Robust multivariate methods for compositional data

Peter Filzmoser

Department of Statistics and Probability Theory
Vienna University of Technology

Compstat – Paris, France

August 23, 2010

Vienna University of Technology

Contents

Joint work with . . .

Karel Hron, Univ. Olomouc, Czech Republic

Clemens Reimann, Geological Survey of Norway

Robert Garrett, Geological Survey of Canada

Example household expenditures

Household Expenditures in former HK$ (Aitchison, 1986)

Person	Housing	Foodstuff	Alcohol	Tobacco	Other goods	Total
1	640	328	147	169	196	1480
2	1800	484	515	2291	912	6002
3	2085	445	725	8373	1732	13360
4	616	331	126	117	149	1339
5	875	368	191	290	275	1999
6	770	364	196	242	236	1808
7	990	415	284	588	420	2697
8	414	305	94	68	112	993
⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	1195	443	329	974	523	3464
19	2180	521	553	2781	1010	7045
20	1017	410	225	419	345	2416

Characterization of compositional data

Definition: **Compositional data** consist of real-valued vectors $\mathbf{x} = (x_1, \ldots, x_D)^t$ 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 Euclidean geometry).

Most important reference:

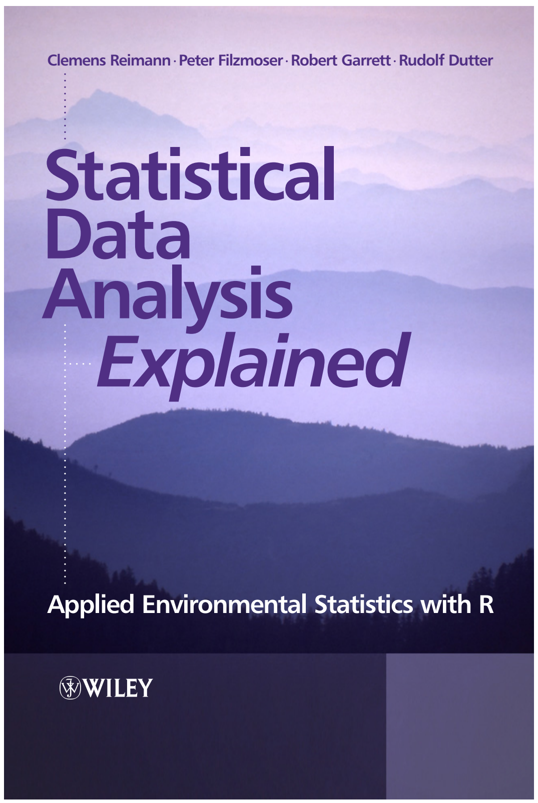
J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, U.K., 1986.

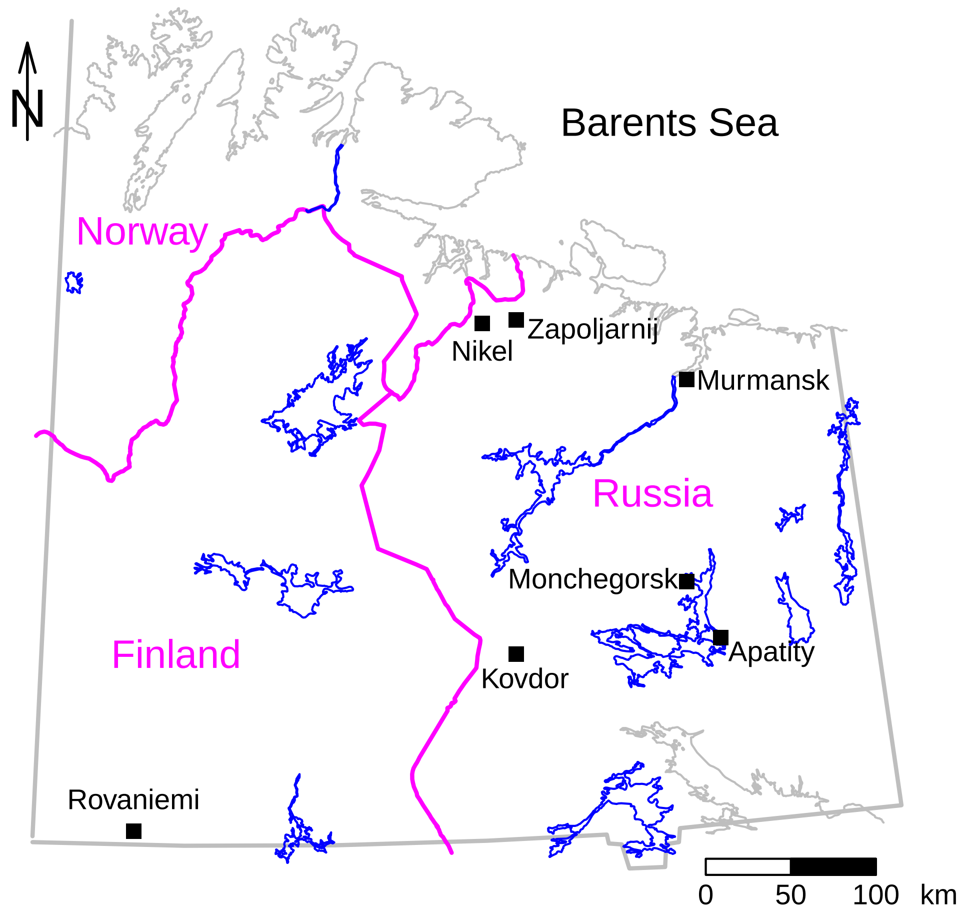
Example Kola data

Kola data: `library(StatDA)`

about 600 samples

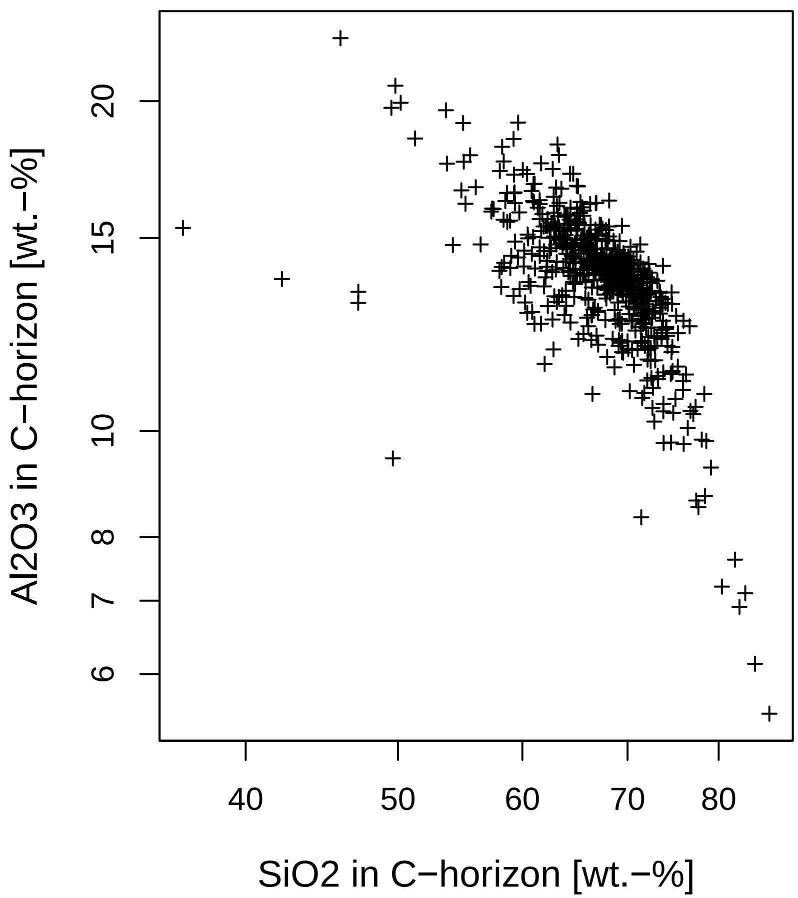
from 4 soil layers

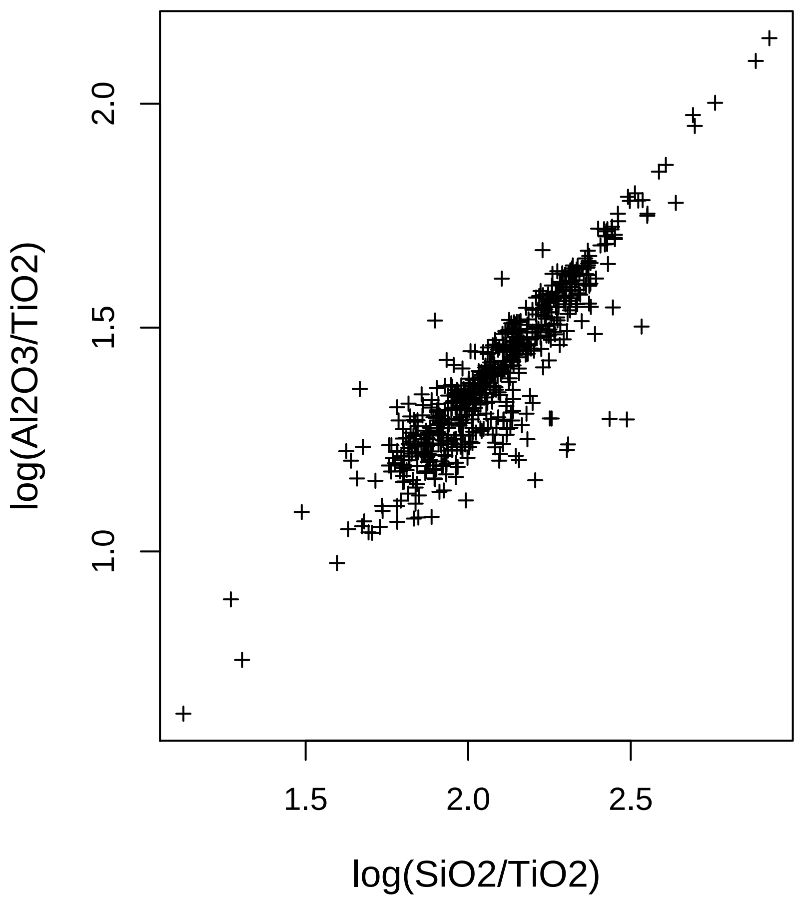




Example Kola data

Two dominant parts in the C-horizon:





Example factor analysis

(Reimann, Filzmoser, Garrett, 2002, *Appl. Geochem.*)

Kola moss data:

```
library(StatDA)
data(moss)
```

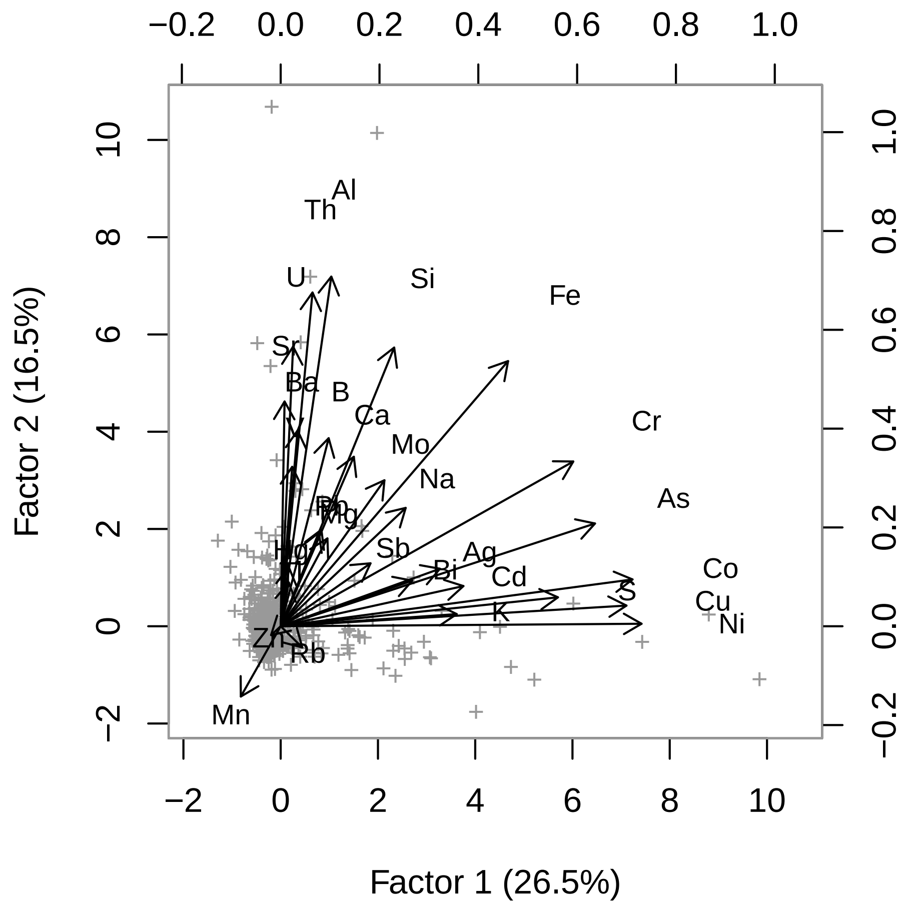
594 samples

31 variables

Factor analysis:

- log-transformation
- results presented in biplots

⟹ industrial contamination!



BUT: We have compositional data!

Example household expenditures

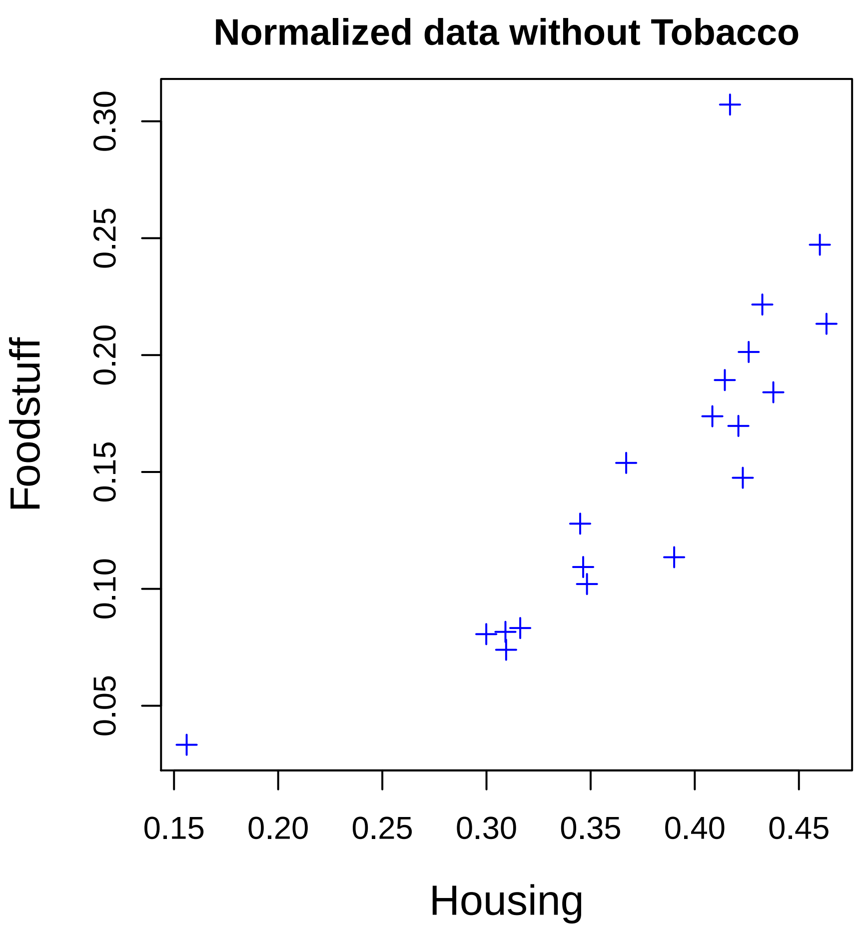
Household Expenditures in former HK$ (Aitchison, 1986)

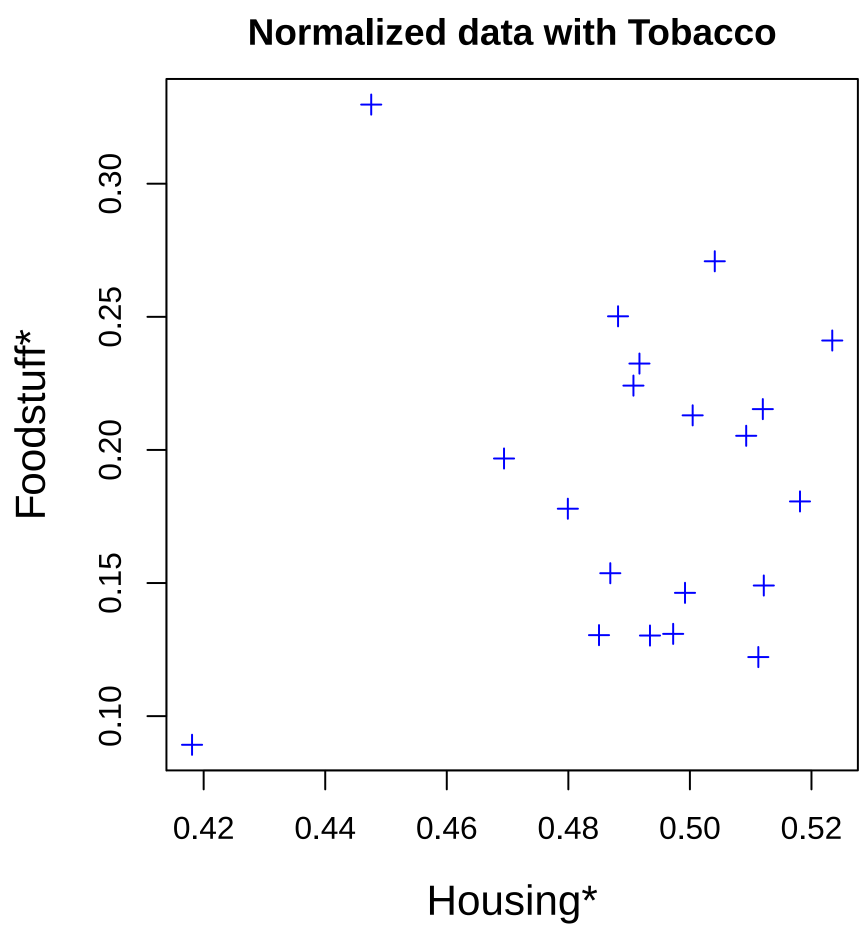
Person	Housing	Foodstuff	Alcohol	Tobacco	Other goods	Total
1	640	328	147	169	196	1480
2	1800	484	515	2291	912	6002
3	2085	445	725	8373	1732	13360
4	616	331	126	117	149	1339
5	875	368	191	290	275	1999
6	770	364	196	242	236	1808
7	990	415	284	588	420	2697
8	414	305	94	68	112	993
⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	1195	443	329	974	523	3464
19	2180	521	553	2781	1010	7045
20	1017	410	225	419	345	2416

Example household expenditures

Two versions: Data with and without `Tobacco`

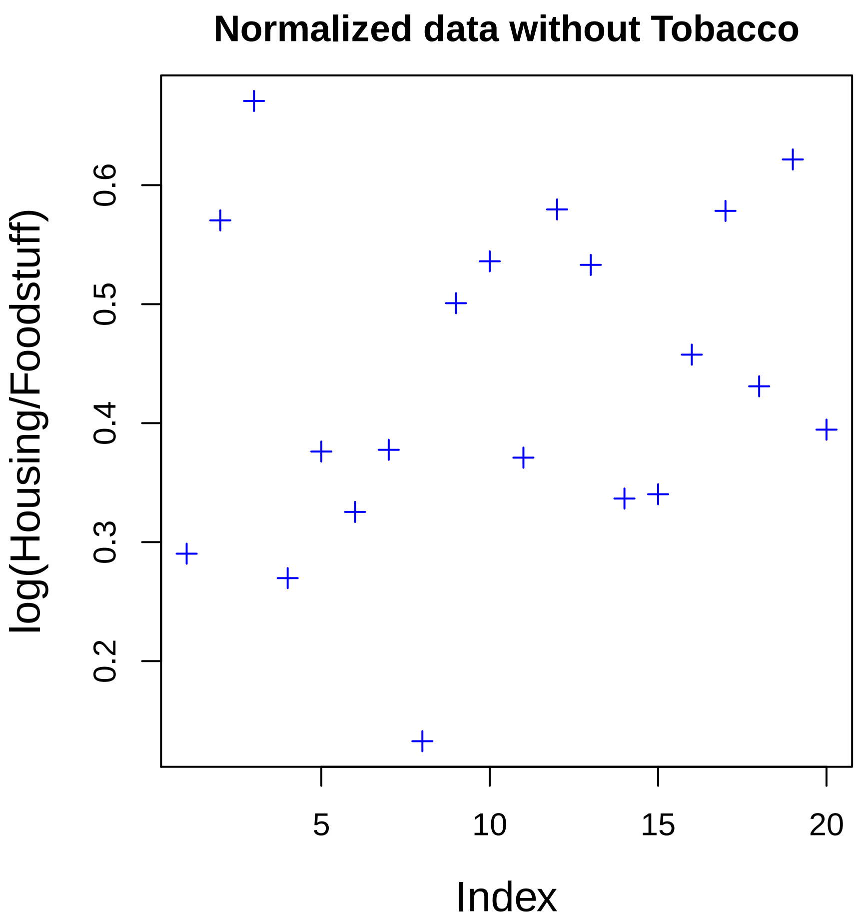
Data are normalized with the total expenditures

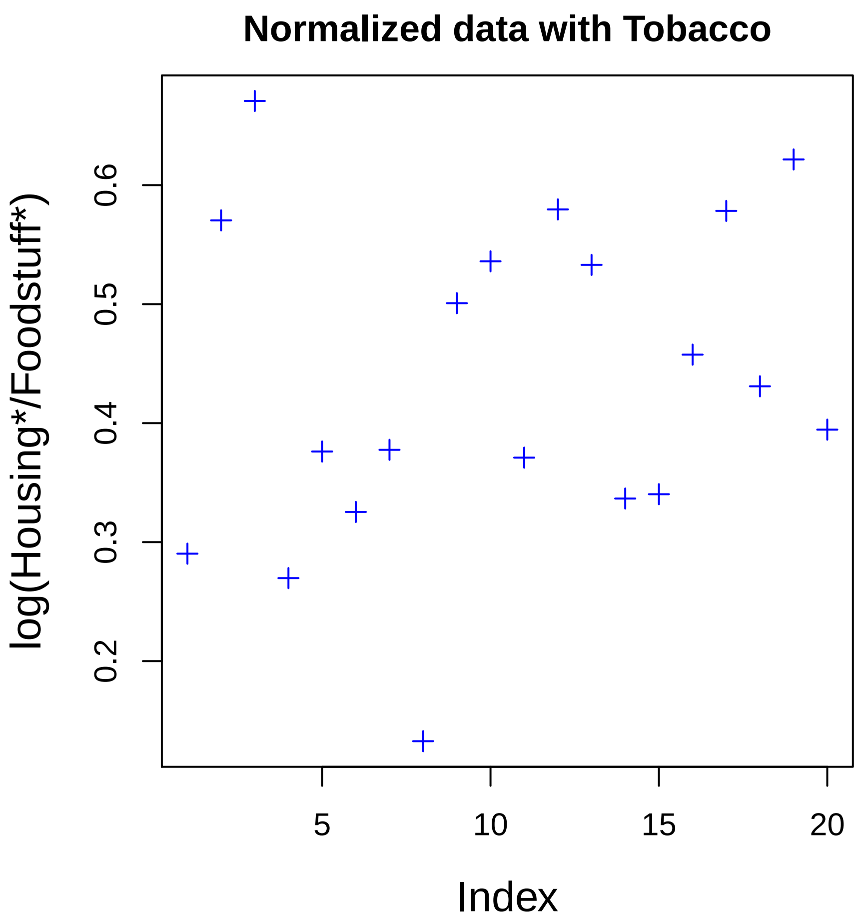




Example household expenditures

Solution: consider (log-)ratios

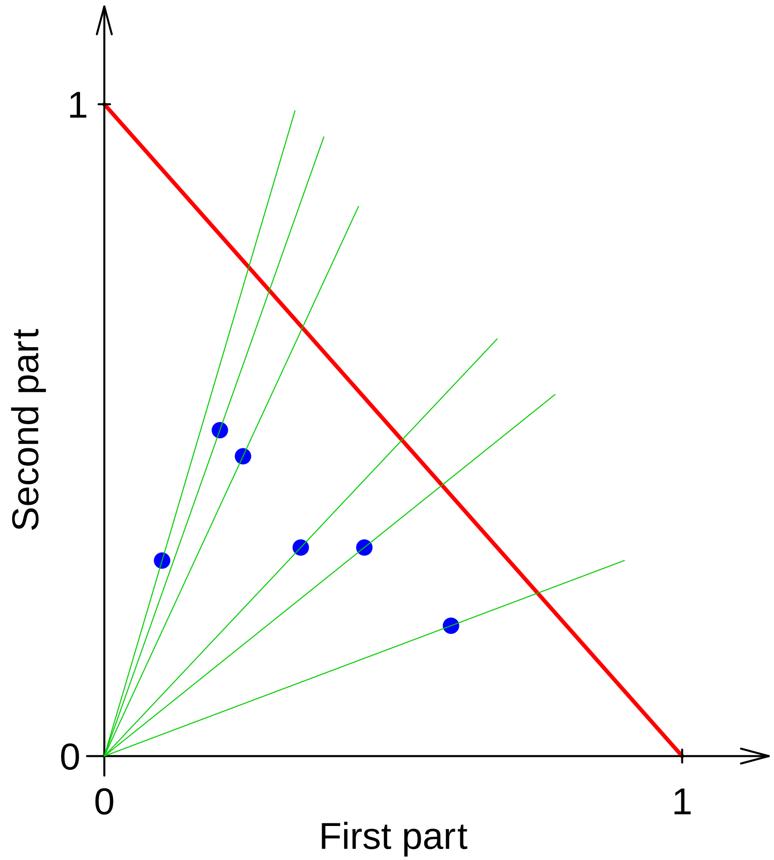


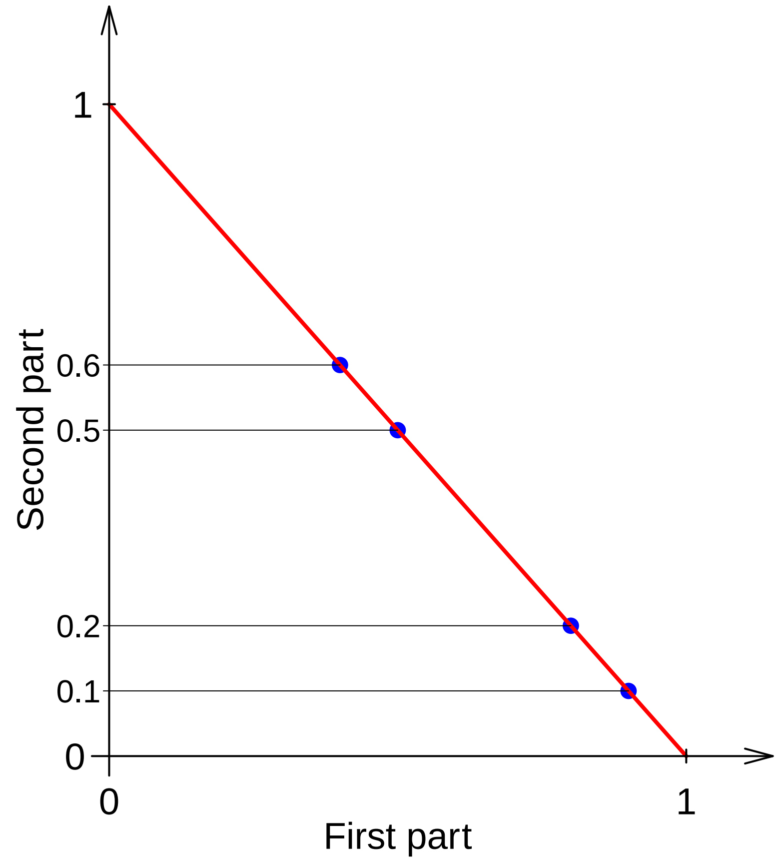


Normalization not necessary: same result with original data in HK$

Geometrical properties

Compositional data with only 2 parts





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

Transformations

Special transformations from the simplex to the Euclidean space:

- **alr (*additive logratio*) transformation:**
  Divide values by the j-th part, $j \in \{1, \ldots, D\}$:

$$\mathbf{x}^{(j)} = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right)^t$$

- **clr (*centered logratio*) transformation:**
  Divide values by the **geometric mean**:

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

- **ilr (*isometric logratio*) transformation:**
  take an orthonormal basis in the clr-space $\Longrightarrow$ **difficult to interpret**

Factor analysis for compositional data

Given a D-dimensional random variable $\mathbf{y}$.

FA model: $\mathbf{y} = \mathbf{\Lambda f} + \mathbf{e}$

with

$\mathbf{\Lambda}$ . . . loadings matrix

$\mathbf{f}$ . . . "factors" of dimension $k < D$

$\mathbf{e}$ . . . error term

With the usual assumptions this results in

$$\text{Cov}(\mathbf{y}) = \mathbf{\Lambda}\mathbf{\Lambda}^t + \mathbf{\Psi}$$

with the diagonal matrix $\mathbf{\Psi} = \text{Cov}(\mathbf{e})$ (*uniquenesses*).

Factor analysis for compositional data

(Filzmoser, Hron, Reimann, Garret, 2009, *Comp. & Geosci.*)

For an interpretation, FA **must** be related to the **original variables!**

$\Longrightarrow$ ilr transformation $(\mathbf{z})$, covariance estimation $(\text{Cov}(\mathbf{z}))$,
back-transformation to the clr-space: $\text{Cov}(\mathbf{y}) = \mathbf{V}\text{Cov}(\mathbf{z})\mathbf{V}^t$

Next problem: $\text{Cov}(\mathbf{y})$ is singular, which is in conflict with

$$\text{Cov}(\mathbf{y}) = \mathbf{\Lambda}\mathbf{\Lambda}^t + \mathbf{\Psi}$$

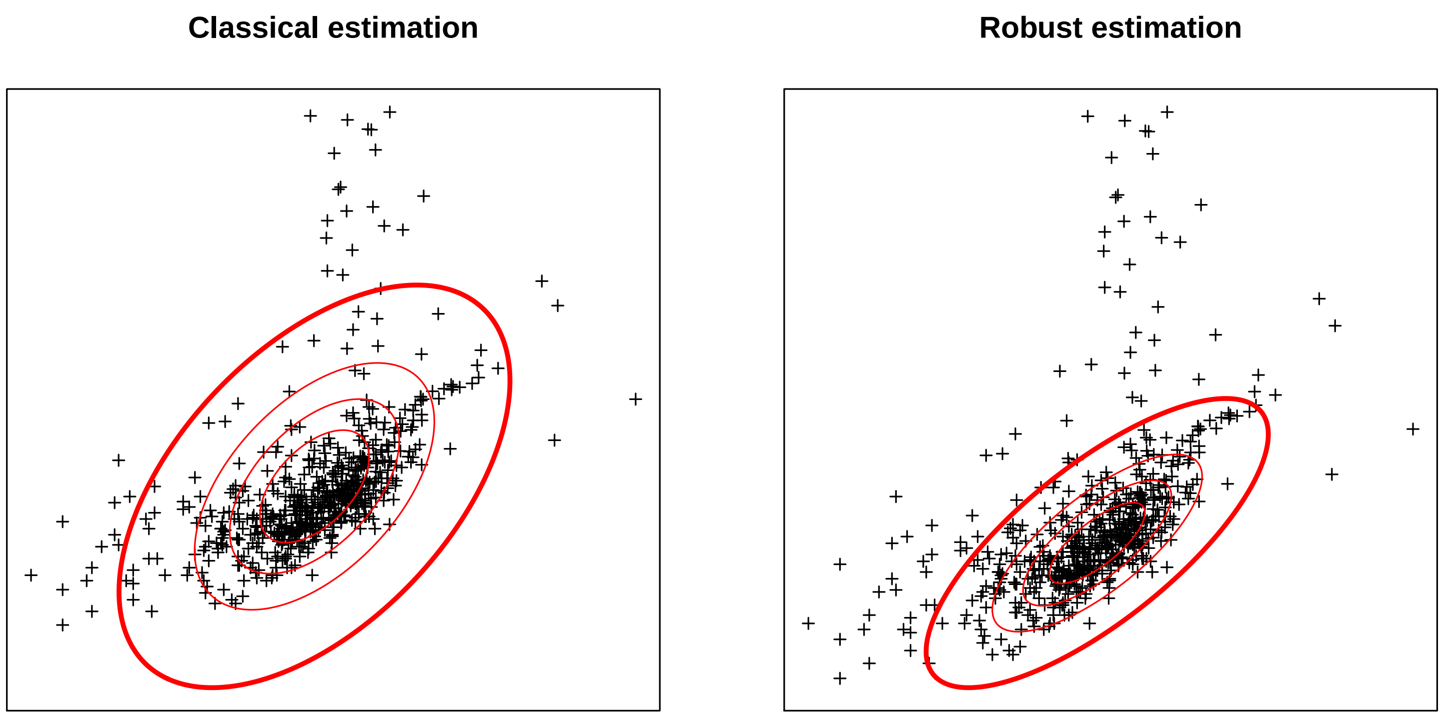
with a diagonal form of $\mathbf{\Psi}$.

Solution: Projection of the diagonal matrix $\mathbf{\Psi}$ on the hyperplane
$y_1 + \ldots + y_D = 0$ formed by the clr-space.
$\Longrightarrow$ resulting $\mathbf{\Psi}^*$ is no longer a diagonal matrix

Robust parameter estimation

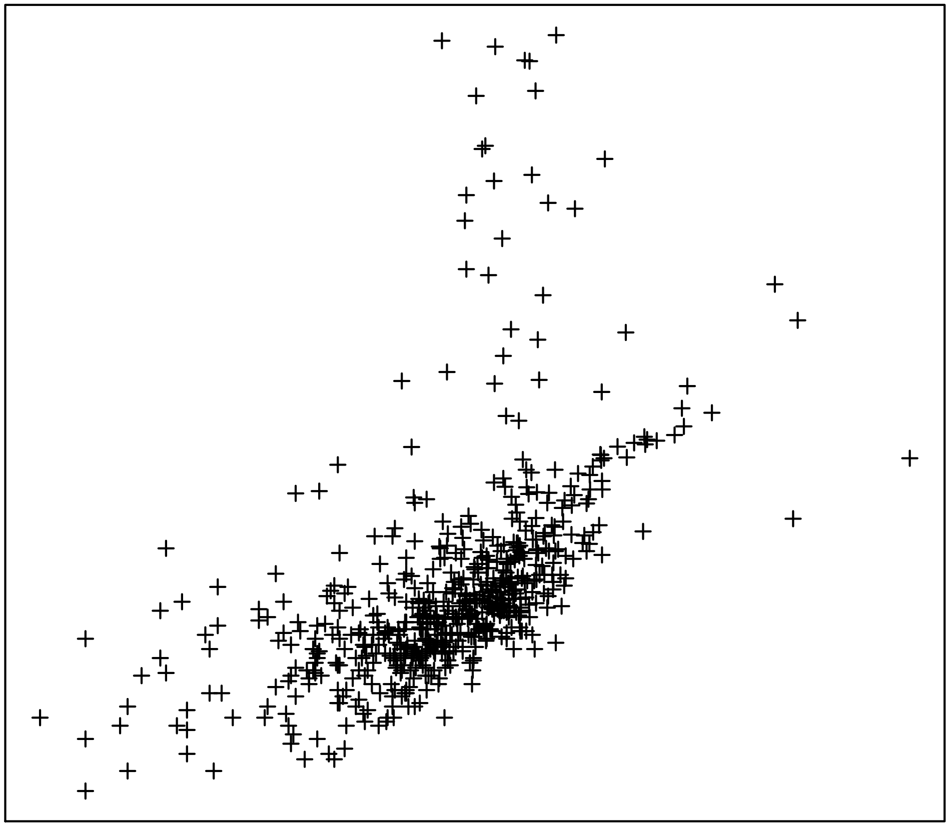
The basis for parameter estimation in the FA model is the estimation of the **covariance matrix**. The classical estimation is **sensitive with respect to outliers**.

$\Longrightarrow$ robust estimation of the covariance matrix leads to robust estimation of the parameters for FA (Pison, Rousseeuw, Filzmoser, Croux, 2003, *J. Multiv. Anal.*)



Robust covariance estimation

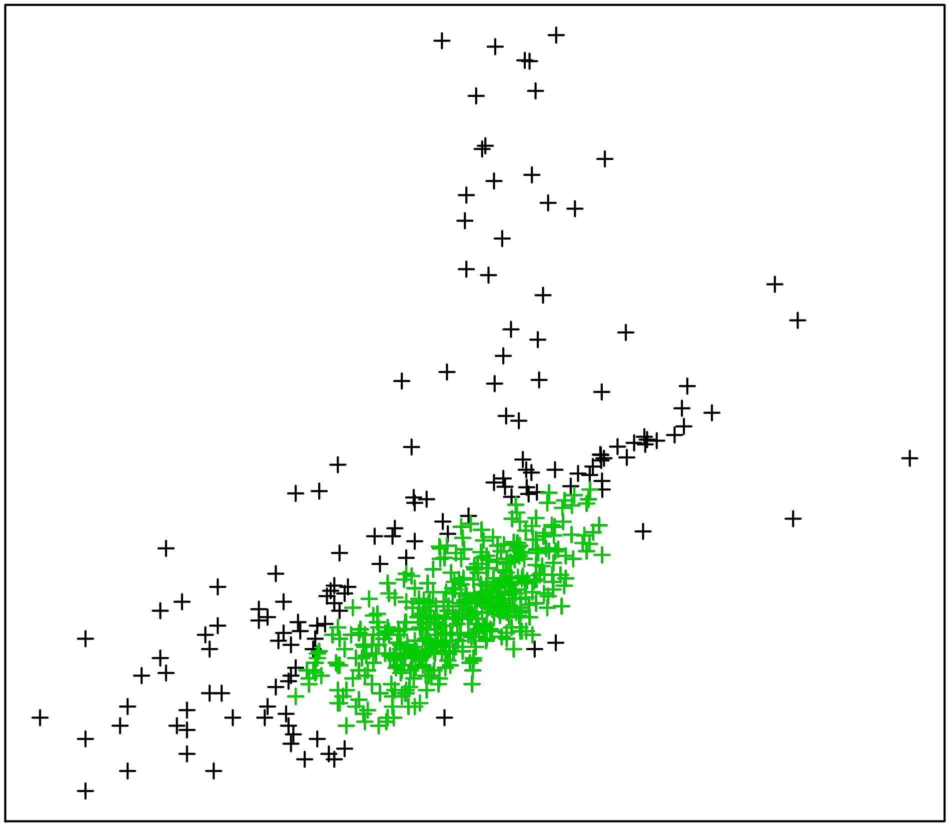
Minimum Covariance Determinant estimator (MCD):



Robust covariance estimation

Minimum Covariance Determinant estimator (MCD):

Search those 75% of data points having the smallest determinant of their classical covariance matrix



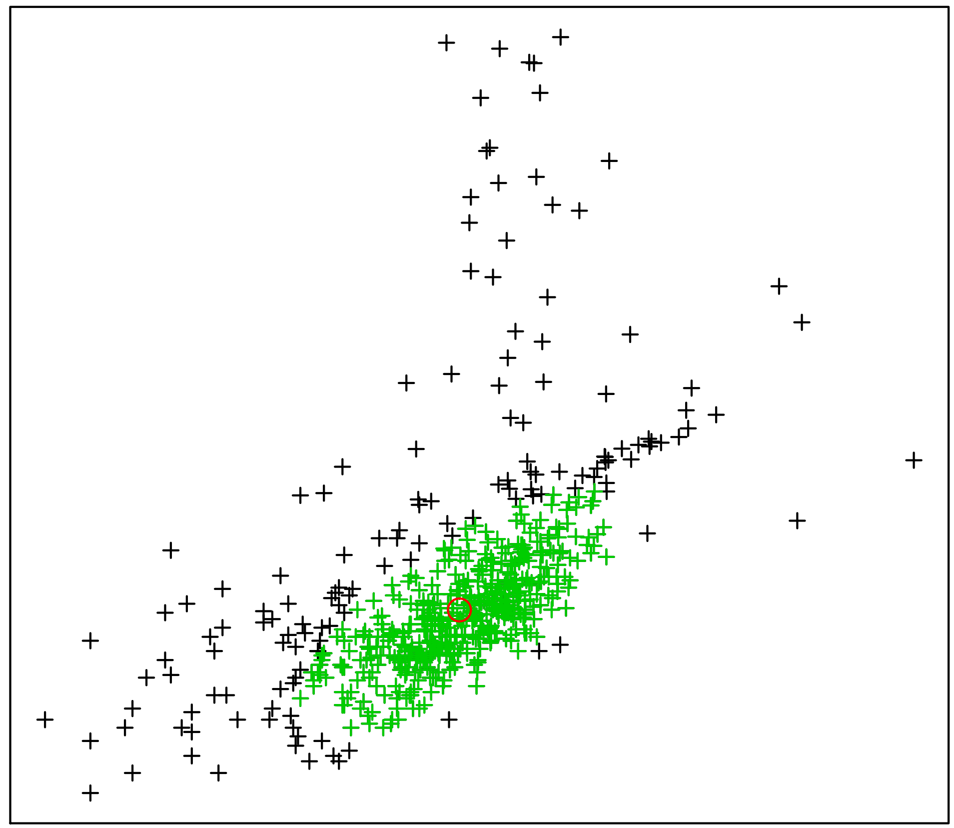
Robust covariance estimation

Minimum Covariance Determinant estimator

(MCD):

Search those 75% of data points having the smallest determinant of their classical covariance matrix

⟶ **Arithm. mean** is robust estimator of location



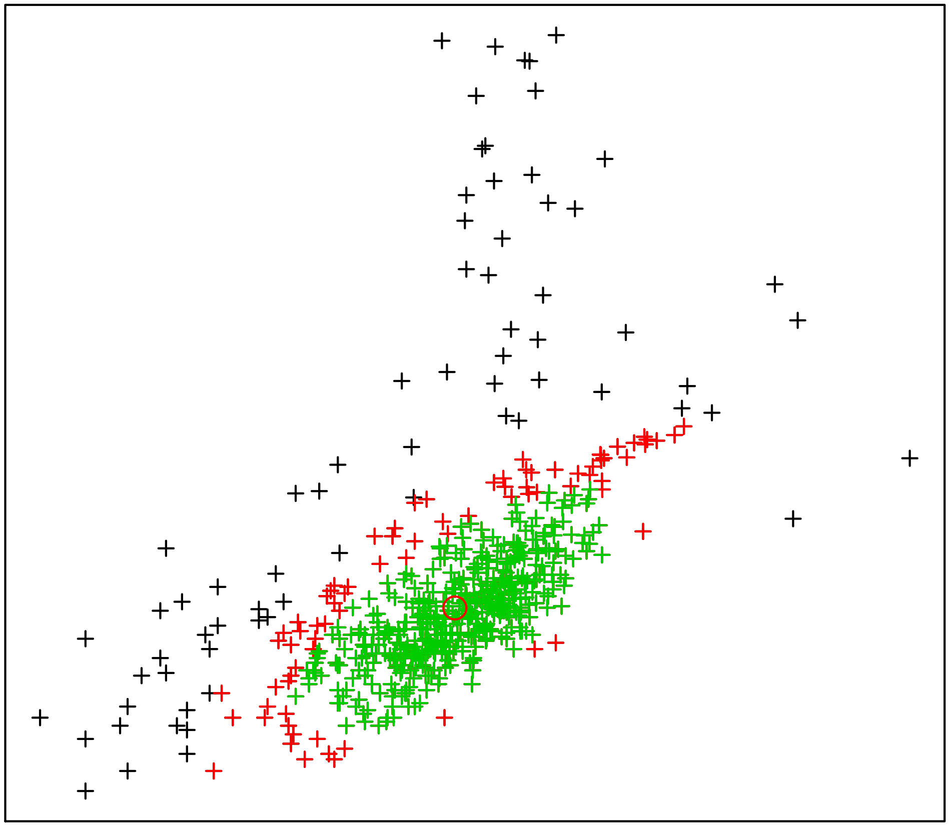
Robust covariance estimation

Minimum Covariance Determinant estimator (MCD):

Search those 75% of data points having the smallest determinant of their classical covariance matrix

$\longrightarrow$ **Arithm. mean** is robust estimator of location

$\longrightarrow$ **classical covariance**, multiplied by a factor, is robust covariance estimator



Robust FA for compositional data

Kola moss data:

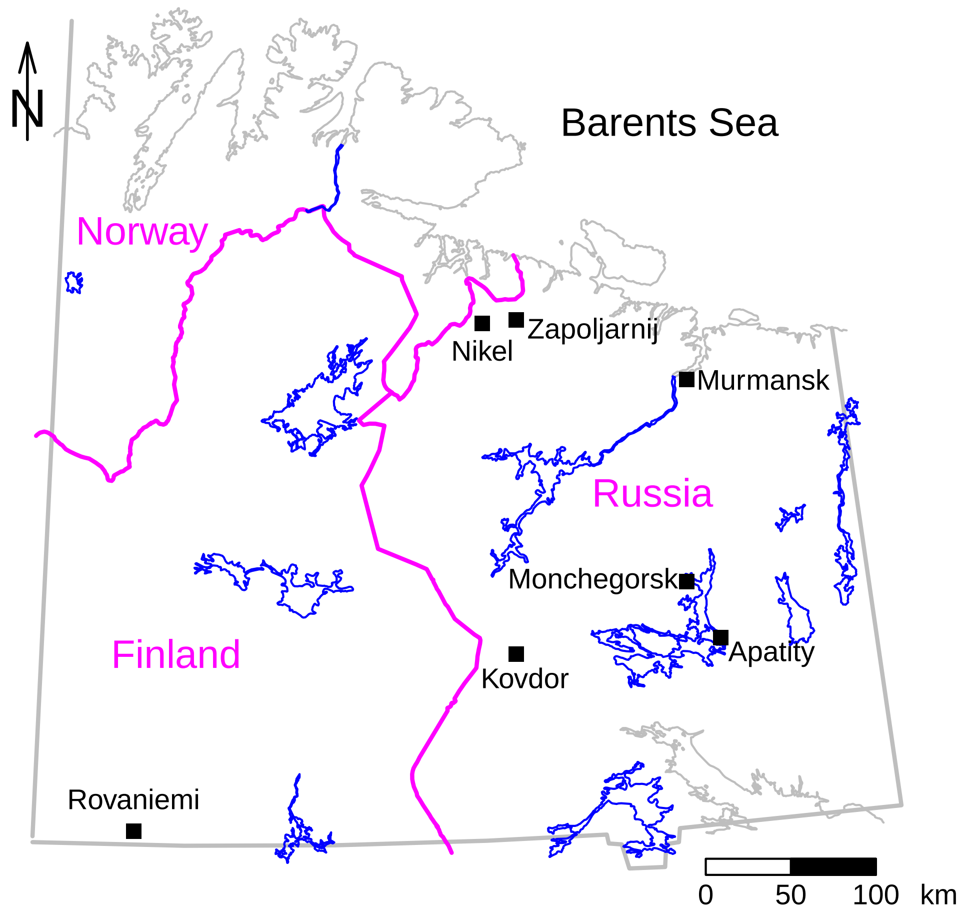
```
library(StatDA)
data(moss)
```

594 samples

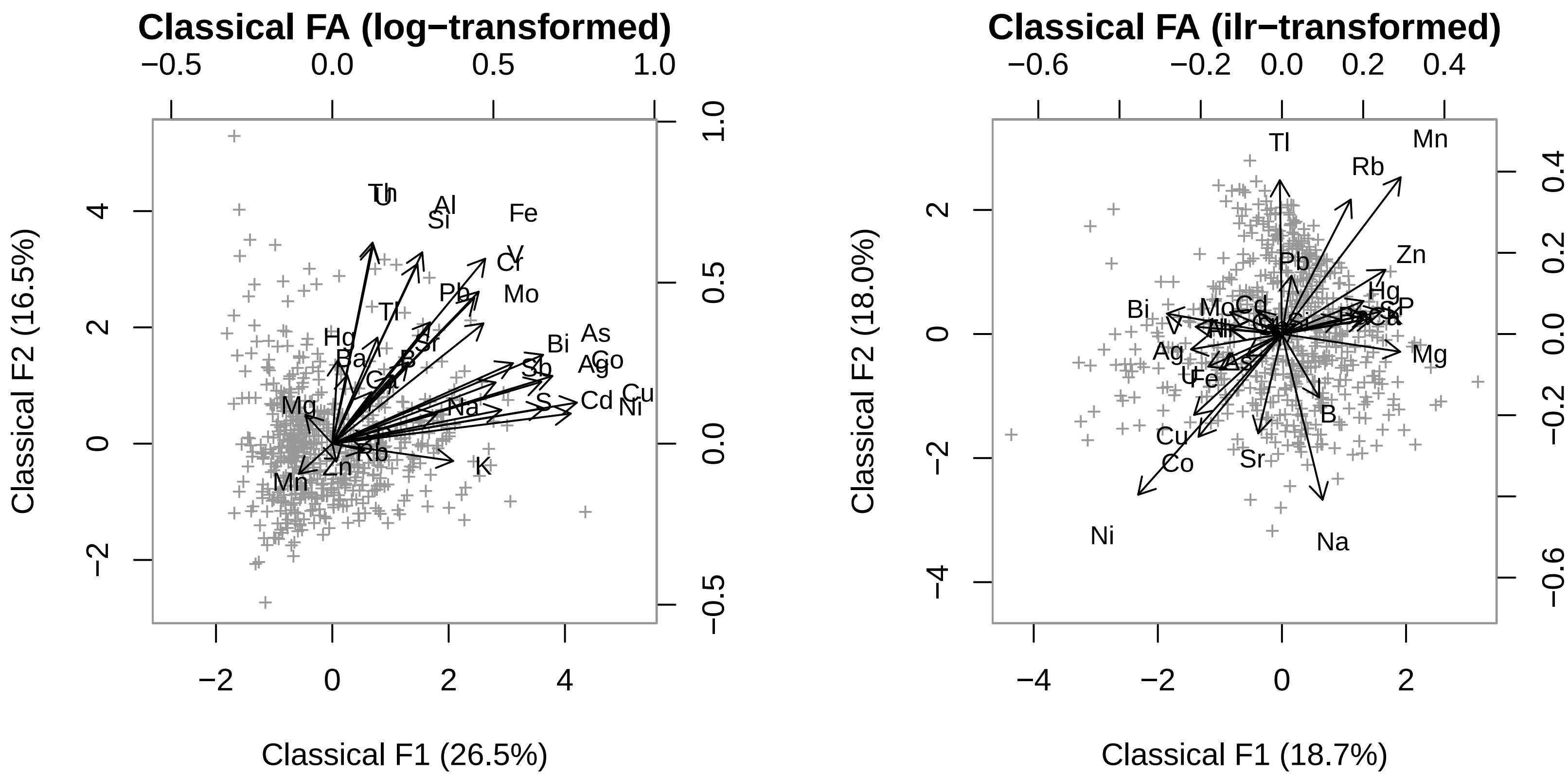
31 variables

Compare:

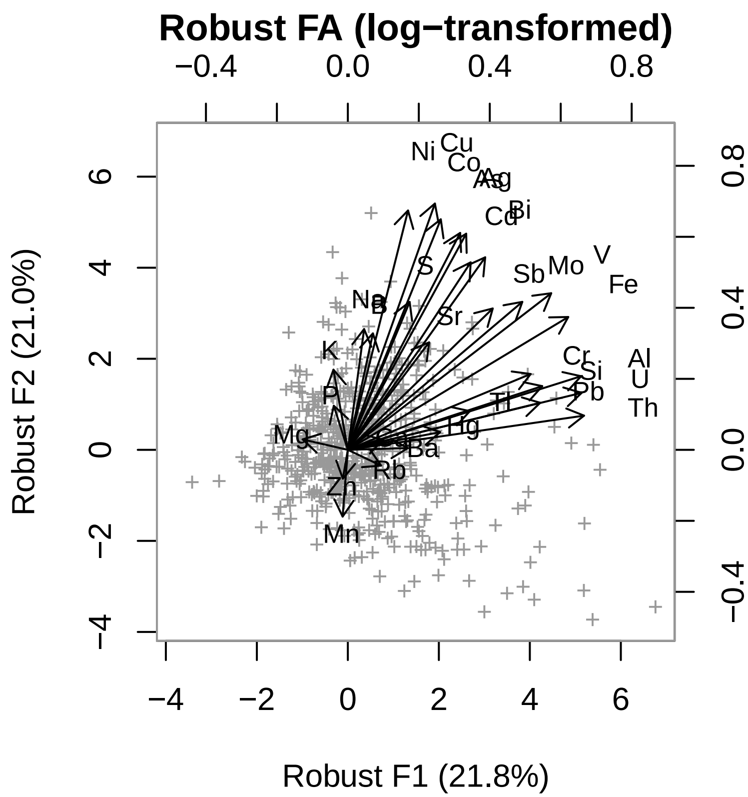
- classical and robust FA for
- log-transformed and ilr-transformed data

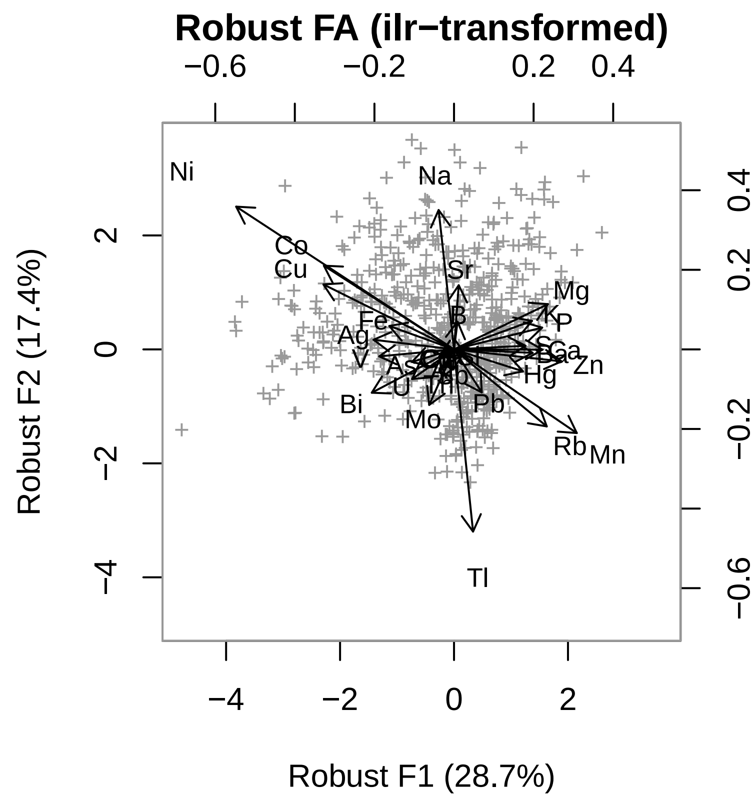


Example


Classical FA (log−transformed)
Classical F1 (26.5%)
Classical F2 (16.5%)
Classical FA (ilr−transformed)
Classical F1 (18.7%)
Classical F2 (18.0%)

Example





Relations between variables would indicate a dominance of industrial contamination.

Interesting processes:
sea spray, relations to plant nutrients, contamination.

Summary

- Compositional data are NOT characterized by a constant sum constraint. Rather, the compositional nature is an inherent data property.

- The sample space of compositional data is the simplex. For applying methods developed for the Euclidean geometry, the data first have to be transformed to the Euclidean space (ilr).

- Robust statistical methods cannot "repair" an incorrect geometrical representation of the data.

- Software is available; e.g. in the R packages `compositions`, `robCompositions`

Further work on this issue

M. Templ, P. Filzmoser, and C. Reimann (2008). Cluster analysis applied to regional geochemical data: Problems and possibilities. *Applied Geochemistry*. 23(8):2198-2213.

P. Filzmoser, K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3):233-248.

P. Filzmoser and K. Hron (2009). Correlation analysis for compositional data. *Mathematical Geosciences*, 41:905-919.

P. Filzmoser, K. Hron, and C. Reimann (2009). Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment*, 407:6100-6108.

P. Filzmoser, K. Hron, and C. Reimann (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, 20:621-632.

P. Filzmoser, K. Hron, C. Reimann, and R.G. Garrett (2009). Robust factor analysis for compositional data. *Computers and Geosciences*, 35:1854-1861.

K. Hron, M. Templ, P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*. To appear.

P. Filzmoser, K. Hron, and M. Templ (20??). Discriminant analysis for compositional data and robust parameter estimation. Under review.

Important references

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

A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (2006), editors, *Compositional data analysis in the geosciences: From theory to practice*. Geological Society, London.

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

. . . and many more . . .