# Methodological Concepts for Source Apportionment

**Peter Filzmoser**

**Institute of Statistics and Mathematical Methods in Economics**
**Vienna University of Technology**

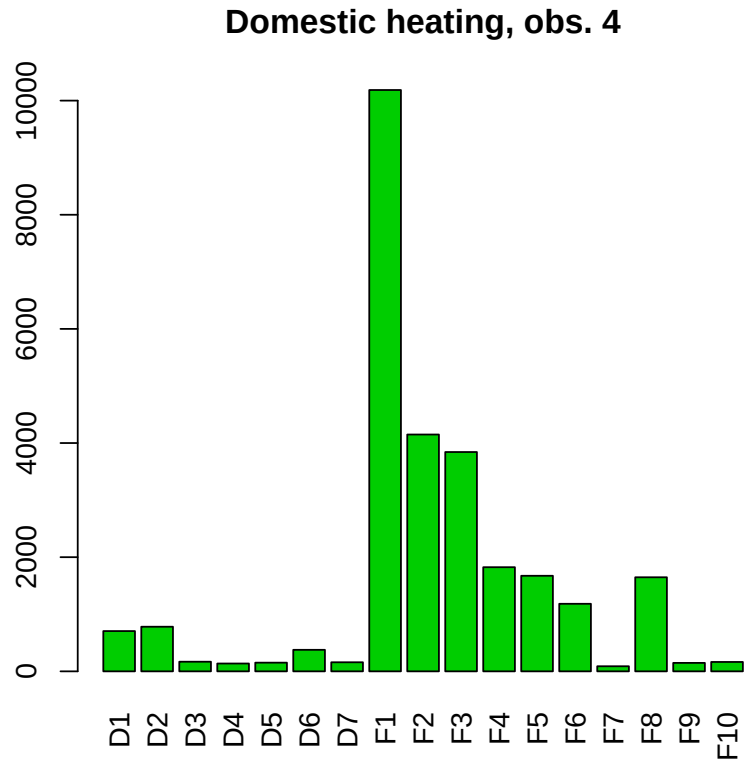
*UBA Berlin, Germany*

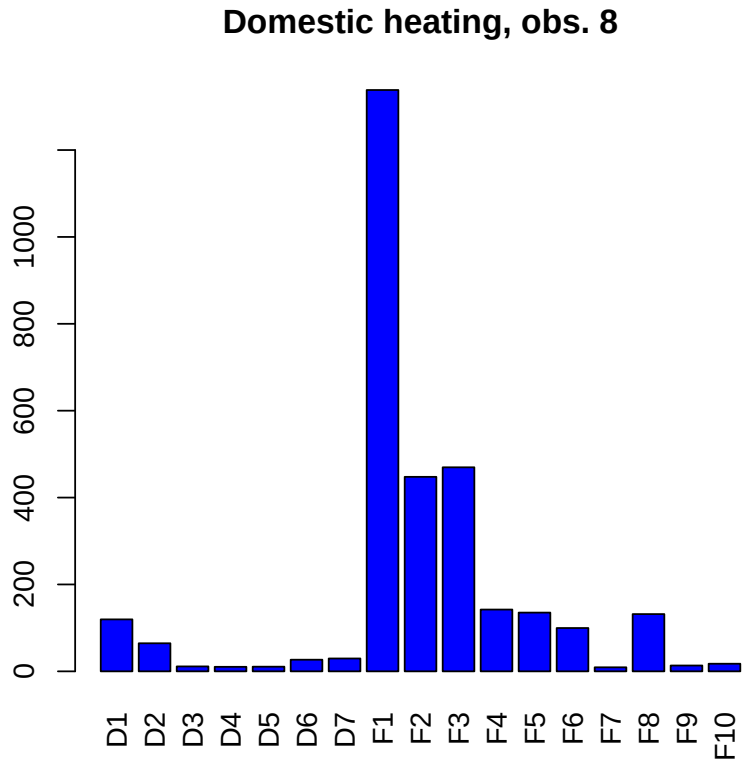
November 18, 2016

in collaboration with

# Example data: POPs

POPs: Persistent Organic Pollutants
7 dioxine compounds (D1-D7), 10 fouran compounds (F1-F10)
Two “similar” observations, with very different scale (concentration levels)!




Should we normalize to row sum 1, or analyze (log-)ratios?

# Compositional data

**Definition:** **Compositional data** consist of vectors $\mathbf{x} = (x_1, \ldots, x_D)$ with $D$ strictly positive components describing the parts on a whole, and which carry only relative information (Aitchison, 1986; Egozcue, 2009).

**Consequences:**

- The values $x_1, \ldots, x_D$ as such are not informative, but only their ratios are of interest.
- The parts $x_1, \ldots, x_D$ do not need to sum up to 1.
- Compositional data follow the so-called Aitchison geometry on the simplex (and not the standard Euclidean geometry).

**Key reference:**

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, U.K., 1986.

# Example data: POPs

Two compositions with $D$ parts:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)$$

$$\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D)$$

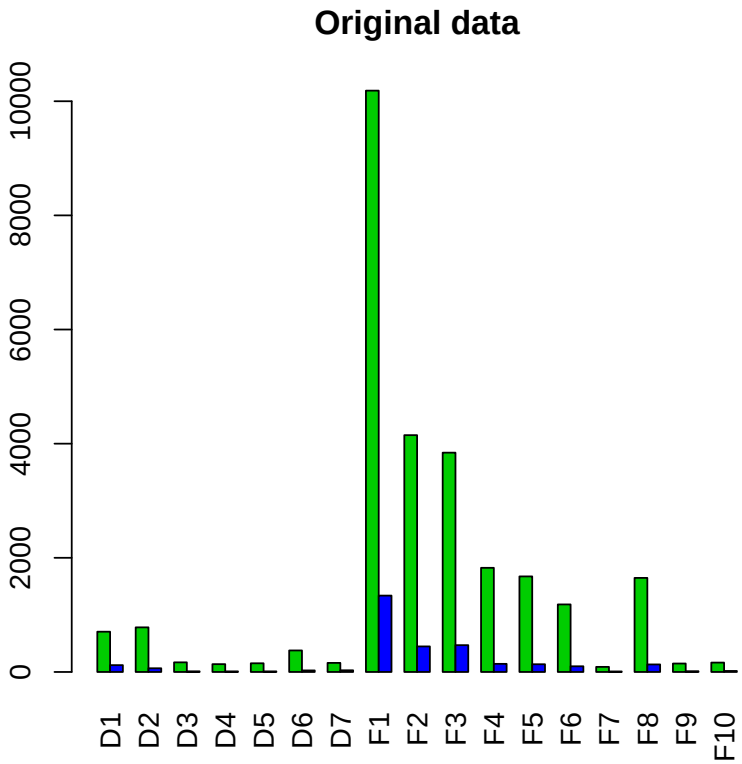
Aitchison distance between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ is defined as:

$$d_A(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \left( \log \frac{x_i}{x_j} - \log \frac{\tilde{x}_i}{\tilde{x}_j} \right)^2$$

For comparison: Euclidean distance:

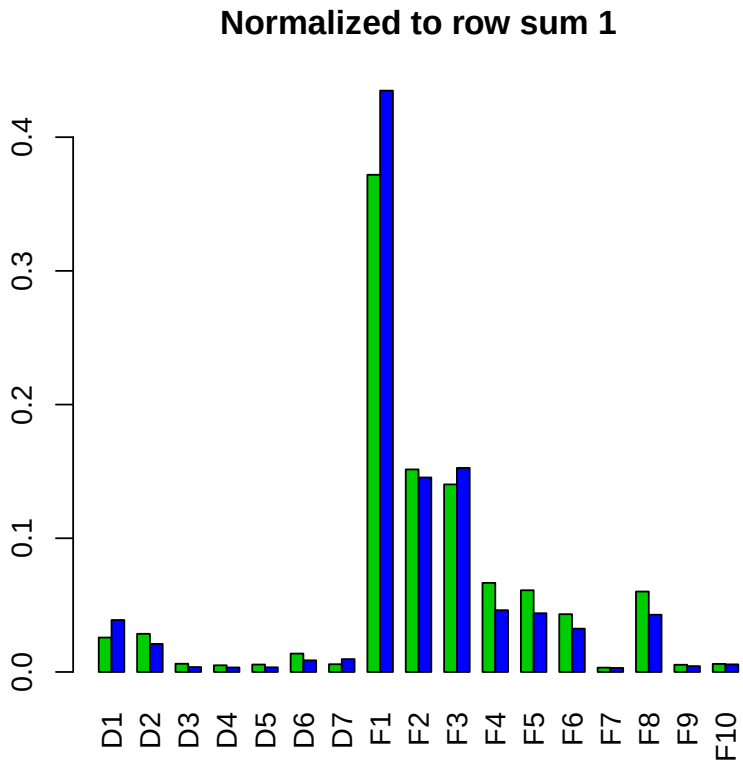
$$d_E(\mathbf{x}, \tilde{\mathbf{x}}) = \sqrt{\sum_{i=1}^{D} (x_i - \tilde{x}_i)^2}$$

# Example data: POPs




Aitchison distance: 1.21

Euclidean distance: 10636




Aitchison distance: 1.21

Euclidean distance: 0.075

# Euclidean geometry

Represent information from the simplex in the Euclidean geometry:

- **clr (*centered log-ratio*) coefficients:**
  Divide values by the **geometric mean**:

$$\mathbf{y} = (y_1, \ldots, y_D) = \left( \log \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)$$

- **ilr (*isometric log-ratio*) coordinates:**
  $\mathbf{z} = (z_1, \ldots, z_{D-1})$ can be defined by

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \qquad i = 1, \ldots, D-1.$$

# clr & ilr

Both, clr and ilr are **isometric**, which means that

$$d_A(\mathbf{x}, \tilde{\mathbf{x}}) = d_E(\text{clr}(\mathbf{x}), \text{clr}(\tilde{\mathbf{x}})) = d_E(\text{ilr}(\mathbf{x}), \text{ilr}(\tilde{\mathbf{x}}))$$

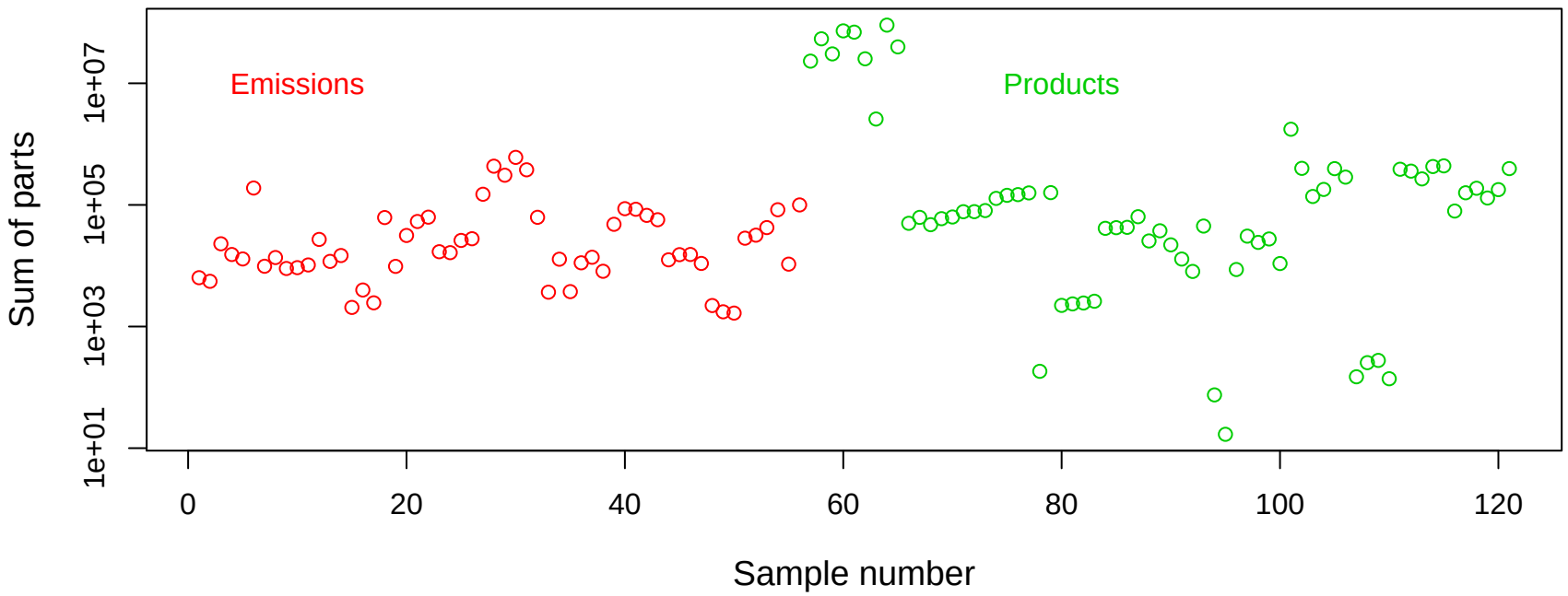
where $d_E$ stands for **Euclidean distance**.

This means that after expressing the information in clr coefficients or ilr coordinates, the data are in the usual **Euclidean geometry**. Most standard statistical methods are designed for this geometry.

A **log-transformation** does not transfer the compositions to this Euclidean geometry!

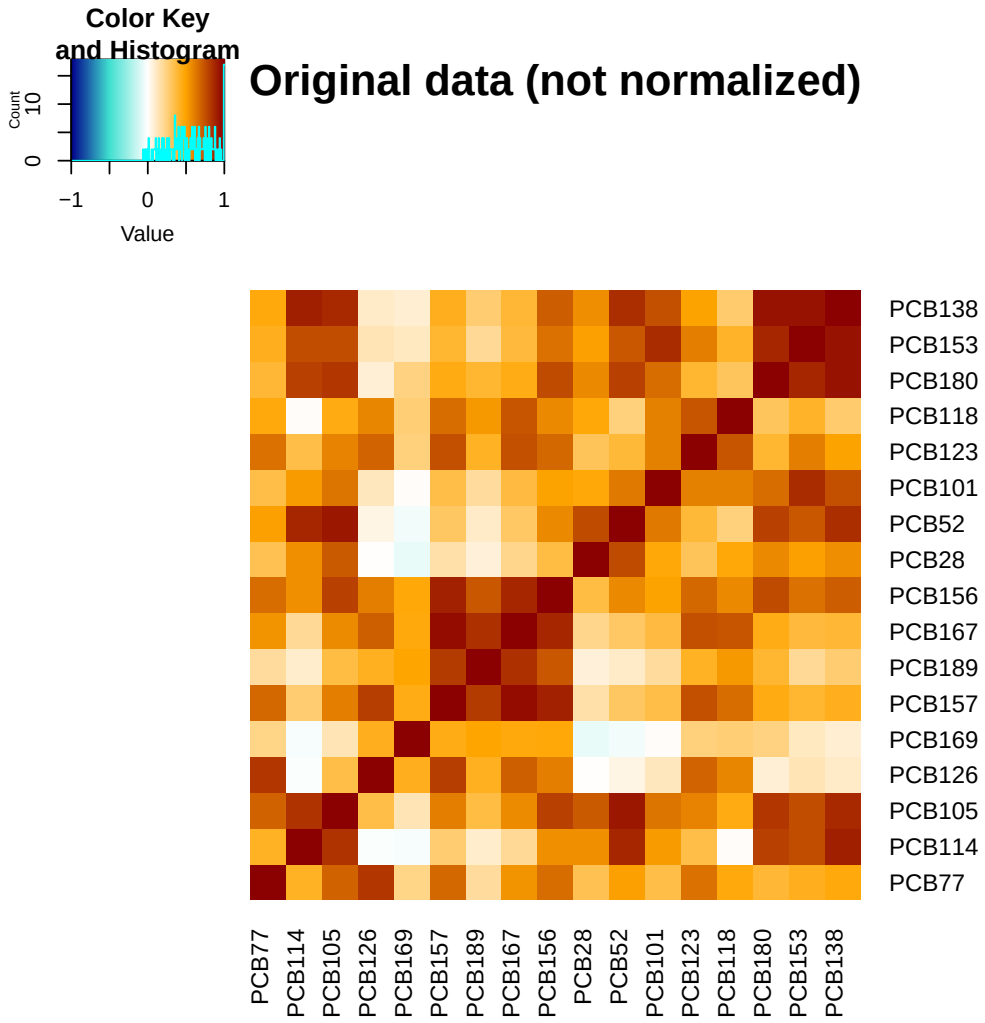
# Example data

Data set with compositional parts (dioxins and indicator PCBs)
PCB77, PCB126, PCB169, PCB105, PCB114, PCB118, PCB123, PCB156, PCB157, PCB167, PCB189, PCB28, PCB52, PCB101, PCB138, PCB153, PCB180
measured in 56 “emissions” (ng/m$^3$) and 65 “products” ($\mu$g/g or ng/g).



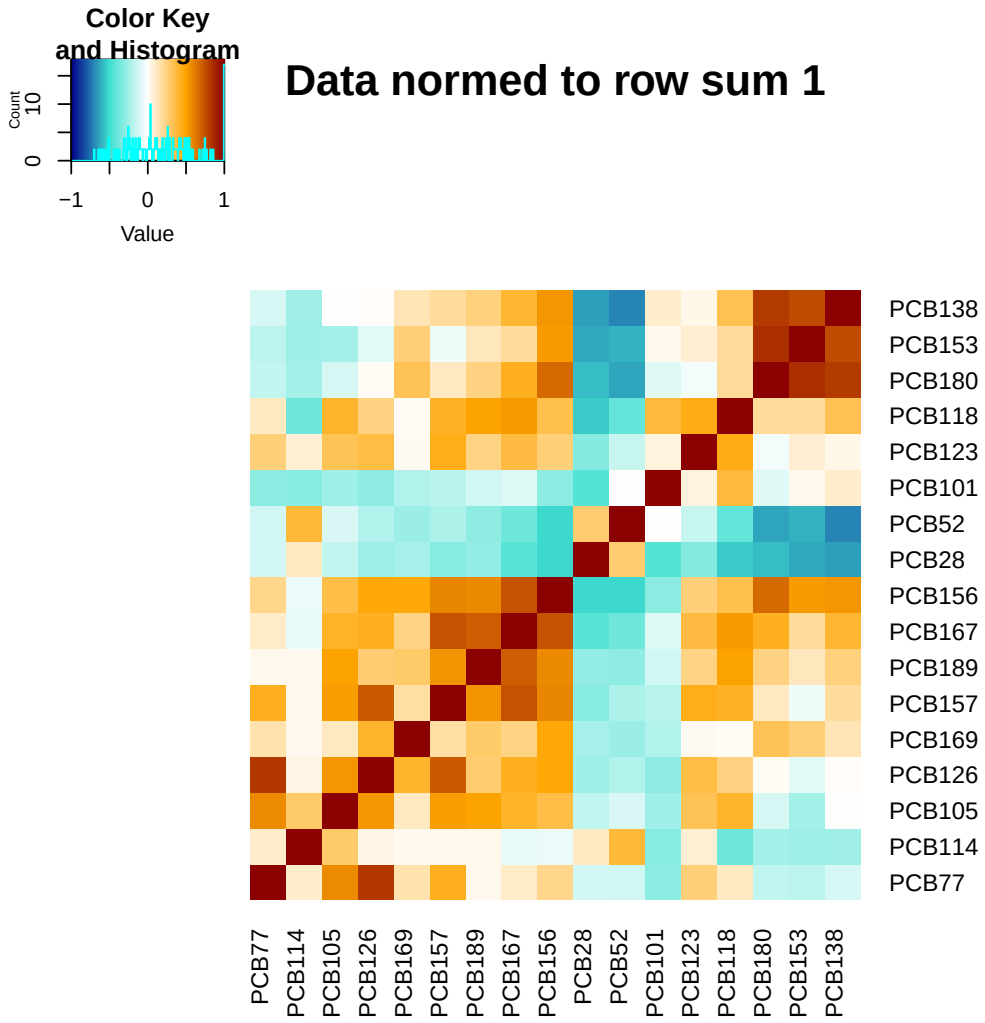
# Correlation matrix

**Correlations** for “Emission” computed “classically” from original data:



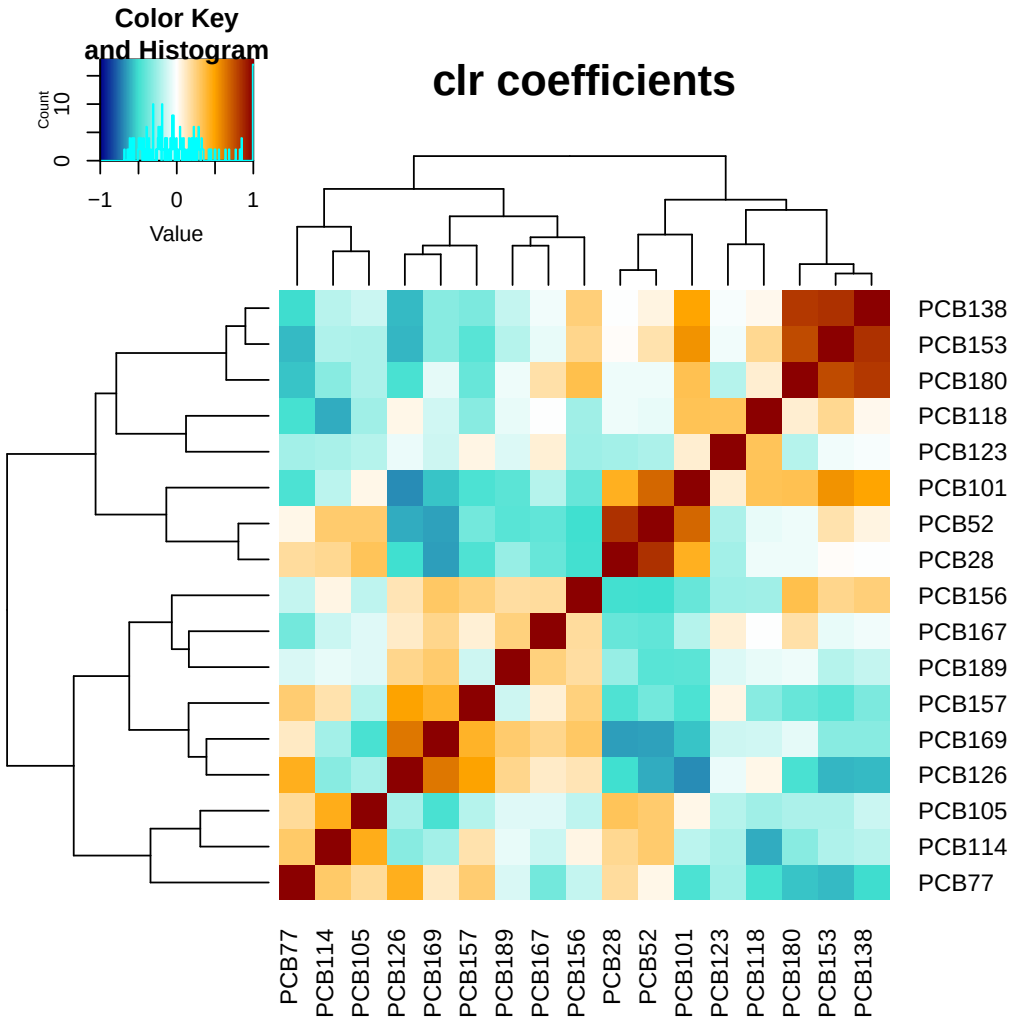
# Correlation matrix

**Correlations** for “Emission” computed “classically” from “length 1 data”:



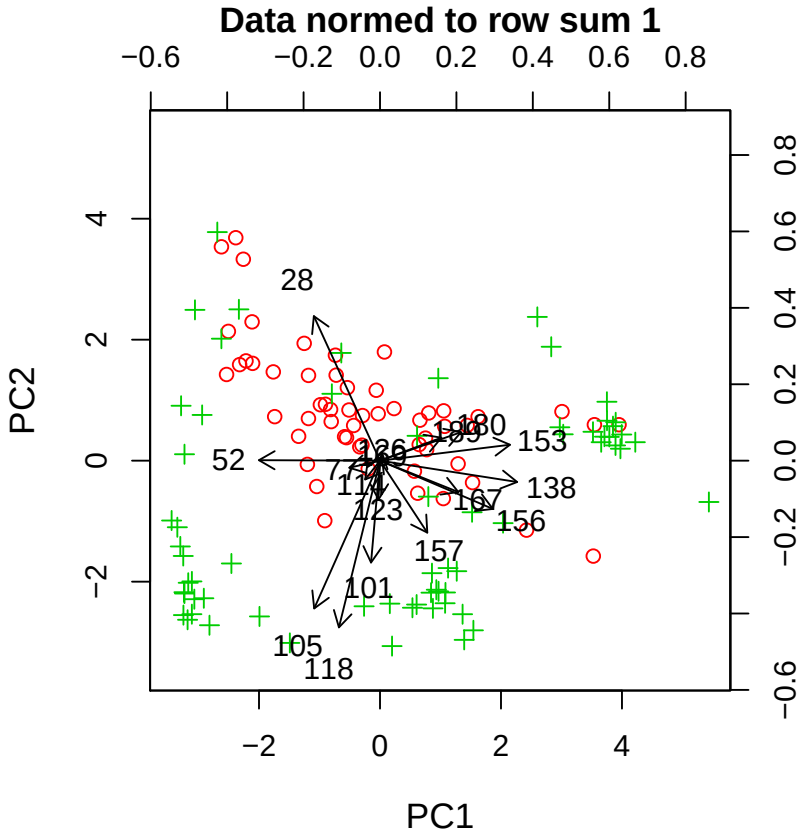
# Correlation matrix

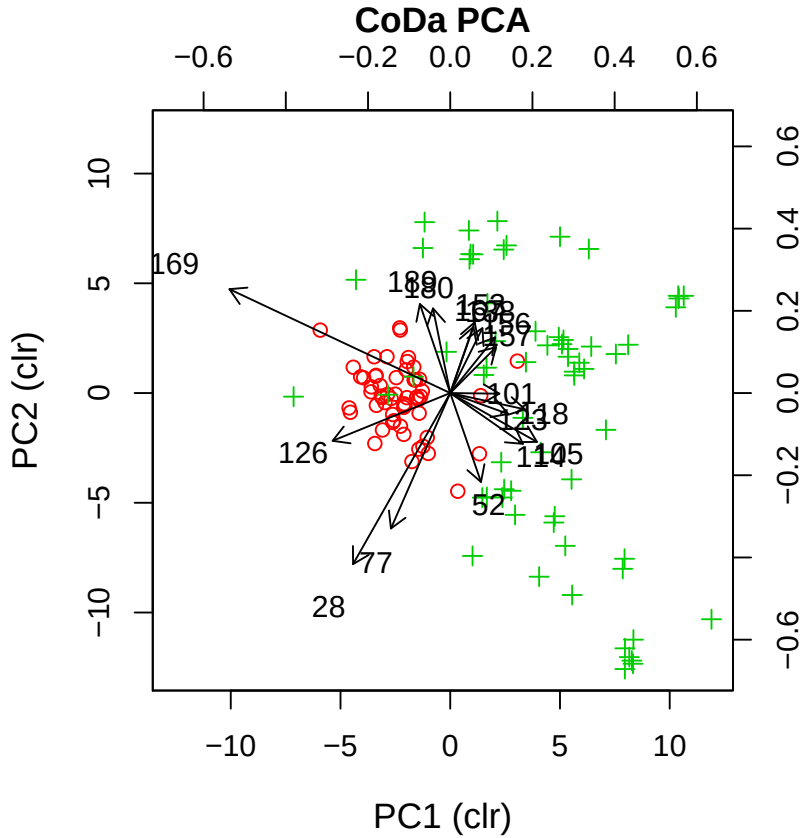
**Correlations** for “Emission” computed from clr coefficients:



# Robust PCA

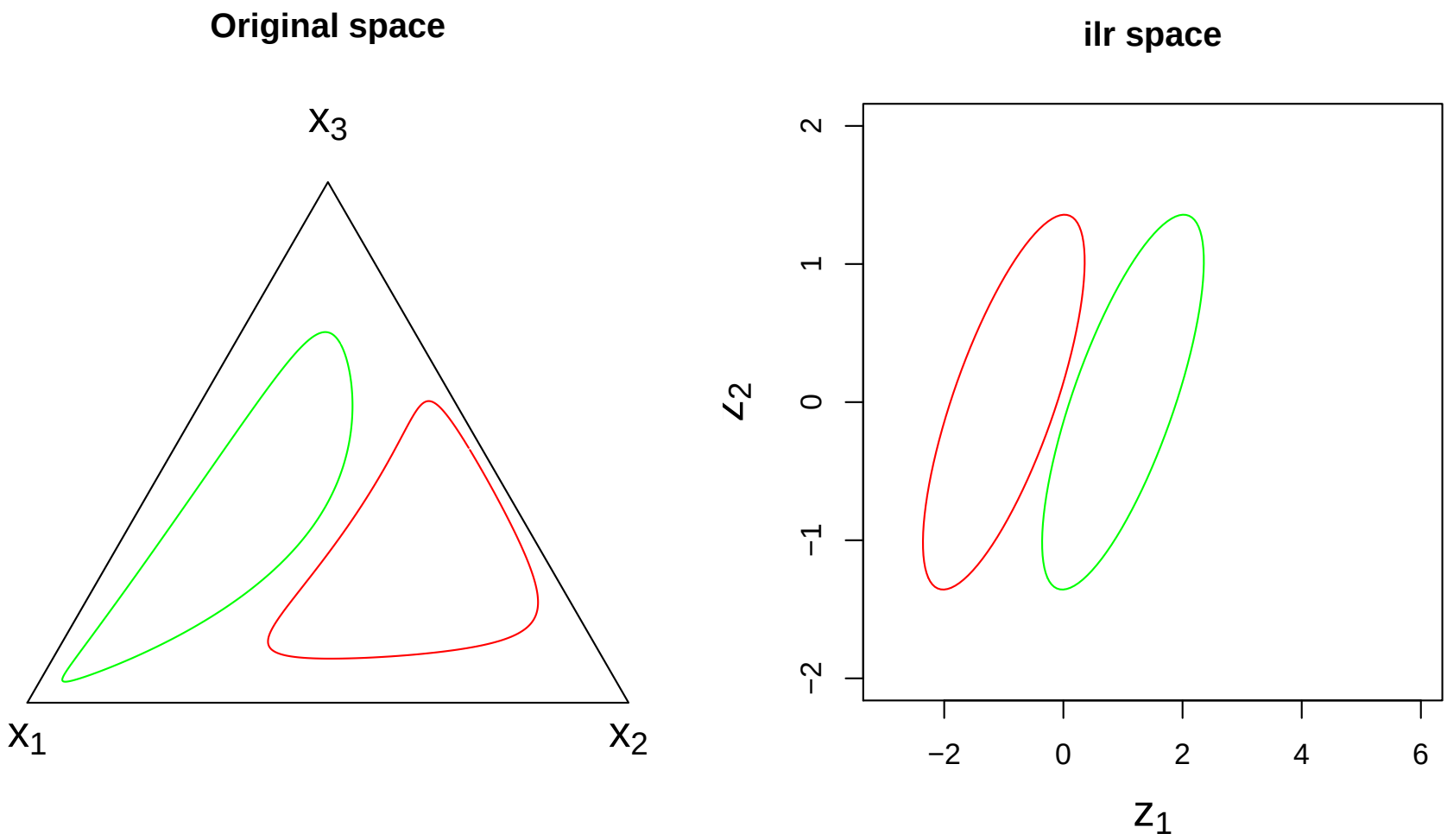
for data normed to row sum 1 (left) and CoDa PCA (right)



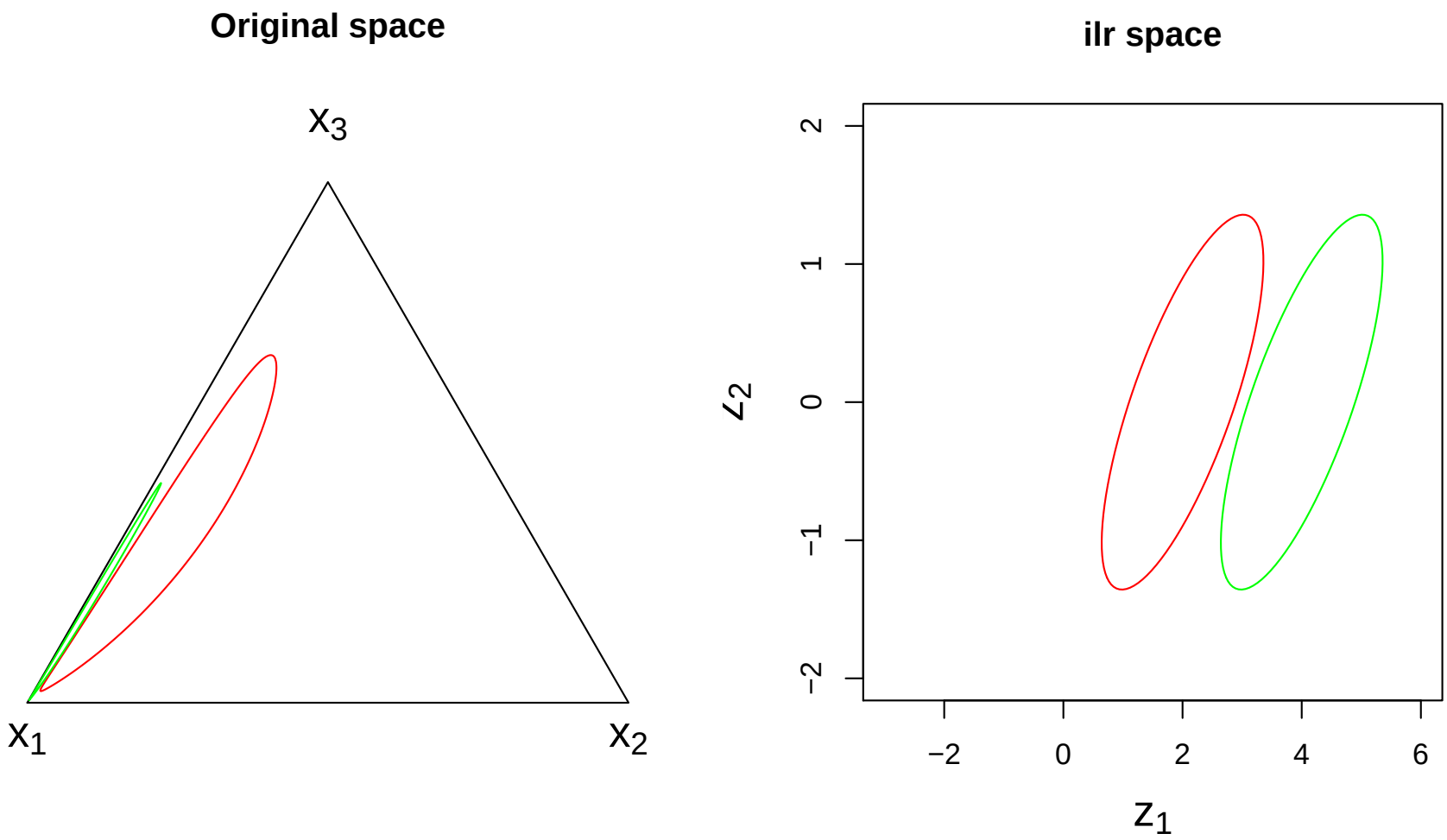
# Discriminate two groups

Two groups (red, green), normally distributed, equal covariance matrix:



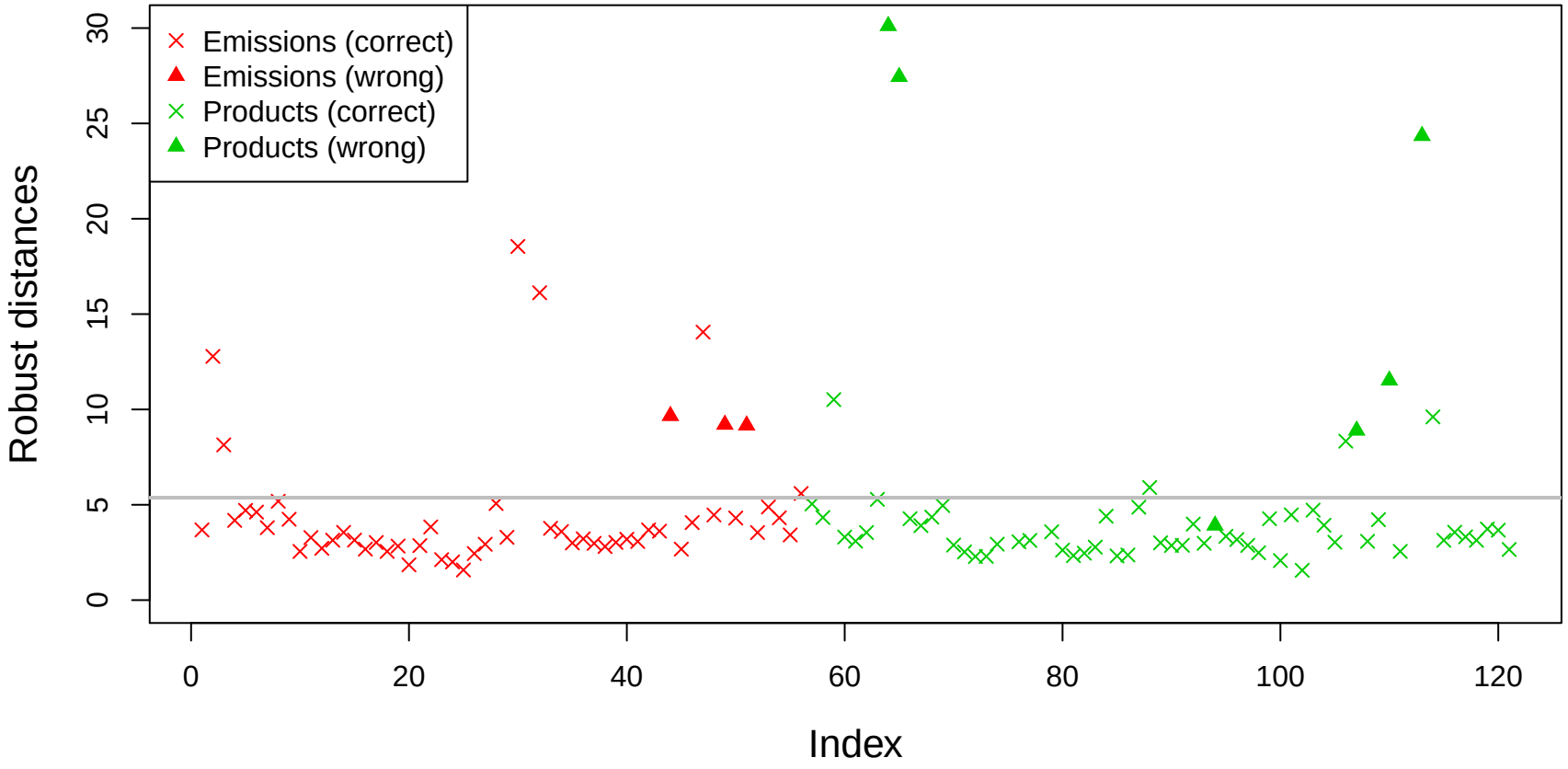
# Discriminate two groups

Two groups (red, green), normally distributed, equal covariance matrix:



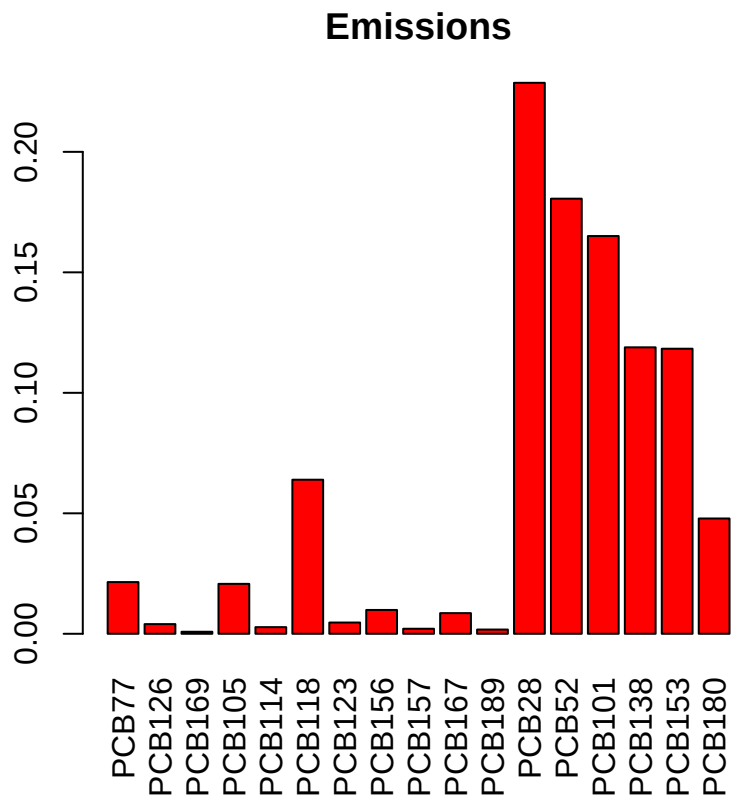
# Robust QDA (all POPs)

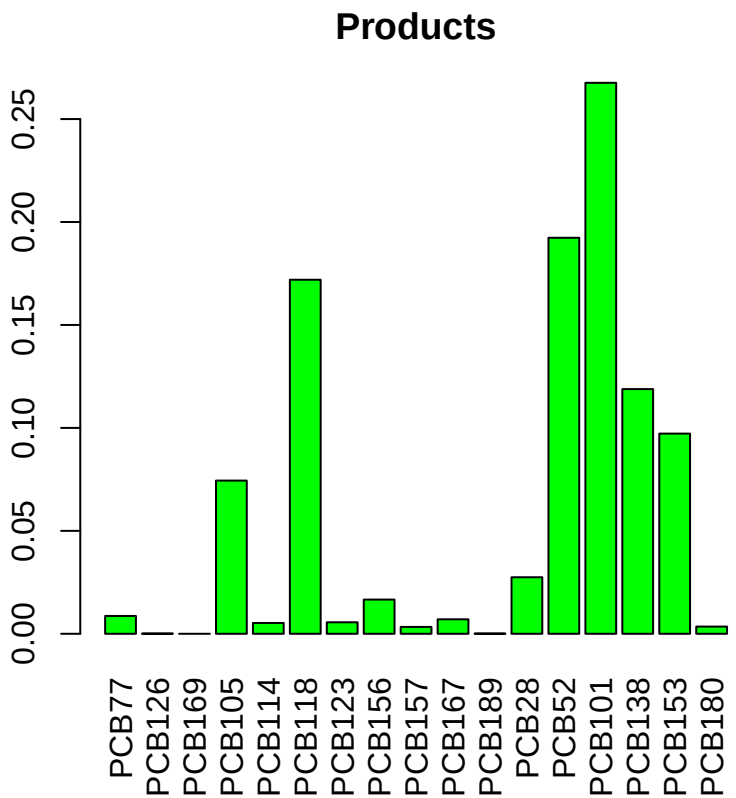
**Robust QDA** is applied to the ilr coordinates of the complete data set:
**Diagnostics** based on robust distances: Mahalanobis distances for each group, using robust estimates of location and covariance from QDA.
Can be used to assign (or not) new observations to a group.



# Group profiles

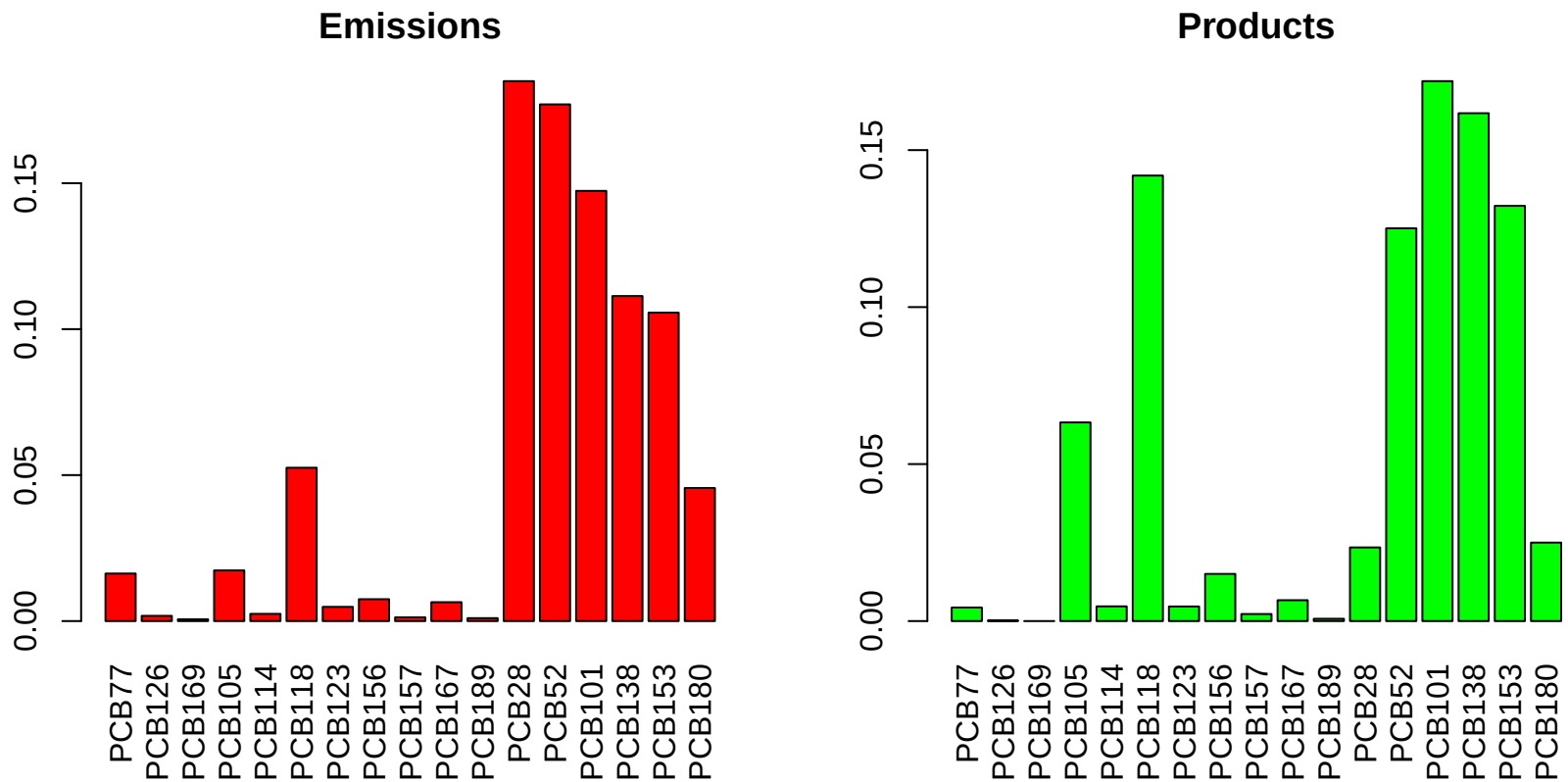
**Robust group centers** can be back-transformed from ilr coordinates to the simplex.



# Group profiles

**Group centers** obtained from median profiles,
incorrectly computed in the simplex.



# Benefit of log-ratios

So, what are ilr coordinates?

They form an orthonormal basis, considering ALL pairwise log-ratios.

They form an isometry, i.e.

Aitchison distances in the simplex = Euclidean distances in ilr.

log-ratios are scale invariant, i.e.

$$\log\left(\frac{x_i}{x_j}\right) = \log\left(\frac{\text{const} \cdot x_i}{\text{const} \cdot x_j}\right)$$

Thus, no matter if the data sum up to 1, and

no matter if the data sum up to completely different concentration levels!

# Why log-ratios?

- Compositions consist of relative contributions to a whole (mg/kg, %, proportions, etc.).
  This relative information is analyzed in form of log-ratios (log guarantees same variance if nominator and denominator change).
- Compositions live in the **simplex sample space**. They do not follow the usual Euclidean geometry, for which most statistical methods are designed.
- Work with **ilr coordinates**, and transform back for an interpretation, if necessary.
- An appropriate treatment does not necessarily lead to better results, but at least the interpretation of the results is correct.

# Data quality issues

- Missing values can be imputed (with CoDa methods!), but they should occur only rarely!
- Values BDL can be estimated (with CoDa methods!), but they should occur only rarely, and the DL has to be known!
- Usually, multivariate methods require more samples than variables!
- If only few samples per group: covariance-based methods (discriminant analysis) are problematic $\longrightarrow$ switch to distance-based methods

# Some references


J. Aitchison (1986). *The statistical analysis of compositional data.* Monographs on statistics and applied probability. Chapman & Hall, London.

J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueraz, C. Barcelo-Vidal (2003). Isometric log-ratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279-300.

P. Filzmoser, K. Hron, and M. Templ (2012). Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 27(4):585-604.

V. Pawlowsky-Glahn and A. Buccianti (2011). *Compositional data analysis: Theory and applications*. Wiley, Chichester.

V. Pawlowsky-Glahn, J.J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. Wiley, Chichester.

K.G. van den Boogaart and R. Tolosana-Delgado (2013). *Analyzing compositional data with R*. Springer, Heidelberg.

# Software R

For the computations, we used the R package “robCompositions”:

- Principal component analysis (here robust):

```
res <- pcaCoDa(x)          # x is the original data matrix
biplot(res)                # shows a biplot
```

- Linear discriminant analysis (LDA):

```
x.ilr <- isomLR(x)         # express data x in ilr coordinates
library(rrcov)
res <- LdaClassic(x.ilr,grp)  # grp contains grouping info
predict(res)               # shows error rate and confusion table
```

Use “Linda” for robust LDA, “QdaClassic” for quadratic discriminant analysis (QDA), and “QdaCov” for robust QDA.