterns. For example, for the first pattern, the log-ratio variance $\mathrm{Var}(\mathit{log}(\texttt{glob\_obesa}/\texttt{neogl\_atl}))) = 1.23$, whereas it takes the value 0.06 for the second pattern. These differences should be tested using the function *zVarArrayTest*. The instruction *zVarArrayTest(data, label = 0, groups = NULL, b = 1000)* returns the P-value for homogeneity of log-ratio variances equal to 0.1459 and 0.169 for the homogeneity of log-ratio means. Consequently, it fails to reject the null hypothesis of homogeneity, the differences observed are not significant.

When the differences among the variation arrays are significant, then one should analyze *where* are these differences. That is, one should analyze which parts and patterns are involved in these differences. Table 3.2 shows the *relative importance* (%) of square relative errors for each pairwise ratio in the variation array returned by the instruction *zVarArrayError(data)*. In the results, we also see that the total squared relative errors (*$TotalSREVars* and *$TotalSREmeans*), being accumulated across all patterns and pairwise logratios, are 0.4629 for the log-ratio variances and 0.1245 for the log-ratio means. According to the P-Values calculated above, both errors are not significant. In Table 3.2, one can observe that the ratio $\frac{\texttt{neogl\_atl}}{\texttt{neogl\_pach}}$ has the largest weight in the differences across log-ratio variances (56.29%), whereas for the expectations the largest value is 47.08%, associated to the ratio $\frac{\texttt{glob\_obesa}}{\texttt{neogl\_pach}}$. When

Table 3.1. Variation array of `foraminiferal` CoDa set by pattern: (a) full observed samples; (b) zero in part `glob_obesa`; (c) zero in part `glob_triloba`; (d) overall variation array.

	neogl_atl	neogl_pach	glob_obesa	glob_triloba
neogl_atl	0.00	0.33	1.23	0.86
neogl_pach	1.10	0.00	1.95	0.66
glob_obesa	2.79	1.69	0.00	2.89
glob_triloba	3.07	1.97	0.28	0.00

(a)

	neogl_atl	neogl_pach	glob_obesa	glob_triloba
neogl_atl	0.00	1.02	0.06	NA
neogl_pach	0.78	0.00	1.57	NA
glob_obesa	0.94	0.15	0.00	NA
glob_triloba	NA	NA	NA	0.00

(b)

	neogl_atl	neogl_pach	glob_obesa	glob_triloba
neogl_atl	0.00	0.15	NA	0.06
neogl_pach	0.51	0.00	NA	0.25
glob_obesa	NA	NA	0.00	NA
glob_triloba	2.65	2.14	NA	0.00

(c)

	neogl_atl	neogl_pach	glob_obesa	glob_triloba
neogl_atl	0.00	0.36	1.14	0.77
neogl_pach	1.02	0.00	1.92	0.61
glob_obesa	2.65	1.58	0.00	2.89
glob_triloba	3.02	1.99	0.28	0.00

(d)

one adds the argument *breakdown = TRUE* at the instruction (that is, *zVarArrayError(data, breakdown = TRUE)*), the information about the square relative errors is provided separately for each pattern. In this case, for the log-ratio variances, the weight (in %) of each pattern is respectively 5.20, 61.15, and 33.65. For the log-ratio means, the percentages of contribution by pattern to the square relative errors are respectively 9.58, 68.93, and 21.49. One can conclude that the second pattern is the responsible of the major contribution to the relative errors in the variation arrays.

Table 3.2. Contribution (in %) of each pairwise logratio to the squared relative errors for the variation arrays of `foraminiferal` CoDa set.

	neogl_atl	neogl_pach	glob_obesa	glob_triloba
neogl_atl	–	56.29	13.99	20.83
neogl_pach	26.81	–	0.52	8.37
glob_obesa	24.23	47.08	–	0.00
glob_triloba	1.38	0.51	0.00	–

3.3. Count zeros

Activities for Section 3.3

The CoDa set `weibo_hotels.xls` aims at comparing the use of Weibo (Facebook equivalent in China, see Appendix) 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 of the 10 hotels are analyzed and coded regarding the count of posts featuring information on facilities, food, events, and promotions. Hotels were coded as large "L" or small "S" in the categorical variable `hotel_size`.

- Use the menu *File* ▷ *Import* ▷ *Import XLS Data* to load the `weibo_hotels.xls` file.
- Calculate the sum of the four variables `facilities`, `food`, `events`, and `promotions` using the menu *Data* ▷ *Manipulate* ▷ *Calculate new Variable*. Denominate `sum` the new variable.
- We want to analyze the variable `sum` by the groups defined by the variable `hotel_size`. Go to the menu *Statistics* ▷ *Classical statistics summary* to calculate the percentiles 0, 25, 50, 75, and 100 of the variable `sum` by the categories in `hotel_size`. Do you see any differences between both groups?
- Use the menu *Data* ▷ *Operations* ▷ *Subcompositions/Closure* to close the data to 100%. Do you see any relative differences between the groups defined by `hotel_size`?

The percentiles of the variable `sum` suggest some differences between the groups. For example, the percentile 0 (= *minimum*) for the group of large hotels ("L") is

49 whereas the the percentile 100 (= *maximum*) for the group of small hotels ("S") is 44. Consequently, to compare the values in the four variables it is recommended to apply the closure operation. Once this operation is applied CoDaPack creates the four variables `clo_facilities`, `clo_food`, `clo_events`, and `clo_promotions`. Table 3.3 shows the percentiles of the four variables by the two categories defined by the hotel size. The results suggest that large hotels take smallest values in the part `facilities`. On the other hand, for the variables `food` and `events` the smallest values are in the group "S". In the part `promotions` the differences are not so relevant.

Table 3.3. Percentiles of variables `facilities`, `food`, `events`, and `promotions` by the categories in `hotel_size` of the `weibo_hotels` data set closed at 100%.

Part	Category	0	25	50	75	100
`facilities`	L	3.64	12.12	17.46	18.18	18.37
	S	53.85	55.56	59.52	63.33	63.64
`food`	L	28.57	33.33	34.55	36.73	43.64
	S	6.67	7.16	9.09	13.89	15.38
`events`	L	20.41	21.81	25.40	30.30	34.55
	S	0.00	0.00	2.56	3.33	5.56
`promotions`	L	18.18	24.24	24.49	25.45	28.57
	S	25.00	26.67	27.27	28.21	33.33

- Note that the `events` part has two zeros, which must be treated as count zeros with the *Bayesian Multiplicative* (*BM*) method. Replace the zeros by the count-zero replacement methods included in the *Irregular data* ▷ *Bayesian Multiplicative zero Replacement* menu. In the *Output* option the user can choose to close the replaced data as imputed proportions (*prop*, default) or leave them as pseudo-counts (*p-counts*). Several procedures are available under the *Method* option: Geometric Bayesian Multiplicative (GBM, default); square root BM (SQ); Bayes-Laplace BM (BL); and count zero multiplicative (CZM).
  - ○ Introduce the four parts [`facilities`, `food`, `events`, `promotions`] with the raw counts in the *Selected data* window.
  - ○ Select the *GBM* method.
  - ○ Select the *prop* output.
  - ○ Click the *Accept* button.
- Observe the values in the columns for the replaced parts [`facilities.imp`, `food.imp`, `events.imp`, `promotions.imp`]. Do you think that the replacement method have replaced the zeros by an *adequate* value?
- Plot the quaternary diagram of the replaced parts [`facilities.imp`, `food.imp`, `events.imp`, `promotions.imp`] (Fig. 3.6). Select `hotel_size` in *Groups*. What are the main differences between large and small hotels?

The two zeros have been replaced by 0.02. Once compared to the other values in the part `events` for the small hotels one may conclude that 0.02 is a *reasonable* value. Figure 3.6 shows that the two groups of samples are separated. As commented above, the small hotels (red circles) are close to the vertex `facilities`, whereas they take smaller values in the parts `food` and `events` than the large hotels (blue circles).

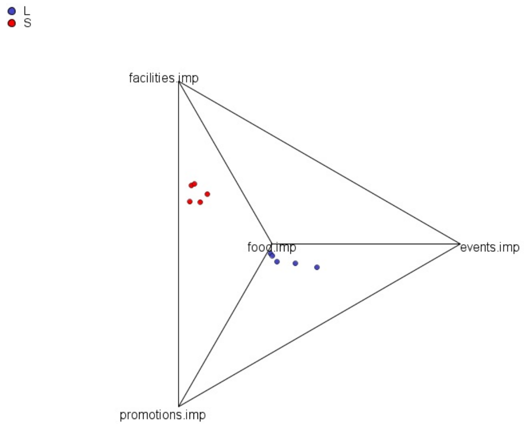


Figure 3.6. Quaternary diagram after GBM count zero replacement in `weibo_hotels` CoDa set. Blue circles are large hotels (L), red circles are small hotels (S)

3.4. Censored data: rounded zeros

Activities for Section 3.4

A. **Rounded zero replacement**

We load the CoDa set `foraminiferal` (see Appendix). In this section, we assume that zeros in the data set are rounded zeros. The data set has four parts `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`, and the 5 rounded zeros are recorded as 0 in parts `glob_obesa` and `glob_triloba` (samples 7, 17, 21, 25 and 30).

Because rounded zeros are censored data then one has to define the detection limit or threshold. This information is used in the replacement procedure for imputing a value below the detection limit.

- Go to the *Irregular data* ▹ *Set detection limit* menu.
  - ○ Select the variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`.
  - ○ Select *Take Minimum of Each Column* into the *Option* window.
  - ○ Click the *Accept* button.

By coincidence, both parts —`glob_obesa` and `glob_triloba`— with zeros have a minimum non-zero value equal to 0.01 which becomes the detection limit. CoDaPack shows zeros with their detection limit as 0[0.01] in the data frame. The menu *Irregular data* ▹ *Set detection limit* could have been used by selecting one or more parts and providing a *Detection limit* defined by the user for those particular parts.

The zero-patterns plot (Fig. 3.5) shows that the overall percentage of cells with zero values is 4.17%. When the percentage of zeros is small, the non-parametric multiplicative replacement method introduces minor distortion in the covariance structure.

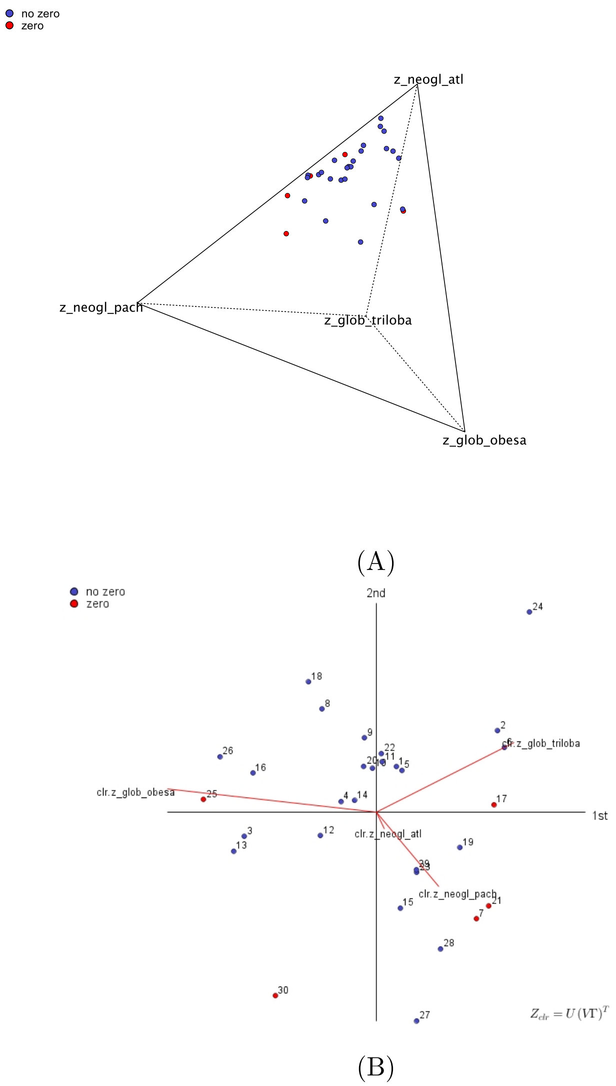
- To replace the zeros for a fraction (65%) of the detection limit (0.01) by using the the non-parametric multiplicative replacement method go to the menu *Irregular data* ▹ *Non-parametric zero Replacement.*
  - ○ Select the variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`
  - ○ Write 0.65 into the box *DL proportion* of the *Options* window.
  - ○ Click the *Accept* button.

CoDaPack creates four new variables `z_neogl_atl`, `z_neogl_pach`, `z_glob_obesa` and `z_glob_triloba` without zeros (replaced by 0.0065 before closure). Note that the area *Table Format* in the menu *File* ▹ *Configuration* allows to extend the number of decimals for the values in the data.

- Go to menu *Graphs* and plot the quaternary diagram (Fig. 3.7A) with the replaced variables `z_neogl_atl`, `z_neogl_pach`, `z_glob_obesa` and `z_glob_triloba`. Use variable `zeros` as grouping variable.
- Plot the covariance *clr*-biplot of the replaced variables (Fig. 3.7B). Use variable `zeros` as grouping variable.
  - ○ Click on *Data, Show observation names.*
- Interpret the biplot, especially with respect to the position of samples with replaced values (samples 7, 17, 21, 25 and 30).

As expected the imputed samples are far away from the center of the data set because they take small values in some parts. Samples 7, 17, and 21, with zeros in part *glob_obesa*, are far from the vertex in the quaternary diagram (Fig. 3.7A) and in the negative part of the ray *clr.z_glob_obesa* in the biplot (Fig. 3.7B). Samples 25 and 30 show an analogous behaviour as regards the part *glob_triloba*.

Now we will replace the zeros in the original data set *foraminiferal* applying the modified log-ratio EM algorithm.



(A)

(B)

Figure 3.7. `foraminiferal` CoDa set after the non-parametric multiplicative zero replacement: (A) Quaternary diagram; (B) Covariance biplot (94.52% variance accounted).

CoDaPack requires at least one part to have complete data for all samples. We know that the `foraminiferal` data set fulfills this condition. If this is not the case, a preliminary zero replacement can be performed with the simpler non-parametric replacement method. The part with the fewest replaced zeros is substituted as if it was the raw data and plays the role of the complete part.

- Go to the menu *Irregular data* ▷ *Logratio-EM zero Replacement.*
    - ○ Select the variables `neogl_atl`, `neogl_pach`, `glob_obesa`, `glob_triloba`

  - ○ Keep default options.
  - ○ Click the *Accept* button.
- Go to the *File* ▷ *Configuration* menu and write 0.0000 (four decimals) in the *Table format* cell.
  - ○ Click the *Accept* button.
- Look at the data frame. Compare the replaced values with both methods.
- Select the new replaced variables `z_neogl_atl.c`, `z_neogl_pach.c`, `z_glob_obesa.c` and `z_glob_triloba.c`, and plot their covariance *clr*-biplot (Fig. 3.8). Use the variable `zeros` as grouping variable.
  - ○ Click on *Data* ▷ *Show observation names.*

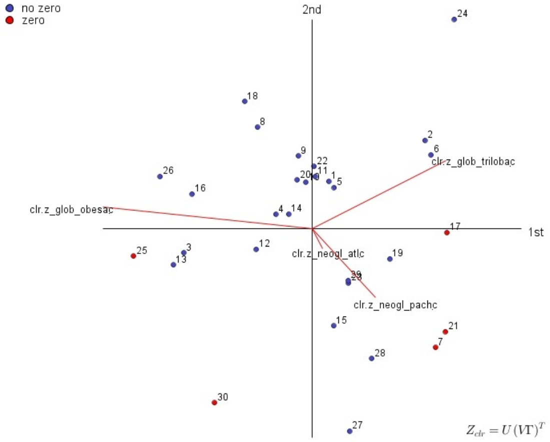


Figure 3.8. Covariance biplot of the `foraminiferal` CoDa set after the modified log-ratio EM zero replacement (94.86% variance accounted).

No large differences are detected between the biplot after the multiplicative replacement (Fig. 3.7B) and the biplot of the `foraminiferal` data set after the modified log-ratio EM zero replacement (Fig. 3.8). Both rays and samples show a similar behaviour.

B. Cell-specific detection limits

Using the menu *Irregular data* ▷ *Set detection limit*, CoDaPack can only define the same detection limit for all zeros in a part. However, one can import `Excel` or text files in which a possibly different detection limit is specified in each cell with a rounded zero. For this purpose, rounded zeros are not coded as "0" but as the detection limit preceded with the symbol "<", for instance "< 0.008". The dot "." is always used as a decimal separator in detection limits, even if in some languages `Excel` uses commas "," as a decimal separator in the data.

- Open the original `Excel` data frame `foraminiferal.xls` with the `Excel` program.
  - ○ Code the zeros in samples 7, 17, 21, 25 and 30 as being below the detection limits 0.005, 0.008, 0.007, 0.009 and 0.01, respectively.
  - ○ Save the `Excel` file with a different file name.

- Import it into CoDaPack and check that detection limits are properly indicated as 0[0.005], 0[0.008] and so on.
- Go to the menu *Irregular data* ▷ *Non-parametric zero Replacement* to replace the zeros for a fraction (65%) of the detection limit by using the non-parametric multiplicative replacement method.

The procedure works correctly. The rounded zeros have been respectively replaced by the 65% of 0.005, 0.008, 0.007, 0.009 and 0.01.

C. **Amalgamation**

We shall construct a new composition from the raw components by amalgamating the parts `glob_obesa` and `glob_triloba`. No rounded zeros exist in this amalgamated three-part composition.

- Select the menu *Data* ▷ *Manipulate* ▷ *Calculate New Variable.*
  - ○ Introduce `glob_obesa` and `glob_triloba` as *Selected Data.*
  - ○ Click the *Accept* button.
  - ○ Write the expression *x1+x2* into the *Enter expression* cell of the *Create new Variable* window.
  - ○ Write `glob_obesa_triloba` into the *Enter new variable name* cell of the *Create new Variable* window.
  - ○ Click the *OK* button.
- Plot the covariance *clr*-biplot from the variables `neogl_atl`, `neogl_pach` and `glob_obesa_triloba` (Fig. 3.9).
  - ○ Click on *Data* ▷ *Show observation names.*
- Compare the results with those obtained with both replaced data sets.
- Why are samples 24 and 13 close together in the amalgamated biplot and distant in the replaced biplots?

The data set `foraminiferal` after the amalgamation is formed by 3-part compositions. Once this two dimensional data set is represented into a *clr*-biplot (Fig. 3.9), 100% of variance is accounted. The structure of variance of this data set is different from the replaced data set (e.g., Fig. 3.8). For example, the ray of *crl.neogl_atl* in Fig. 3.9 is larger than the ray in Fig. 3.8. Moreover, samples 13 and 24 are separated in Fig. 3.8. Sample 13 is in the negative part of the horizontal axis whereas the sample 24 is in the positive part. They take different values in the parts *glob_obesa* and *glob_triloba* and they take similar values in the other two parts. On the other hand, these two samples are very close together after amalgamation (Fig. 3.9) because the two samples take similar values in the three parts (`neogl_atl`, `neogl_pach`, `glob_obesa_triloba`).

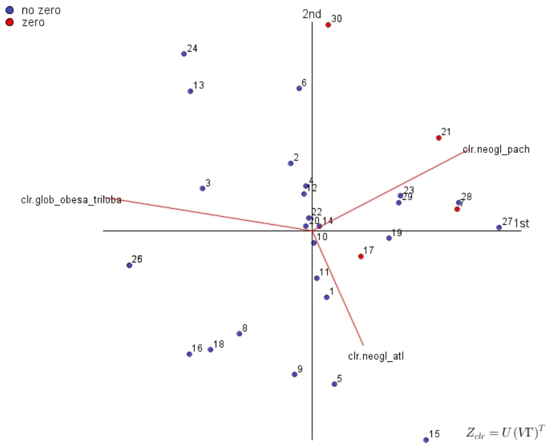


Figure 3.9. Covariance biplot of the CoDa set `foraminiferal` after the amalgamation of parts `glob_obesa` and `glob_triloba` (100% variance accounted).

3.5. Dealing with missing values and zeros

Activities for Section 3.5

We shall import the `Excel` file `foraminiferalNA.xlsx` (see Appendix) with rounded zeros that we already analyzed in previous sections (Fig. 3.5). We forced (randomly) MCAR in the data set to create a data set with zeros and NA.

- Go to the menu *Irregular data* ▷ *ZPatterns plot*. Select all the parts. Introduce NA as the label into the box *Label*.

Figure 3.10 shows the distribution of missing values among samples and parts. Note that 40% of samples have at least one NA and no part is free of missing values.

To deal simultaneously with rounded zeros and NA:

- Go to the menu *Irregular data* ▷ *Logratio EM Zero & missing replacement*.
  - Introduce the four parts as *Selected Data*.
  - Because all parts have NA, select *TRUE* at the *Rob Option* and the option *complete.obs* at the area *IniCov Option*.
  - Click the *Accept* button.
- Plot the covariance *clr*-biplot from the variables `neogl_atl_NA.imp`, `neogl_pach_NA.imp`, `glob_obesa_NA.imp` and `glob_triloba_NA.imp` (Fig. 3.11). Use variable `missing` as grouping variable.
  - Click on *Data* ▷ *Show observation names*.

As expected most of the samples with imputed rounded zeros (7, 17, 21, 25 and 30) are far from the center of the data set. Samples with imputed NA (red circles) are close to the center except the sample 24. For this sample, which has a large value in

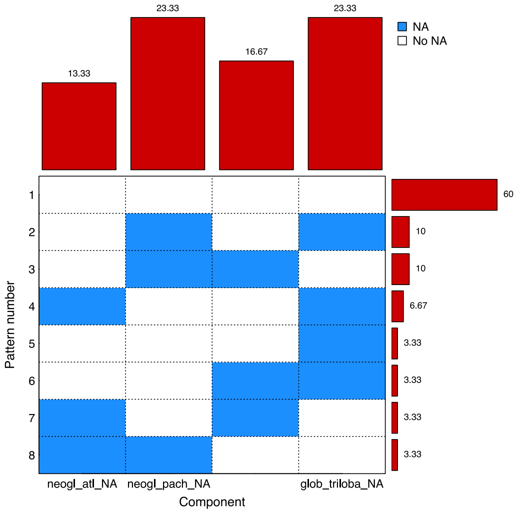


Figure 3.10. Patterns of forced missing values in the CoDa set `foraminiferal`.

part *glob_triloba*, the algorithm imputed a very small value ($3 \cdot 10^{-5}$) for the missing value in the part *glob_obesa*. Angles and lengths of rays of *clr*-variables (Fig. 3.11) are very similar to the biplot after zero replacement (Fig. 3.8), suggesting that the algorithm has introduced a minor distortion on the covariance structure.

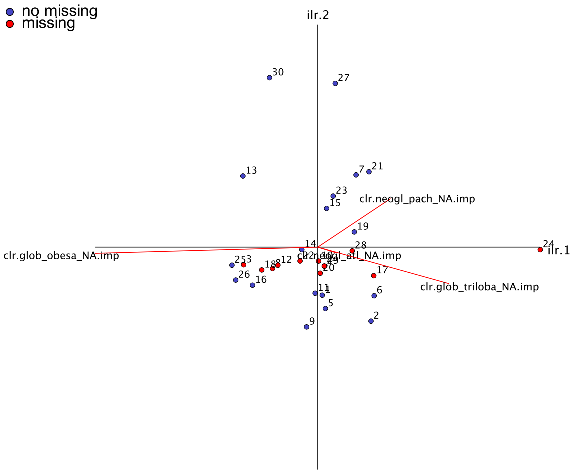


Figure 3.11. Covariance biplot of the CoDa set `foraminiferal` after the replacement of zeros and NA (97.44% variance accounted).

3.6. Potential outliers

Activities for Section 3.6

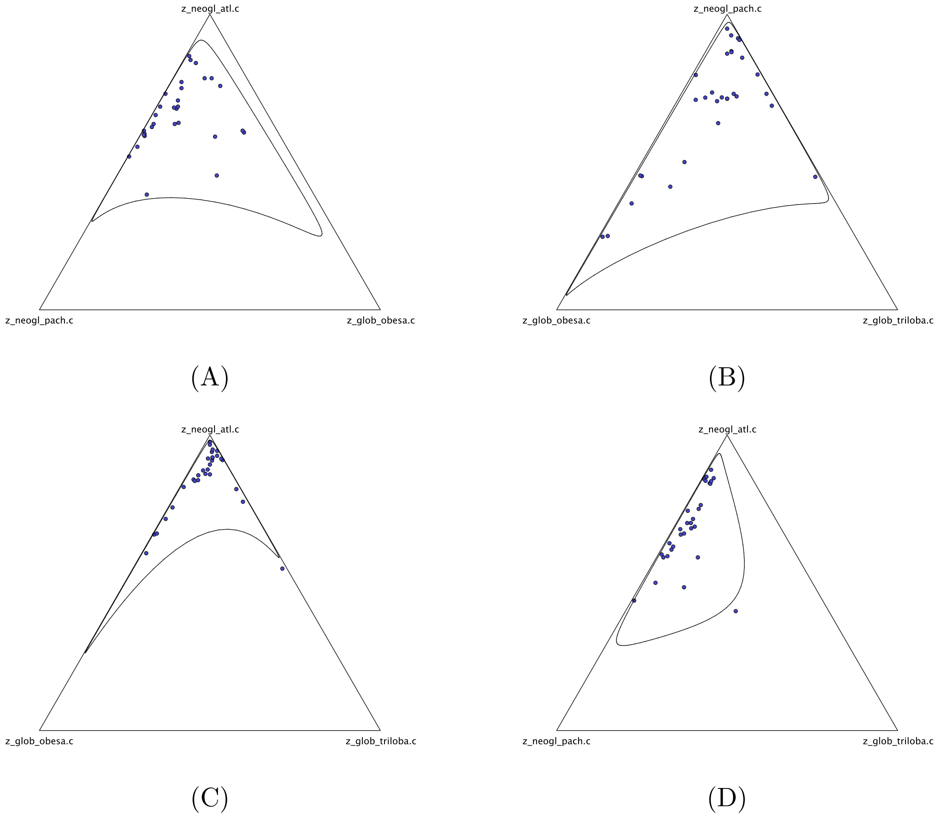
We use the CoDa set `foraminiferal` (see Appendix) after zero replacement with the modified log-ratio EM algorithm.

Assuming log-ratio normality, the menu *Irregular data* ▷ *Atipicality index* provides the χ^2 quantile of the squared Mahalanobis distances. The *Level of confidence* value represents the threshold that decides if a sample is a potential outlier. CoDaPack shows which samples have an index above the threshold and labels them as *Atypical* in the variable `Atyp`. The `Chisq` variable contains the quantiles themselves.

- In the *clr*-biplot (Fig. 3.8) of `foraminiferal` CoDa set (after zero replacement with the modified log-ratio EM algorithm), is there any sign of potential outliers?
- Go to *Irregular data* ▷ *Atipicality index* menu.
  - Select the variables `z_neogl_atl.c`, `z_neogl_pach.c`, `z_glob_obesa.c` and `z_glob_triloba.c`.
  - Write *0.975* into the *Level of Confidence* cell of the *Options* window.
  - Click the *Accept* button.
- Is there any potential outlier in this data set?
- Repeat the *clr*-biplot (Fig. 3.8) using the variable *Atyp* as a factor.
  - Click on *Data, Show observation names.*

Only the sample 24 is classified as potential outlier.

- Why is it extreme? Look at the value of this sample and compare them with the percentiles of the four parts.
- Go to the *Graphs* ▷ *Predictive Region* menu to draw the predictive regions of selected 3-part subcompositions.
  - Select three of the four variables `z_neogl_atl.c`, `z_neogl_pach.c`, `z_glob_obesa.c` and `z_glob_triloba.c`.
  - Write *0.975* into the *Predictive level* cell of the *Options* window.
  - Click the *Accept* button.
  - Repeat the diagram with a three different selected variables until you create the four possible ternary diagrams (Fig. 3.12).



(A) (B)

(C) (D)

Figure 3.12. Ternary diagrams with the four 3-part subcomposition of **foraminiferal** data set after the modified log-ratio EM zero replacement: 97.5% predictive region. (A)(**neogl_atl**, **neogl_pach**, **glob_obesa**); (B) (**neogl_pach**, **glob_obesa**, **glob_triloba**); (C) (**neogl_atl**, **glob_obesa**, **glob_triloba**); (D) (**neogl_atl**, **neogl_pach**, **glob_triloba**)

The sample 24 has the percentage composition (neogl_atl, neogl_pach, glob_obesa, glob_triloba)= (40, 27, 1, 32). The minimum value in the part neogl_atl is 38.74%, where the sample 40 takes the second smallest value. The sample takes the maximum value in the part glob_triloba (32%), whereas the 75% percentile is 4%. The values in the other two parts are in the middle of the distribution. Consequently, the pairwise ratio $\frac{\texttt{glob\_triloba}}{\texttt{neogl\_atl}}$ takes an extreme value in sample 40 as regards the other compositions in the data set. Figures 3.12A–D show the four possible 3-part subcompositons in the ternary diagrams. In Fig. 3.12A and B, the sample 40 is not classified as outlier because one of the parts glob_triloba and neogl_atl is not represented. On the other hand, when the two parts glob_triloba and neogl_atl) are represented (Fig. 3.12C and D) the sample 40 is outside the predictive region.

The chapter's key concepts

- ✓ There are appropriate methods for dealing with irregular CoDa: missing data, outliers and zeros.
- ✓ Different types of zeros require different techniques for dealing with them.

Chapter 4

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

4.1. LRM for a compositional response and scalar predictor

Activities for Section 4.1

The principle of working on coordinates allows researchers to apply any multivariate statistical method (regression, ANOVA, generalized linear models, discriminant analysis, clustering methods, time series, geostatistics, etc.) available in any statistical software. The most common statistical tools are implemented in CoDaPack so that exporting coordinates to an external statistical platform is not needed. To use a statistical method in an external package you can proceed in a simple way: compute *olr*-coordinates of data in CoDaPack, export them to your statistical platform, get results, interpret them on coordinates, bring results back as compositions and interpret them in the simplex.

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 3-part compositions [`sand`, `silt`, `clay`]. The file *arcticlake.cdp* (see Appendix) records the compositions of 39 sediment samples at different water depths in an Arctic lake. Samples are ordered by the values in the variable water depth. Here we want to understand how the sediment composition can be explained by the water depth. For that reason we are going to create a LRM in which the explanatory variable (X) is the water depth and the response (Y) is the 3-part composition:

- Load the file `arctic_lake.cdp`.
- Plot the CoDa set in the ternary diagram.
  - ○ Use the variables `num_sedim` or `depth` as observation names to investigate if there is some association between composition and depth (*Data* ▷ *Add observation names* and *Data* ▷ *Show observation names*).

  Figure 4.1 shows that concentration of sand decreases as depth increases, whereas clay concentration increases.
- Calculate the *olr*-coordinates of the compositional system sand-silt-clay according the default SBP of CoDaPack: `ilr.1` and `ilr.2`.
- Inspect the two scatterplots of each *olr*-coordinate (`ilr.1` and `ilr.2`) against `depth` through the menu *Graphs* ▷ *Scatterplots 2D/3D*.
- Calculate the Pearson correlation between `ilr.1` and `depth` through the menu *Statistics* ▷ *Classical statistics summary*. (Result: $r = -0.7752$)

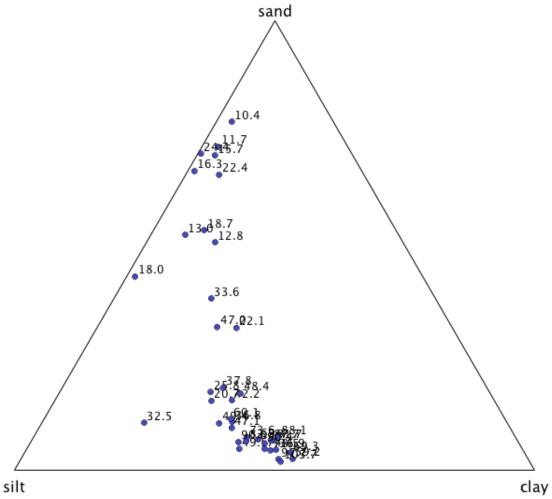


Figure 4.1. `Arctic_lake` dataset in the ternary diagram. Labels of samples are the value of variable water depth

- Calculate the Pearson correlation between `ilr.2` and `depth` through the menu *Statistics* ▷ *Classical statistics summary*.(Result: $r = -0.8223$)
  Both scatterplots and Pearson correlation coefficients suggest that the variable `depth` has a reasonable *linear* correlation with `ilr.1` and `ilr.2`.
- Repeat the previous steps (scatterplot and Pearson correlation) using the logarithm of `depth` (`ln(depth)`) instead of `depth`. To create the new variable `logdepth`:
  - go to the menu *Data* ▷ *Manipulate* ▷ *Calculate New Variable*,
  - select the variable `depth` and click the button *Accept*,
  - write the expression *log(x1)* into the box *Enter expression* of the window *Create new Variable* and write *logdepth* into the box *Enter new variable name*, and
  - click the button *Ok*.

Figure 4.2 shows the scatterplots of each *olr*-coordinate `ilr.1` and `ilr.2` against the variable `logdepth`. Both scatterplots and Pearson correlation coefficients indicate that the variable `depth` log-transformed is preferable as a predictor in the LRM.

We estimate the two LRM between the *olr*-coordinates and `logdepth`:

- Go to the menu *Statistics* ▷ *Multivariate Analysis* ▷ *Regression* ▷ *X real Y composition*.
  - ○ Choose `logdepth` on the box *Selected X*. Introduce the three components `sand`, `silt` and `clay` (in this order) into the box *Selected Y*.
  - ○ Activate the options *Residuals* and the *Fitted*.
  - ○ Click the button *Set Y partition*, and select the option *Default partition*.
  - ○ Click the button *Accept*.

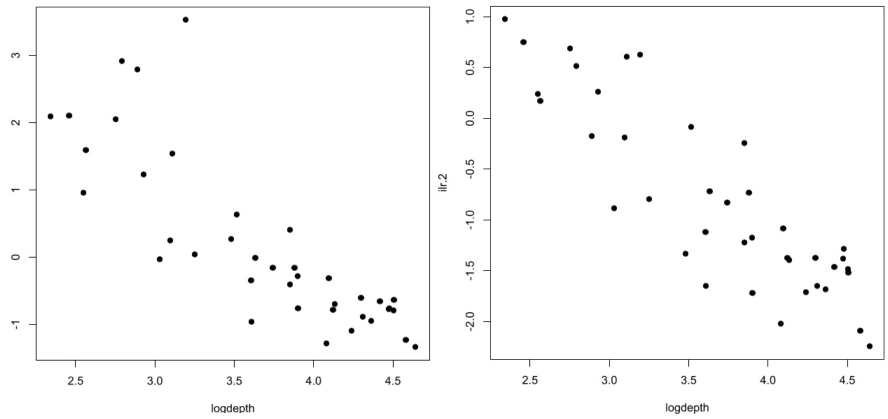


Figure 4.2. Scatterplots of the two *olr*-coordinates against `logdepth` in the `Arctic_lake` data set: (left) with `ilr.1`, where $r = -0.8314$; (right) with `ilr.2`, where $r = -0.8755$.

- ○ Click the button *Accept* in the window *X real Y composition regression Menu.*

In the window *Output* of CoDaPack we can see the estimates of the two regression models:

$$ilr.1 = 5.9206 - 1.5673 * logdepth$$
$$ilr.2 = 3.4593 - 1.1644 * logdepth$$

Note that CoDaPack creates a new table (*coefficients*) to store the coefficients of the two regression models (columns *ilr.1* and *ilr.2*). Also, in the window *Output*, we can see the typical information of any regression model: estimates (values, standard errors, *p*-values), residual standard errors, the R^2 of each model . . . At the end of the output CoDaPack provide the overal R^2 value (Result: $r^2 = 71.62748$).

- Are all the coefficients of the regression model significant?

  Yes, coefficients can be considered significant because all *p*-values are below 5%.

- How is the goodness of fit of the model?

  The quality can be considered quite reasonable because the overall R^2 takes the value 71.63%, that is, the variable ln`(depth)` explains 71.63% of the variation of composition [`sand`, `silt`, `clay`].

Moreover, two graphic windows (*ilr.1* vs *logdepth* and *ilr.2* vs *logdepth*) allow to graphically assess the goodness of the fit of the LRM, and the normality of the residuals. Both homoscedasticity and normality are reasonable but few samples can be considered as potential outliers.

Now we will back-transform the coefficients of the regression model to the simplex.

- Make sure you have the *coefficients* table selected.
- Go to menu *Data* ▷ *Transformations* ▷ *ILR-Raw.*
  - ○ Select variables `ilr.1` and `ilr.2` .

- ○ Select *Default Partition* in the window *Options*.
- ○ Click the button *Accept*.

The new variables `inv.ilr.1`, `inv.ilr.2` and `inv.ilr.3` store the coefficients of the LRM expressed in compositional form:

$$\texttt{[sand, silt, clay]} = [0.9925, 0.0074, 0.0001] \oplus \\ \oplus \texttt{logdepth} \odot [0.0460, 0.2389, 0.7151]$$

Note that a closure to 1 has been applied to the coefficients.

- Is the expression of the model coherent with the pattern observed in the ternary diagram?

  Once one compares the gradient $[0.0460, 0.2389, 0.7151]$ to the uniform perturbation $[1/3, 1/3, 1/3]$ then we expect that the concentration of sand decreases and an increase of clay proportion when depth increases. Because the intercept is $[0.9925, 0.0074, 0.0001]$ one can interpret that the regression *straight* line departs from vertex `sand`. The line arrives to vertex `clay` because the third component of the gradient takes the largest value (0.7151).

Now we will calculate the fitted values and the residuals of the LRM:

- Select the *articlake* table.

Columns `ilr.1.r` and `ilr.2.r` store the residuals, and `ilr.1.f` and `ilr.2.f` store the fitted values. Take into account that the residuals and the fitted values are expressed in *olr*-coordinates.

- Use the menu *Data* ▷ *Transformations* ▷ *ILR-Raw* to back-transform the fitted values using the default partition.
- Draw the compositional fitted samples (variables `inv.ilr.1`, `inv.ilr.2` and `inv.ilr.3`) in a ternary diagram and compare them with the original values in the data set. You can use the `depth` variable as observation names (Fig. 4.3).

  The shape of the cloud in Fig. 4.3 represents quite well the shape of the original data set but few samples that can be considered as potential outliers (Fig. 4.1).
- Back-transform the residuals using the default partition.
- Draw the compositional residuals `inv.ilr.1.c`, `inv.ilr.2.c` and `inv.ilr.3.c` in the ternary diagram.
  - ○ Activate the options *Centered* and *Show Center* on the border of the graph.
  - ○ What do you observe? What is the justification?

    Once centered the residuals data set remains unaltered because the center of the original residuals is already the center of the simplex.
- Go to the menu *Statistics* ▷ *Log-Ratio Normality Test* to test the log-ratio normality of the back-transformed residuals `inv.ilr.1.c`, `inv.ilr.2.c` and `inv.ilr.3.c`.
  - ○ Do you reject the multivariate log-normality hypothesis of residuals?

    As regards radius test, we see that p-value of Anderson-Darling test is below 5% but p-values of both Cramér-von Mises and Watson test are greater than 5%. This disagreement could be caused by the potential

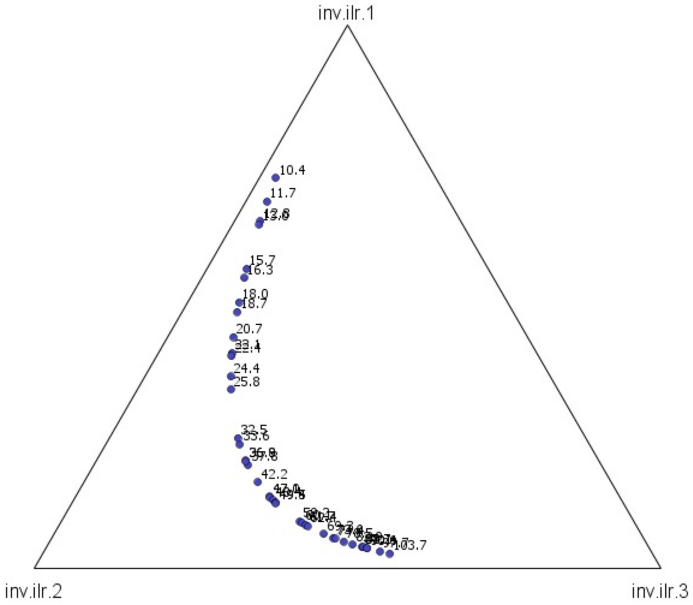


Figure 4.3. Ternary diagram with fitted values (`arcticlake` data set)

> outliers. In any case, it is a decision of the analyst to reject or not the multivariate normality of the residuals. Once we analyze the univariate normality of *olr*-coordinates then we detect that the variable `ilr.1` has a lack of normality because the *p*-values of the three tests are below 2.5%.

Does the estimated regression model depend on the basis used to calculate the *olr*-coordinates of compositions? Let's see how not.

- Load again the original *arcticlake.cdp* data set.
- Go to menu *Statistics* ▷ *Multivariate Analysis* ▷ *Regression* ▷ *X real Y composition* use a partition different from the default to estimate the LRM [Note: we suggest the partition: $1\ -1\ 1||\ 1\ 0\ -1$].
- Is the global R^2 value different from the one obtained previously?
- Back-transform the LRM coefficients.
- Does the compositional LRM change?

The overall R^2 takes the same value (71.63%) and the coefficients of the LRM model, once back-transformed, are the same. The LRM model expressed in the simplex doesn't change, it is invariant under a change of the *olr*-basis.

4.2. LRM for a scalar response and compositional predictor

Activities for Section 4.2

The *BMIPhisActi.cdp* CoDa set (see Appendix) 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. Remember that the body mass index (BMI) is a measure of body fat based on height and weight. It is useful for detecting overweight and obesity. The higher your BMI, the higher your risk for certain diseases (heart disease, high blood pressure, type 2 diabetes...). In the file `BMIPhisActi.cdp`, the variable BMI has been standardized as a z-score (`zBMI`) using age- and sex-specific reference data provided by the World Health Organisation (WHO). So, `zBMI` can theoretically take any value from $-\infty$ to $+\infty$. Finally, a categorical variable `gender` identifies the gender (boy or girl) of each child.

We intent to create from the *BMIPhisActi.cdp* data set an LRM to predict `zBMI` from the composition [`sleep`, `sedent`, `Lpa`, `Mpa`, `Vpa`]. We also want to analyze how changes in physical activity influence `zBMI`.

- Load the file *BMIPhisActi.cdp*.
- Analyze the classical statistics summary of `zBMI`.
  We see that the mean (0.5866) is greater than the median (0.4900) suggesting skewness at the right side. However, the standard deviation (1.1314) is quite large as regards the mean, suggesting a large variability. Consequently, the difference between mean and median is not large.
- Go to menu *Graphs* ▷ *Boxplot* to plot de boxplot of of `zBMI`. Select *Draw mean: Arithmetic* under *Options*.
  - ○ Globally, do the children in this sample have a balanced weight?

  The boxplot corroborates the numerical statistics. No relevant skewness is detected.
- Perform a Shapiro-Wilk normality test of `zBMI` by means of the menu *Statistics* ▷ *Classical Univariate Normality test*.
  The p-value (0.3474) indicates that we fail to reject the normality of the data.
- Analyze the compositional statistics of the composition [`sleep`, `sedent`, `Lpa`, `Mpa`, `Vpa`].
  - ○ Which parts dominate the geometric center?
  - ○ Which parts dominate the total variance?

  `sleep` (40.09%) and `sedent` (33.78%) dominate the proportion of time use. Note that the uniform center is $[1/5, 1/5, \ldots, 1/5]$, that is, it is doubled by the proportion of `sleep`. As regards variability, `Vpa` (0.1788) largely dominates the weight in the total variance (0.3389).
- Draw the biplot of the full composition.
  - ○ Does the first axis explain a large percent of variance?
  - ○ How are the 5-parts separated according to the positive and negative sides of the two axes?

The biplot has a high quality because it accounts for 93.40% of variance. The first axis retains 81.15% of variance, whereas the second only 12.25%. The rays of parts `sleep`, `sedent` and `Lpa` are in the positive side of first axis, whereas `Mpa` and `Vpa` are in the negative part. As regards the second axis, the rays of `sleep`, `sedent` and `Vpa` are in the negative part and the other parts (`Lpa` and `Mpa`) are in the positive part (Fig. 4.4). No relevant differences between boys (blue dots) and girls (red dots) are detected.

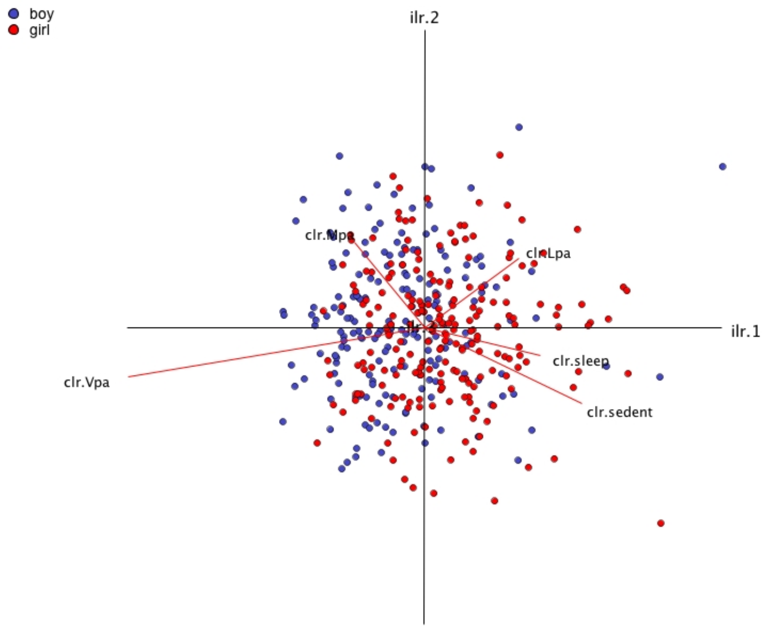


Figure 4.4. clr-biplot of `BMIPhisActi` data set. Red dots are samples of girls and blue dots are boys.

- Do the CoDa-dendrogram and calculate the corresponding *olr*-coordinates using the following SBP in *Defined partition* (select variables `sleep`, `sedent`, `Lpa`, `Mpa` and `Vpa` in this order):

$\mathrm{olr}_{\mathrm{k}}$	`sleep`	`sedent`	`Lpa`	`Mpa`	`Vpa`
x_1^*	$+1$	-1	$+1$	$+1$	$+1$
x_2^*	-1	0	$+1$	$+1$	$+1$
x_3^*	0	0	-1	$+1$	$+1$
x_4^*	0	0	0	-1	$+1$

  ◦ Which balance dominates the variability?

The variance of coordinate $\sqrt{\frac{2}{3}}\ln\left(\frac{(\mathtt{Mpa}\cdot\mathtt{Vpa})^{1/2}}{\mathtt{Lpa}}\right)$ is the largest (0.1116), dominating the total variance (0.3389).

- Go to the menu *Statistics* ▷ *Classical statistics summary* and find the correlation coefficient of each *olr*-coordinate against the `zBMI` variable in the correlation matrix.

- Go to the menu *Graphs* ▷ *Scatterplot 2D/3D* and enter the `zBMI` variable and one of the *olr*-coordinates at a time.
  - Assess the level and type of correlation of `zBMI` with each *olr*-coordinate.

  The correlations are respectively -0.19, -0.15, -0.17 and -0.24. All of them are negative but none approaches -1. The scatterplots corroborate this behaviour, that is, no relevant correlation is suggested because none shows a linear trend.

Let's estimate the LRM to explain the variability of `zBMI` from the 5-part composition [`sleep`,`sedent`,`Lpa`,`Mpa`,`Vpa`]:

- Go to the menu *Statistics* ▷ *Multivariate Analysis* ▷ *Regression* ▷ *X composition Y real.*
  - Enter the variables `sleep`, `sedent`, `Lpa`, `Mpa` and `Vpa` into the box *Selected X.*
  - Enter the `zBMI` variable into the box *Selected Y.*
  - Click the options *Residuals* and *Fitted.*
  - Set the same partition as before into the box *Set X partition.*
  - Click the button *Accept.*

Note that CoDaPack stores the regression coefficients in a new table called *coefficients.*

The *Output* window provides the standard information of a LRM. The equation of the model is:

$$\texttt{zBMI} = 0.7101 - 1.0662 * ilr.1 + 0.6337 * ilr.2 + 0.5111 * ilr.3 - 1.2928 * ilr.4 \quad ,$$

that is,

$$\begin{aligned} \texttt{zBMI} = 0.7101 - 1.0662\sqrt{\frac{4}{5}}\ln\left(\frac{(\texttt{sleep}\cdot\texttt{Lpa}\cdot\texttt{Mpa}\cdot\texttt{Vpa})^{1/4}}{\texttt{sedent}}\right) + \\ +0.6337\sqrt{\frac{3}{4}}\ln\left(\frac{(\texttt{Lpa}\cdot\texttt{Mpa}\cdot\texttt{Vpa})^{1/3}}{\texttt{sleep}}\right) + \\ +0.5111\sqrt{\frac{2}{3}}\ln\left(\frac{(\texttt{Mpa}\cdot\texttt{Vpa})^{1/2}}{\texttt{Lpa}}\right) - \\ -1.2928\sqrt{\frac{1}{2}}\ln\left(\frac{\texttt{Vpa}}{\texttt{Mpa}}\right). \end{aligned} \tag{4.1}$$

Observe that only the coefficients of coordinates `ilr.1` and `ilr.4` are significantly different from 0. The negative sign of them matches with the sign of the corresponding correlation coefficient seen above. The same cannot be said in relation to the coefficients of `ilr.2` and `ilr.3`. The goodness of fit is very low (adjusted $\mathrm{R}^2 = 0.067$) but the model is significant (p-value= $2.894e^{-06}$). The graphs of residuals show that the theoretical hypotheses of an LRM are fulfilled, at least approximately (Fig. 4.5).

- Compute the compositional gradient using the new coefficient table.
  - Make sure you have the table *coefficients* selected.

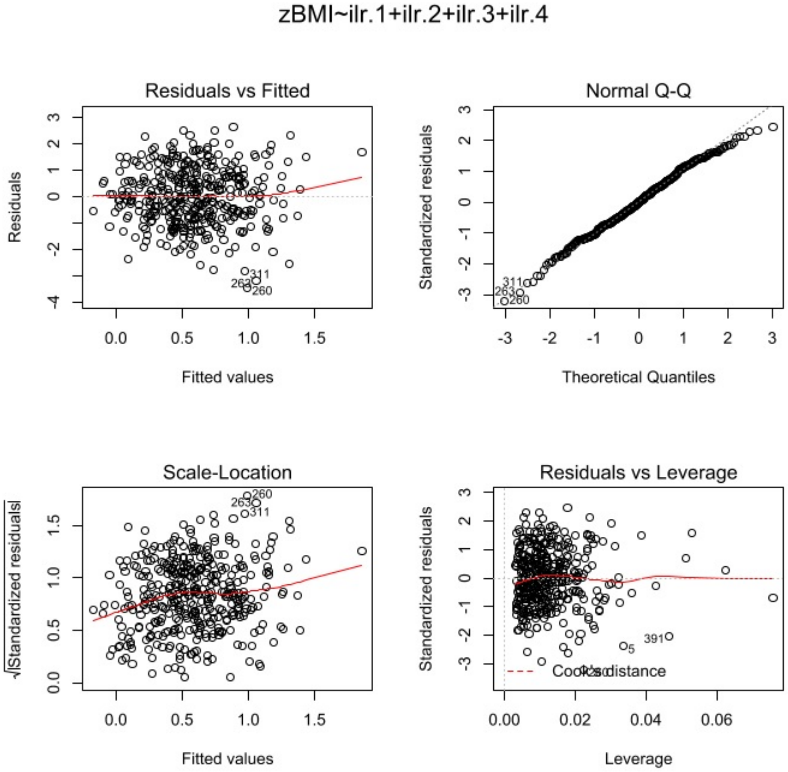


Figure 4.5. LRM diagnostics plots for `BMIPhisActi` data set

- ○ Go to the menu *Data* ▷ *Transformations* ▷ *ILR-Raw* and select the four *olr*-coordinates. The SBP must be the same than the partition used for creating the LRM.
- ○ Click the button *Accept*.

The gradient for the composition the composition [`sleep`, `sedent`, `Lpa`, `Mpa`, `Vpa`] is [0.0646, 0.3682, 0.0884, 0.4125, 0.0663] thus showing increases in `zBMI` to be associated mainly to reductions in `sleep` and to do more `Mpa`.

Besides interpreting the gradient, in medical research these LRM are used to study the differences in `zBMI` for reallocations of time between sleep, physical activity and sedentary behaviour. Thus, we will use this example to estimate how much does `zBMI` change if we reduce sedentary time by 30 minutes and increase time spent on `Lpa`, `Mpa` and the `Vpa` by 10 minutes each (without changing the sleeping time).

Due to the compositional nature of time-use data, the estimations for changes in `zBMI` related to reallocations between components must be made with reference to a baseline or starting compositions. We will use the mean time-use composition as the baseline, that is, the center:

$$[\texttt{sleep}, \texttt{sedent}, \texttt{Lpa}, \texttt{Mpa}, \texttt{Vpa}] = [0.4009, 0.3378, 0.2182, 0.0291, 0.0140].$$

The LRM (Eq. 4.1) predicts zBMI to be 0.587 for this baseline composition. If we reduce the daily sedentary time by 30 minutes (30/1440 in proportion) and increase the daily time spent on Lpa, Mpa and the Vpa by 10 minutes each (10/1440 in proportion), we are moving from the baseline composition to the following time-use composition:

$$[\texttt{sleep}, \texttt{sedent}, \texttt{Lpa}, \texttt{Mpa}, \texttt{Vpa}] = [0.4009, 0.3170, 0.2251, 0.0360, 0.0209].$$

The zBMI estimated from the LRM (Eq. 4.1) is now equal to 0.434, lower than that of the baseline composition. This reduction seems logical since the increase in time spent on physical activity helps to reduce weight.

Let's see if this reduction depends on the gender.

- Select the original table *BMIPhisActiv*.
- Go to the menu *Data* ▷ *Filters* ▷ *Categorical filter*.
  - Select the variable gender.
  - Click the button *Accept*.
  - Select *boy* and click the button *Accept*.

A new table *MMIPhisActiv_Fil_gender* –including only the boys will appear in the data frame.

- Go to the menu *Statistics* ▷ *Compositional statistics summary* to calculate the compositional center of [sleep,sedent,Lpa,Mpa,Vpa] in boys.

The new baseline composition is now:

$$[\texttt{sleep}, \texttt{sedent}, \texttt{Lpa}, \texttt{Mpa}, \texttt{Vpa}] = [0.3969, 0.3330, 0.2202, 0.0329, 0.0170].$$

The prediction for zBMI in Eq. (4.1) is now equal to 0.561. If as before we reduce the daily sedentary time by 30 minutes and increase the daily time spent on Lpa, Mpa and the Vpa by 10 minutes each the predicted zBMI is equal to 0.482.

We repeat the calculations for girls.

- Delete the table *MMIPhisActiv_Fil_gender*.
- Select the girls in the original *MMIPhisActi*.

The new baseline composition is now:

$$[\texttt{sleep}, \texttt{sedent}, \texttt{Lpa}, \texttt{Mpa}, \texttt{Vpa}] = [0.4040, 0.3417, 0.2162, 0.0261, 0.0118].$$

For girls the reallocation of 30 minutes from sedentary behaviour to physical activity reduces zBMI from 0.611 to 0.435. So it seems that the 30-minute time-reallocation is more effective in girls than in boys. This is due to girls having a different center.

4.3. LRM extensions

Activities for Section 4.3

We use the CoDa set *BMIPhisActi* (see Appendix) including the *olr*-coordinates analysed in the previous section. We want to further check the importance of gender from another point of view by adding it as a non-compositional explanatory variable into the LRM. For this purpose we save the table `BMIPhisActiv` including the *olr*-coordinates as an `R` workspace.

- Go to the menu *File* ▷ *Export* ▷ *Export Data to R Data*.
- Select all variables.
- Click the button *Accept*.
- Open the data file with R and run the following commands:

```
attach(data)
model1<-lm(zBMI∼ilr.1+ilr.2+ilr.3+ilr.4+factor(gender),data=data)
summary(model1)
plot(model1)
```

The coefficient of gender is non significant, having a p-value 0.410901. Thus, as shown before, boys and girls have a different center composition, but, now we see that conditional on the composition there is no evidence than boys and girls have a different `zBMI` value. In any case, the equation is:

$$\texttt{zBMI} = 0.59188 - 1.03512 * ilr.1 + 0.56882 * ilr.2 +$$

$$0.45223 * ilr.3 - 1.24335 * ilr.4 - 0.09951 * girl.$$

The `girl` dummy variable is 1 for girls and 0 for boys. Therefore, the intercept term for boys is 0.59188 and for girls $0.59188 - 0.09951 = 0.49237$. Girls thus have a lower `zBMI` value than boys with the same time-use composition, albeit not significantly so (p-value=0.4109).

The chapter's key concepts

✓ The principle of working on coordinates applied to linear models allows us to fit compositional regression models by ordinary least squares.

✓ Coefficients can be interpreted in coordinates or back to the simplex.

✓ The composition may be the response or the predictor variable.

✓ The LRM can be extended to other general models.

Chapter 5

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.

5.1. Cluster analysis

Activities for Section 5.1

We will again use the `alimentation.cdp` file already described in the activity of Chapter 2 for the normal distribution (see Appendix). The variables of the CoDa set are: `RM` (red meat), `WM` (white meat), `E` (eggs), `M` (milk), `F` (fish), `C` (cereals), `S` (starch), `N` (nuts), `FV` (fruits and vegetables).

The file also contains some categorical variables: `Country` (name of the country), `CountryID` (two letters identifying the country), `NorthMed` (North: Northern country; Med: Mediterranean or South European country) and `EastWest` (East: Eastern country; West: Western country).

We use CoDaPack to analyse the similarities between countries as regards to their consumption of foodstuffs and look for associations with the categorical variables.

- Draw the *clr*-biplot:
  - Use the variable `NorthMed` in the options *Groups*.
  - Once the window *Biplot* is open, label the observations with the names fo the countries (menu *Data* ▷ *Add observation names* and *Data* ▷ *Show observation names*).
  - Repeat the *clr*-biplot using the categories of the variable `EastWest` for the *Groups*.

The quality of the *clr*-biplot is quite high because it retains 77.12% of total variance. Note that the consumption of foodstuffs of the countries seems to be grouped by the variable `NorthMed`, whereas it appears mixed when the variable `EastWest` is used.

In CoDaPack we can cluster CoDa using the k-means algorithm, which is a simple method to classify numerical data in different numbers of groups. We can fix the number of clusters or find the best clustering according to the Calinski index or to the Average Silhouette index. Twenty-five random starts of the k-means algorithm are selected in each run, in order to prevent local optima.

- Go to the menu *Statistics* ▷ *Multivariate Analysis* ▷ *Cluster* ▷ *K-means*.
  - Select all nine parts in our CoDa set.
  - Choose the *Find the optimal number between 2 and* and write *6* into the box next to 2 in the *Options* menu.
  - Confirm the Calinski index in the *Optimal method* box.
  - Click the *Accept* button.

CoDaPack plots two graphics, for the Calinski and the Silhouette indices. Depending on the *Optimal method* selected by the user, CoDaPack writes into the *Output window* the information about the application the cluster procedure: size of clusters, *olr*-coordinates of clusters center, cluster assigned to each sample, and the ratio between the *between variation* and the *within variation*. In our case, according to the Calinski index the optimal number of clusters for this CoDa set is 3 (see Fig. 5.1). Moreover, CoDaPack creates a new variable *Group* with the cluster number –1, 2 or 3– assigned to each composition.

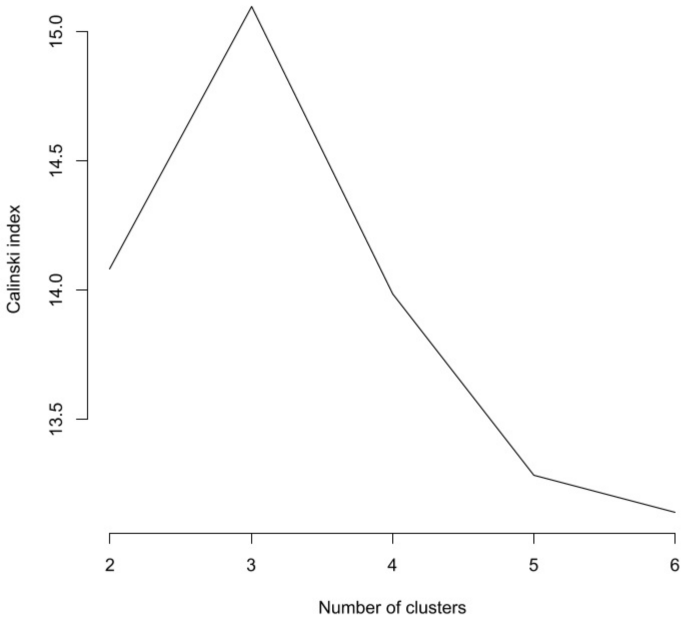


Figure 5.1. Calinski index plot (`alimentation` CoDa set)

- Draw a *clr*-biplot of the CoDa set using the cluster variable `Group` as a group in the *Groups* box. Change from *Cov.* to *Form* biplot. Do you appreciate big differences?
    - To improve the interpretations, label the samples in the *clr*-biplot using the name of the country.

No big differences are appreciated when one moves from *covariance* to *form clr*-biplot. Bulgaria, Romania, Yugoslavia, Albania and Hungary are the countries in group 1, associated with the `clr_C` and `clr_N` axes. Group 2 is composed by Portugal, Spain, Greece and Italy, that is, countries with a Mediterranean diet. This group is associated with the `clr_F` and `clr_FV` axes. Other countries are clustered in group 3, associated with the remaining axes (`clr_RM`, `clr_WM`, `clr_E`, `clr_M` and `clr_S`).

The biplot suggests an SBP (Table 5.1) of the parts to characterise the groups.

Using this SBP one can create the CoDa-dendrogram:

- Go to the *Graphs* ▷ *Balance dendrogram* menu to obtain the balances associated to this SBP using the `Group` as group in the *Groups* box. The option *Add balances* must be active.
- Go to the *Statistics* ▷ *Classical statistics summary* menu to obtain the percentiles of the *olr*-coordinates `ilr.1`, `ilr.2`,..., `ilr.8` by the factor `Group`.

Table 5.1. Sign matrix for the SBP

olr $_i$	RM	WM	E	M	F	C	S	N	FV
x_1^*	-1	-1	-1	-1	-1	+1	-1	+1	-1
x_2^*	0	0	0	0	0	+1	0	-1	0
x_3^*	-1	-1	-1	-1	+1	0	-1	0	+1
x_4^*	0	0	0	0	+1	0	0	0	-1
x_5^*	-1	+1	-1	-1	0	0	-1	0	0
x_6^*	+1	0	-1	-1	0	0	-1	0	0
x_7^*	0	0	+1	-1	0	0	-1	0	0
x_8^*	0	0	0	+1	0	0	-1	0	0

The dendrogram (see Fig. 5.2) indicates that the first *olr*-coordinate (e.g. the balance between the ln(C·N) and the logarithm of the product of the other parts) serves to separate group 1 from the other two groups. Indeed, this balance (stored in the variable `ilr.1`) varies from 1.7252 to 2.9436 in the five countries of the group 1, whereas it varies from -0.7929 to 1.5131 in the remaining countries. Therefore, the ratio between the average consumption of cereals (`C`) and nuts (`N`) and the average consumption of the other types of foodstuffs in the countries in group 1 (Bulgaria, Romania, Yugoslavia, Albania and Hungary) is greater than the same ratio in the other countries. The fourth balance ln(`F`/`FV`) (stored in the variable `ilr.4`) also separates group 1 from the other two groups. It varies from -1.8661 to -0.7281 in the countries in group 1, and from -0.5575 to 0.4146 in the remaining countries. Thus, the ratio `F`/`FV` between the consumption of fish and fruits&vegetables in the countries in group 1 is lower than the same ratio in the other countries. The third balance in the dendrogram (e.g. the balance between the ln(`F` · `FV`) and the logarithm of the product of `RM`, `WM`, `E`, `RM` and `M`) separates group 2 from the other two groups. Indeed, the third balance (stored in the variable `ilr.3`) varies from -0.1042 to 1.2449 in the four countries in group 2, whereas it varies from -1.9914 to -0.3090 in the remaining countries. This means that the ratio between the average consumption of fish (`C`) and fruits&vegetables (`FV`) and the average of consumption of `RM`, `WM`, `E`, `RM` and `M` in the countries in group 2 (Portugal, Spain, Greece and Italy) is greater than the same ratio in the other countries.

This interpretation can be completed by plotting the multiple box plot of the 9-part composition by the factor *Group*:

- Go to the *Graphs* ▷ *Boxplot* menu to obtain boxplot of all raw parts using the factor `Group` as group in the *Groups* box. The option *Draw mean* for the `Geometric` mean should be active.

For example, observe the large values in the part `C` for counties in group 1 or that the group 2 takes the largest values in the parts `F` and `FV`, whereas countries assigned to group 3 take the largest values in the parts `RM`, `WM` and `M`. With all this information on hand one can interpret why the numerator or the denominator in a particular balance takes large or small values. Another graphical technique which is very useful to characterize the clusters as regards the average values of the parts is the *Geometric mean bar plot*.

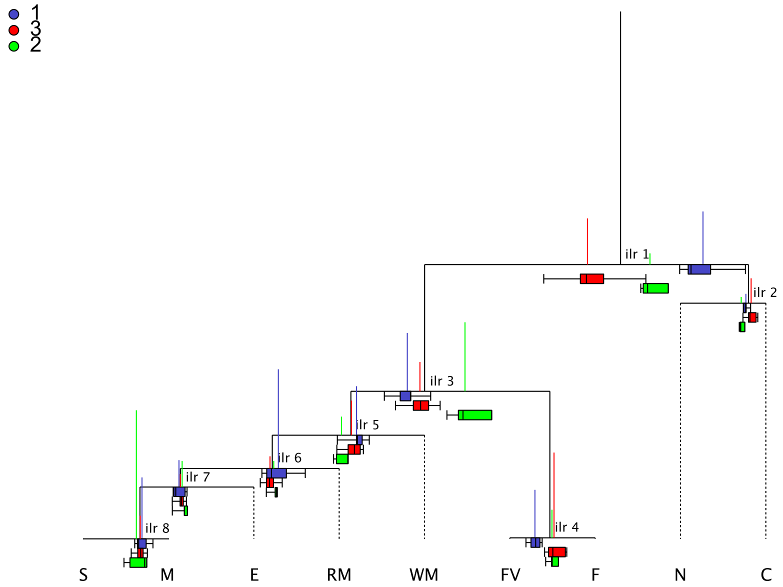


Figure 5.2. Balance dendrogram by cluster of `alimentation` CoDa set (Group 1: blue boxplot; Group 2: green boxplot; Group 3: red boxplot).

- Go to the *Graphs* ▷ *Geometric mean barplot* menu to obtain bar plot of all raw parts using the factor `Group` as group in the *Groups* box.

Here, for example, we appreciate that countries in group 1 (black bars) also are characterized by a low value, in average, in the part `F`. In addition, we conclude that group 2 (red bars) is the *champion* as regards the part `FV`.

5.2. Discriminant analysis

Activities for Section 5.2

In this activity, we use the `petrafm` CoDa set described in the Appendix to perform a LDA.

- Load the `petrafm.cdp` file.

The data frame (*Tables*) `Petrafm` contains the 3-part compositions of 100 samples of volcanic rocks from Ontario (Canada). The parts are: $[A : Na_2O + K_2O; F : FeO + 0.8998 \cdot Fe_2O_3; M : MgO]$. The samples are already classified by the factor `group`: $n_{calc} = 25$ from the calc-alkaline magma series, and $n_{thol} = 75$ from the tholeiitic magma series. The two groups of samples are respectively split on two data frames in the file: *Tables*: `Petrafm.calc` and *Tables*: `Petrafm.thol`.

A. **Exploratory analysis**

- Select the `petrafm` data frame
- Use the *Graphs* ▷ *Boxplot* menu to draw a boxplot of the original parts [`F`,`A`,`M`] using the variable `group` to distinguish the samples (Fig. 5.3).

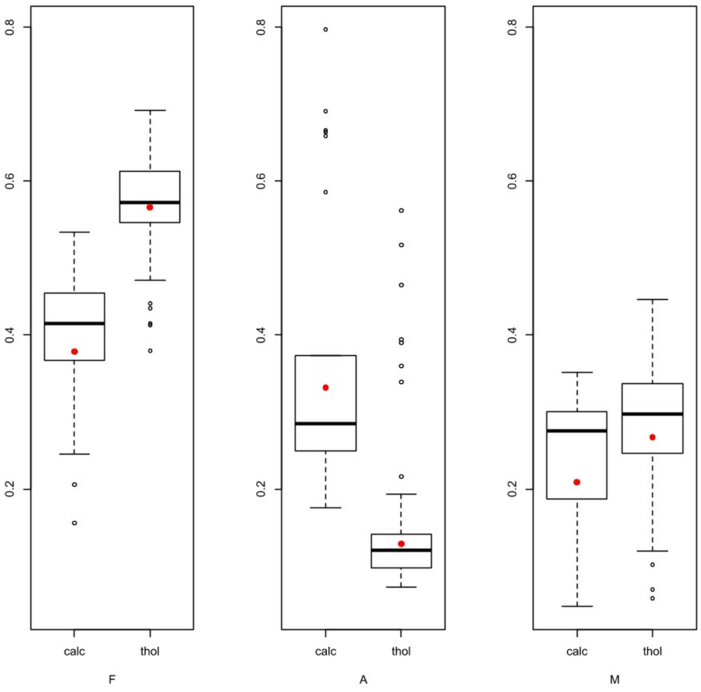


Figure 5.3. Boxplots of raw parts by group. `Petrafm` CoDa set

- Plot a ternary diagram of the compositional variables using the variable `group` to distinguish the samples.
  - ○ Is there any evidence of good separation between groups?

Observe that there is evidence of good separation between groups both with the boxplots and the ternary diagram. Tholeiitic samples contain relatively more of `F` oxides and relatively less of `A`. In the ternary diagram the groups are not overlapped. However, the *border* between them is not *linear* in a typical Euclidean sense, no typical straight line separate the two groups.

- Calculate the *olr*-coordinates of the `petrafm` compositions (use the default SBP).
- Go to *Graphs* ▷ *Scatterplot 2D/3D*. Plot `ilr.1` vs `ilr.2` using the variable `group` to distinguish the samples.

The figure suggests that, in *olr*-coordinates, the separation between the two groups seems more *linear*.

- Calculate the compositional statistics by groups of the variables `F`, `A` and `M`.

The centers of the classes seem to confirm the difference between the two groups suggested by the graphs.

- Plot the balance dendrogram of `F`, `A` and `M` using the the variable `group` to distinguish the samples (use the default SBP).

There is evidence of the differences between the means of $\ln((\mathtt{A} \cdot \mathtt{F})^{1/2}/\mathtt{M})$ in the two groups. The same evidence is present in relation to the logratio $\ln(\mathtt{F}/\mathtt{A})$. Moreover, the variances of these two logratios in the two groups are quite similar.

- Select the `petrafm.calc` data frame (it includes only the calc-alkaline samples)
- Test the log-ratio normality of the variables `F`, `A` and `M` (use the default SBP).

Observe that if we only take into account the *radius test*, the log-ratio normality of `F`, `A` and `M` in the calc-alkaline group cannot be rejected (the p-values of all three available tests are above 0.15).

- Select the `petrafm.thol` data frame (it only includes the tholeiitic samples).
- Test the log-ratio normality of the variables `F`, `A` and `M` (use the default SBP).

Observe that the p-values associated to the *radius test* are between 0.010 and 0.025. Therefore the log-ratio normality of `F`, `A` and `M` in the tholeiitic group could be rejected. Consequently, one should be cautious when interpreting results of LDA and consider applying other discriminant techniques, such as a non-parametric technique.

B. **LDA function**

CoDaPack allows to apply LDA to CoDa.

- Select the `petrafm` data frame (it includes all compositions).
- Go to *Statistics* ▷ *Multivariate Analysis* ▷ *Discriminant Analysis*.
  - ○ Select the composition `F`, `A` and `M` in the *Selected X* box.
  - ○ Select the `group` variable in *Groups*.
  - ○ Click the *Set X Partition* button, select the *Default Partition* and click the *Accept* button.
  - ○ Don't select any option into the *Options* window.
  - ○ Click the *Accept* button.

CoDaPack applies LDA using prior probabilities proportional to the group sizes. In the *Output* window we can read:

– the default prior probabilities (0.25 and 0.75) used to estimate the LDA function: 0.25 and 0.75;

– the means of the `ilr.1` and `ilr.2` variables in the two groups:

$$\overline{\mathbf{x}}^*_{calc} = (0.433, 0.093)\ ,\ \overline{\mathbf{x}}^*_{thol} = (0.008, 1.046)\ ;$$

– the coefficients of the `ilr.1` and `ilr.2` variables in the LDA function LD1:

$$\beta^*_1 = 2.981\ ,\ \beta^*_2 = 4.738.$$

- How can you interpret the coefficients of the LDA function?

The LDA function is

$$\hat{L}_{calc;thol}(\mathbf{u}) := 2.981u_1 + 4.738u_2 + K = 0 ,$$

where the value of constant $K = -4.167$ can be calculated using that $\hat{L}_{calc;thol}(\overline{\mathbf{x}}^*) = 0$, being $\hat{\mathbf{x}}^*$ the center of the data set calculated using the prior probabilities (0.25 and 0.75):

$$\overline{\mathbf{x}}^* = 0.25 \cdot \overline{\mathbf{x}}^*_{calc} + 0.75 \cdot \overline{\mathbf{x}}^*_{thol} = (0.114, 0.808),$$

which coincides, in this case, with the overall center of the CoDa set expressed in *olr*-coordinates. Because the sign matrix for the default SBP is

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix} ,$$

and the coefficients are both positive ($\beta_1^* = 2.981$, $\beta_2^* = 4.738$), we can expect large LDA-scores for samples with $\sqrt{F \cdot A} > M$ and $F > A$. We can use the expression of the LDA function in terms of a logcontrast to interpret the LDA-scores as regards the original parts. The LDA function expressed in terms of the original parts is

$$\hat{L}_{calc;thol}(\mathbf{u}) := (2.981, 4.738) \cdot \begin{bmatrix} \frac{1}{2}\sqrt{\frac{2}{3}} & \frac{1}{2}\sqrt{\frac{2}{3}} & -\sqrt{\frac{2}{3}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 \end{bmatrix} \cdot \begin{bmatrix} \ln \mathrm{F} \\ \ln \mathrm{A} \\ \ln \mathrm{M} \end{bmatrix} - 4.167 = 0 ,$$

that is,

$$\hat{L}_{calc;thol}(\mathbf{u}) := 4.567 \ln \mathrm{F} - 2.133 \ln \mathrm{A} - 2.434 \ln \mathrm{M} \;\; - 4.167 = 0 .$$

Relative large values in part F suggests large LDA function scores, that correspond to the group *thol*. In contrast, one expects small LDA function scores for samples in group *calc* because they take small values in part F and large in part A. We can also visualise the results from the panels of histograms and density function plots for the LDA function scores.

- Evaluate the density function plot (Fig. 5.4).
- Evaluate the overlay of the histograms.

Both histograms and density function plots suggest some overlap, that is, one can expect some miscalssification when this LDA function is used.

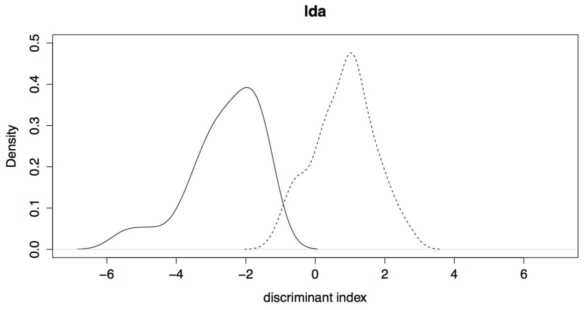


Figure 5.4. Densities of LDA function scores for the `Petrafm` CoDa set. Dotted line: group *thol*. Solid line: group *calc*. Density curves overlap suggests misclassification.

The successive tables in the *Output* window give the classification of samples in the two groups (calc-alkaline and tholeiitic) in accordance with the LDA function. For example, 84% (21/25) of the calc-alkaline samples are well classified in the calc-alkaline group.

- Assess the missclassification rate (percent correct for each group).
- Assess the accuracy of the prediction (addition of diagonal elements of the overall percent table).

Observe that 96% of the tholeiitic samples (72/75) are well classified, whereas only 84% of samples in group *calc* are correctly classified. Overall accuracy is 93%, that is, 7% of samples are missclassified.

C. **To predict the group for a new sample**

Which group can we assign to the new volcanic rock with a 3-part composition [`F`,`M`,`A`] equal to $[0.6, 0.1, 0.3]$?

First we calculate the *olr*-coordinates (default SBP) of the new composition. Then, the LDA function calculates the LD1 score of the new observation and gives the probability of belonging to each of the two groups.

- Go to *Data* ▷ *Create new table.*
  - ○ Write *3* into the *Number of columns* box.
  - ○ Write *2* into the *Number of rows* box.
  - ○ Press the *Ok* button.
  - ○ Write *F*, *A* and *M* in the three cells of the first row of the new table.
  - ○ Write *0.6*, *0.1* and *0.3* in the three cells of the second row of the new table.
  - ○ Press the *Accept* button.
  - ○ Write *Petrafm.new* into the *Enter the name for the new table* box.
  - ○ Press the *Ok* button.
- Go to *Data* ▷ *Manipulate* ▷ *Categorical to Numeric* to change the categorical variables *F*, *A* and *M* of the new table *Petrafm.new* to numeric,
- Select table *Petrafm.*
- Go to *Statistics* ▷ *Multivariate Analysis* ▷ *Discriminant Analysis.*
  - ○ Select the composition `F`, `A` and `M` in the *Selected X* box.
  - ○ Select the `group` variable in *Groups.*
  - ○ Click the *Set X Partition* button, select the *Default Partition* and click the *Accept* button.
  - ○ Select the option *Predicting for a new sample Z* and select the table *Petrafm.new.*
  - ○ Press the *Accept* button two times.
- Select the table *Petrafm.new.*

The new columns `posterior.calc` and `posterior.thol` give the posteriori probabilities of the new observation to belong to *calc* and *thol* groups, (0.0003 and 0.9997 respectively). They are estimated by the the linear discriminant function. The column `LD1` gives the LDA-score of the new observation while the column `class` gives the group to which it is assigned (*thol*).

- Select the table *Petrafm*. Explore the options *Discriminant scores*, *Max. posteriori prob. classification* and *Posterior prob. for the classes* of the *Statistics* ▷ *Multivariate Analysis* ▷ *Discriminant Analysis* menu.
- Go to *Data* ▷ *Create new table* and create a new table which includes the centre of the data set (i.e., the overall geometric mean $[0.5524, 0.1763, 0.2713]$). Investigate the group assigned by the LDA to the centre, that is, which group is assigned to the centre of the data set?
- Repeat the last exercise with the midpoint (geometric average) between the centres of the two groups instead the centre of the data set. Which group is assigned to the midpoint between the centres of the two groups?
- Create the LDA model when the prior probabilities is uniform (i.e. "fifty-fifty"): *Use of uniform prior* option in the *Statistics* ▷ *Multivariate Analysis* ▷ *Discriminant Analysis* menu.
- Which group is assigned to the centre of the data set when the prior probabilities are the uniform (i.e. "fifty-fifty")?
- Which group is assigned to the midpoint between the centres of the two groups when the prior probabilities is uniform?

In the light of the results of exercises above, we can conclude that the *uniform* assumption is only realistic when the group sample sizes are similar, that is, the centre of the data set is similar to the midpoint between the centres of the groups.

5.3. MANOVA

Activities for Section 5.3

The `pollen` CoDa set (see Appendix) contains $n = 30$ fossil pollen compositions whose $D = 3$ parts are [`pinus`, `abies`, `quercus`]. The samples were collected in each of three different locations. This information is considered as a categorical variable (`group`). The question that we face is if the composition of pollen is significantly different from one location to the other.

- Import the file `pollen.txt` using the menu *File* ▷ *Import* ▷ *Import CSV/Text Data*.
  Note that the variable `group`, which records the location, is numerical.
- Go to menu *Data* ▷ *Manipulate* ▷ *Numerical to categorical* to transform the variable `group`.
- Go to menu *File* ▷ *Save workspace* and save the file in a CoDaPack format.
- Plot a ternary diagram of the composition [`pinus`, `abies`, `quercus`] using the variable `group` to distinguish the samples.
  It seems that the separation between groups is quite good.
- Calculate the *olr*-coordinates of the composition [`pinus`, `abies`, `quercus`] using the default SBP.

- Plot `ilr.1` vs `ilr.2` using the variable `group` to distinguish the samples.
  The separation between groups seems more evident than in the ternary representation.
- Calculate the compositional statistics of the 3-part composition [`pinus`, `abies`, `quercus`] by groups.
  Although we cannot draw any conclusion, the (geometric) centre of group 3 looks quite different from the centre of groups 1 and 2. From the variation arrays, it seems that there are no large differences between the variances of the logratios of the three groups, although the differences between the means of the logratios seem evident.
- Draw the *clr*-biplot of the 3-part composition [`pinus`, `abies`, `quercus`] using the variable `group` to distinguish the samples.
  Note that this figure makes the interpretation of groups easier. Group 1 stands out for `abies`, group 2 for `quercus` and group 3 for `pinus`
- Plot the balance dendrogram of the 3-part composition [`pinus`, `abies`, `quercus`] using the variable `group` to distinguish the samples and the default SBP.
  The dendrogram shows that the logratio ln(`pinus`/`abies`) separates samples of group 3 from the other two groups. Moreover, the variances of the two balances are quite similar in the three groups.

To detect if there is evidence in favor of the differences between the centers of groups we will perform a Manova analysis.

- Go to *Statistics* ▷ *Multivariate Analysis* ▷ *Manova.*
  - Select the three parts of the composition [`pinus`, `abies`, `quercus`].
  - Select the `group` variable in the *Groups* box.
  - Activate *Residuals* and *Dif. between pairs of groups* in the *Options* window.
  - Click the *Accept* button.

In the *Output* window there is the summary of the four test statistics (Wilks, Pillai, Hotelling and Roy). From these statistics is clear that the null hypothesis of equal mean vectors in the three groups must be rejected, as the p-values of all four tests are below 0.01. We can also see the *Between* (labeled `$Y` in the output), the *Within* (labeled `$Residuals` in the output) and the *Total* sum of squares matrices. Note that the *Between* sum of squares matrix seems to be large enough (compared to the *Total* sum of squares matrix) to reject the null hypothesis of equal mean vectors in the three groups.

We will make a bar plot to compare all the centres with the whole centre (in terms of geometric means).

- Go to *Graphs* ▷ *Geometric mean barplot.*
  - Select the three parts of the composition [`pinus`, `abies`, `quercus`].
  - Select the `group` variable in the *Groups* box.
  - Click the *Accept* button.

From the bar plot, note that the centre of group 3 (green bars) differs from the centres of the other two groups mainly in the `pinus` and `quercus` parts. The differences between groups 1 (black bars) and 2 (red bars) are mainly in the `quercus` part, whereas the differences between groups 1 and 3, and between groups 2 and 3 are mainly due to the `pinus` and `quercus` parts (Fig. 5.5).

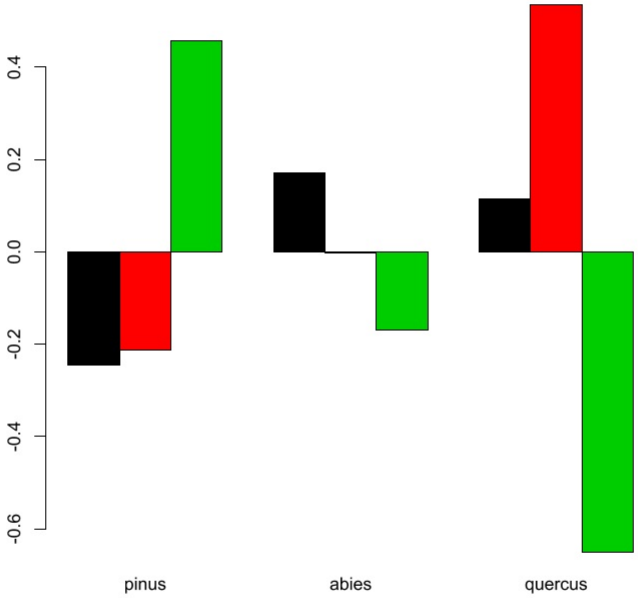


Figure 5.5. Geometric mean bar plot of pollen composition against group in the `Pollen` CoDa set (Group 1: black; group 2: red; group 3: green)

In the *Output* window there are also the p-values of the three comparisons between groups: 0.0003664 (1 vs. 2), 0.0000273 (1 vs. 3) and 0.0002034 (2 vs. 3). After multiplying them by 3 all of them are still lower than 0.05. Therefore, according to the Bonferroni correction, the means of the three groups are different, taking into account that we are testing for 3 pairwise differences.

The new variables `ilr.1.r` and `ilr.2.r` in the CoDaPack data frame are the *olr*-coordinates of the residuals associated to Manova analysis. Let's analyze them.

- Compute the inverse *olr*-transformation of the residuals (*Data* ▷ *Transformations* ▷ *ILR-Raw*) to calculate the 3-part residuals composition. Use the default SBP.
- Calculate the compositional statistics for the back-transformed residuals `inv.ilr.1`, `inv.ilr.2` and `inv.ilr.3` (deactivate the *Groups* option). What is the center of the residuals expressed as proportions?
  The residuals are always centered by construction, and thus the center is $[1/3, 1/3, 1/3]$.
- Plot the predictive regions ($\alpha = 0.1, 0.25, 0.5, 0.9, 0.95$) of the back-transformed residuals on the ternary diagram.
  The plot suggests that the normal distribution fits well the data set of residuals.
- Evaluate the log-ratio normality of the 3-part residual composition `inv.ilr.1`, `inv.ilr.2` and `inv.ilr.3` (*Statistics* ▷ *Log-Ratio Normality Test*, default SBP).

The assumption of normally distributed residuals is tenable, because the p-values of all radius tests are above 0.15.

- What is your final conclusion about the Manova test?
  We can conclude that all groups are significantly different from one another, group 1 standing out for a relatively low `pinus` content, group 2 for a high `quercus` content and group 3 for a high `pinus` and a low `quercus` content.

The chapter's key concepts

- ✓ Multivariate methods can be applied to CoDa when the principle of working on coordinates is assumed.
- ✓ Non-parametric methods require Aitchison distance or Kullback-Leibler dissimilarity. The log-ratio normal distribution model is the most common model for the parametric methods on the simplex.
- ✓ There are specific techniques and plots to facilitate the interpretation of results provided by the multivariate methods.

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.