Compositional data analysis

of our favourite drinks

Matthias Templ

Institute of Data Analysis and Process Design
School of Engineering
Zurich University of Applied Sciences

eRum Budapest, May 15, 2018

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

Outline

Topics of the presentation:

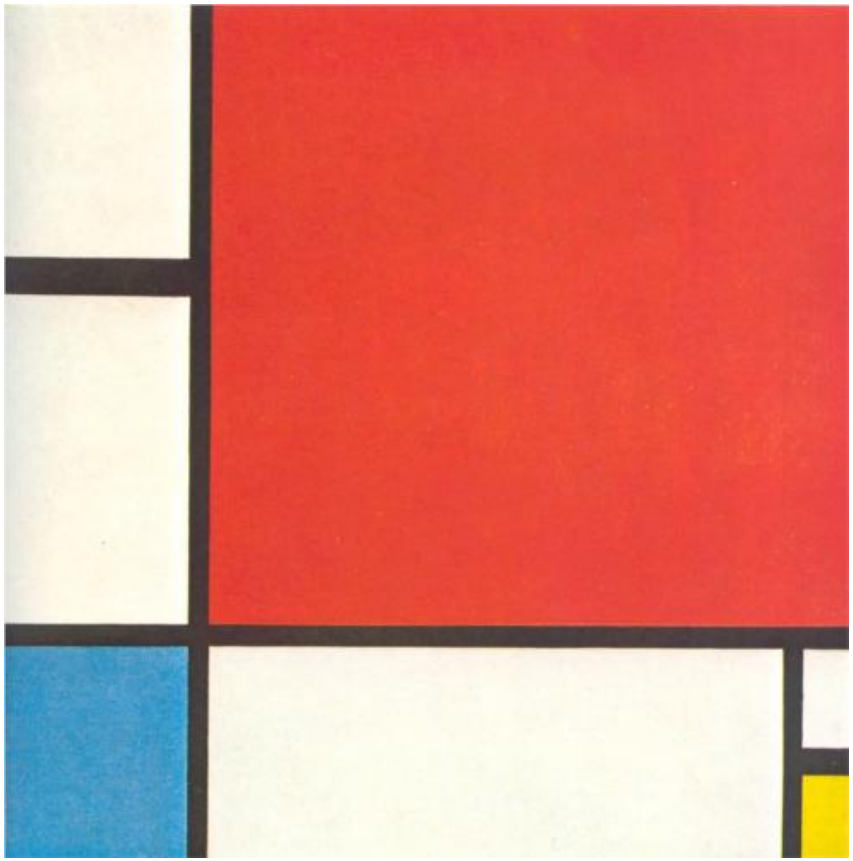
1. Are some of YOUR data compositional?
2. Why to work in *coordinates* and not with the raw data sets
3. Application to our favorite drinks data: orange juice, tea and milk



Compostitional Data

. . . those vectors representing **parts of a whole** which carry relative information

Absolute values are not of particular interest, but **ratios** only



Compositional data . . .

. . . are almost everywhere

Examples:

- expenditure data, tax components, wage components, . . .
- geochemical data
- percentages and ratios of a whole
- . . .

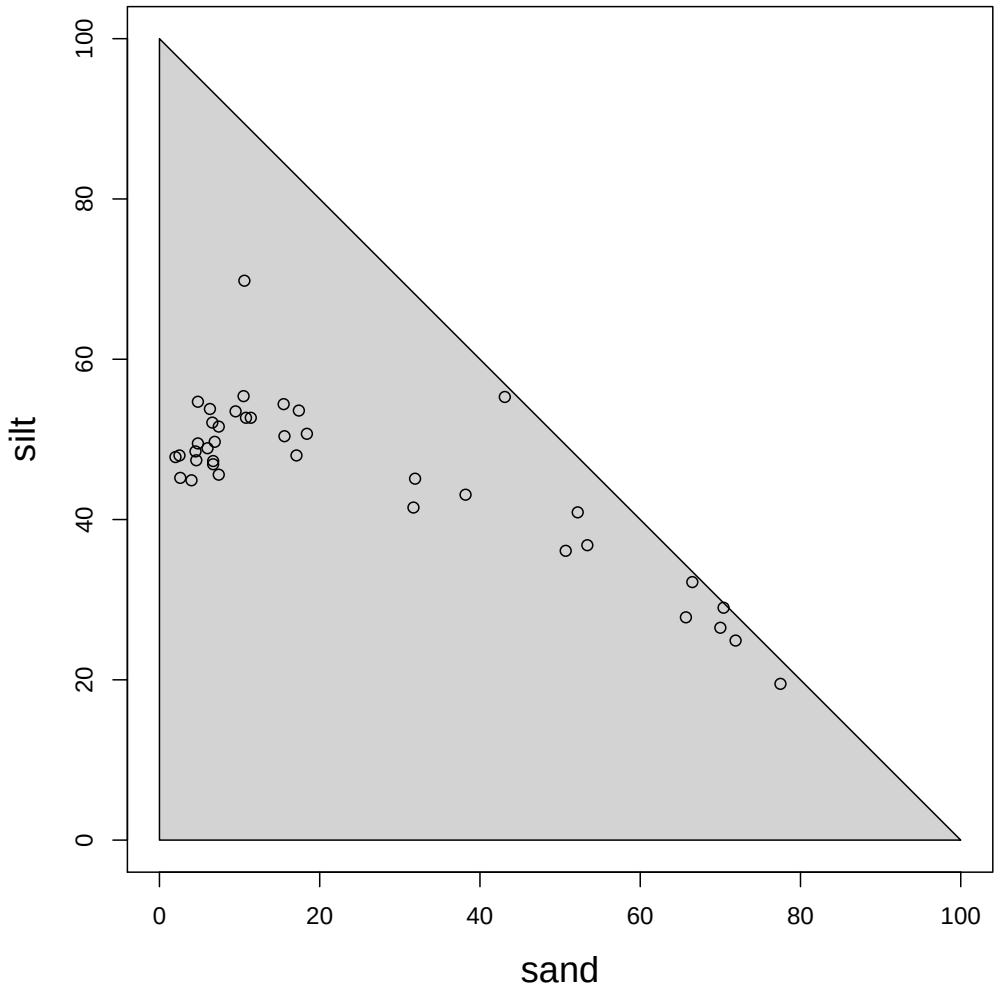
Important facts:

- compositional data must not be restricted to 100% or 1 or any similar constant! (think on expeditures of households)
- do not apply classical statistical methods to these data sets directly. But why? $\longrightarrow$

Compositional Data: Example

```
library("robCompositions")
data("arcticLake")
head(arcticLake)
```

```
##    sand silt clay
## 1 77.5 19.5  3.0
## 2 71.9 24.9  3.2
## 3 50.7 36.1 13.2
## 4 52.2 40.9  6.6
## 5 70.0 26.5  3.5
## 6 66.5 32.2  1.3
```



Something wrong with the data?

- Clearly, something is wrong - results from classical statistics would differ depending if data are used raw, as log's, as percentages or when subcompositions are used. (we can even show that results from raw, log's and percentages are simple wrong)
- There is a special relationship/correlation between variables, if we raise one value, automatically the others must decrease.
- If we do a compositional analysis, results would be always the same and not in contradiction to each other

Let's make a huge jump to a concise methodology

We show only one (out of many) specialized transformation to represent the data in coordinates: *pivot coordinates*.

Pivot coordinates

$$ilr(\mathbf{x}) = (z_1, \ldots, z_{D-1})^t, \quad z_j = -\sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}{x_j} \quad ,$$

$$\text{mit } j = 1, \ldots, D-1.$$

- This guarantees that the values in x_1 does not influence $z_2, \ldots, z_{D-1}$.
- eg 3-part composition:

$$z_1 = -\sqrt{\frac{2}{3}} \ln \frac{\sqrt[2]{x_2 x_3}}{x_1} \quad , \quad z_2 = -\sqrt{\frac{1}{2}} \ln \frac{x_2}{x_3}$$

Coordinate representation

The easy task:

- apply an (log-ratio) transformation first, then employ the standard statistical methodology.

The difficult parts:

- Apply an appropriate, data-specific (log-ratio) transformations (leading to specific sequential binary partitions)
- Correct interpretation of results, thus understanding the transformation and its consequences.

Coda analysis of our favorite drinks

- In the following we apply a few statistical methods on orange juice, tea and milk data.
- I assume knowledge on principal component analysis and correlation
- We only need to define we understand about a correlation measures in the coda context, because correlation measures on compositional data are arbitrary.

Variation matrix

We consider variances of all pairwise logratios. For $\mathbf{X} = (x_{ij})$, the variation matrix is defined as

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1D} \\ t_{21} & t_{22} & \dots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \dots & t_{DD} \end{pmatrix},$$

where $t_{jk},\ j, k = 1, \dots, D$, are sample variances of pairwise logratios between x_j and x_k, i.e.

$$t_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{jk}^i - \overline{z}_{jk})^2$$

with

$$\{z_{jk}^i = \ln \frac{x_{ij}}{x_{ik}},\ i = 1, \dots, n\} \text{ and } \overline{z}_{jk} = \frac{1}{n} \sum_{i=1}^{n} z_{jk}^i.$$

Compositions of beers

)

```
require("Biobase") # not beerbase !
openPDF("beerageing.pdf"); openPDF("poster.pdf")
```

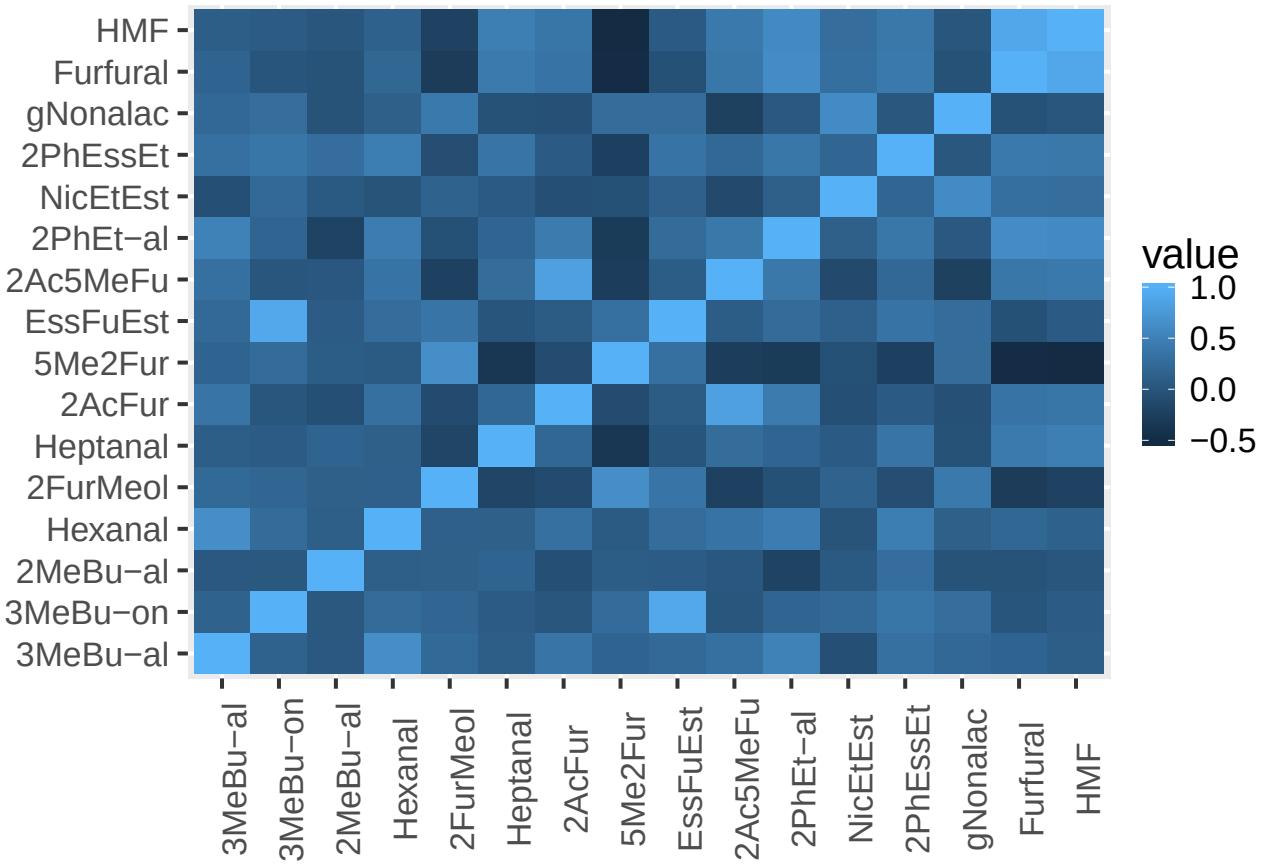
Compositions of beers

```
load("beer.RData")
str(beer)
```

```
## 'data.frame':    86 obs. of  19 variables:
##  $ Betrieb : Factor w/ 45 levels "101RO-M","102RO-C",..
##  $ BetrNr  : Factor w/ 10 levels "1","2","3","4",..: 1 1
##  $ newold  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1
##  $ 3MeBu-al: num  23.4 16.3 101.4 18.3 152.9 ...
##  $ 3MeBu-on: num  10.3 49.5 29.5 7.4 45.4 4 9.1 24.7 6.7
##  $ 2MeBu-al: num  44.4 99.8 223.9 492.2 142.3 ...
##  $ Hexanal : num  180 167 200 137 296 ...
##  $ 2FurMeol: num  3287 1555 2228 1895 3934 ...
##  $ Heptanal: num  4.2 3.9 4.2 4 8 3.8 4.3 4.9 4 4.4 ...
##  $ 2AcFur  : num  40.7 32.2 41.7 29.6 43.3 21.2 29.9 16.
##  $ 5Me2Fur : num  45 41.1 51.9 34 54.6 29.1 34.6 26.7 32
##  $ EssFuEst: num  33.5 39.9 38.8 15.9 49.5 10.5 21 33.2
##  $ 2Ac5MeFu: num  26 14.8 27.8 15.6 8 13.9 24.1 4.5 7.1
```

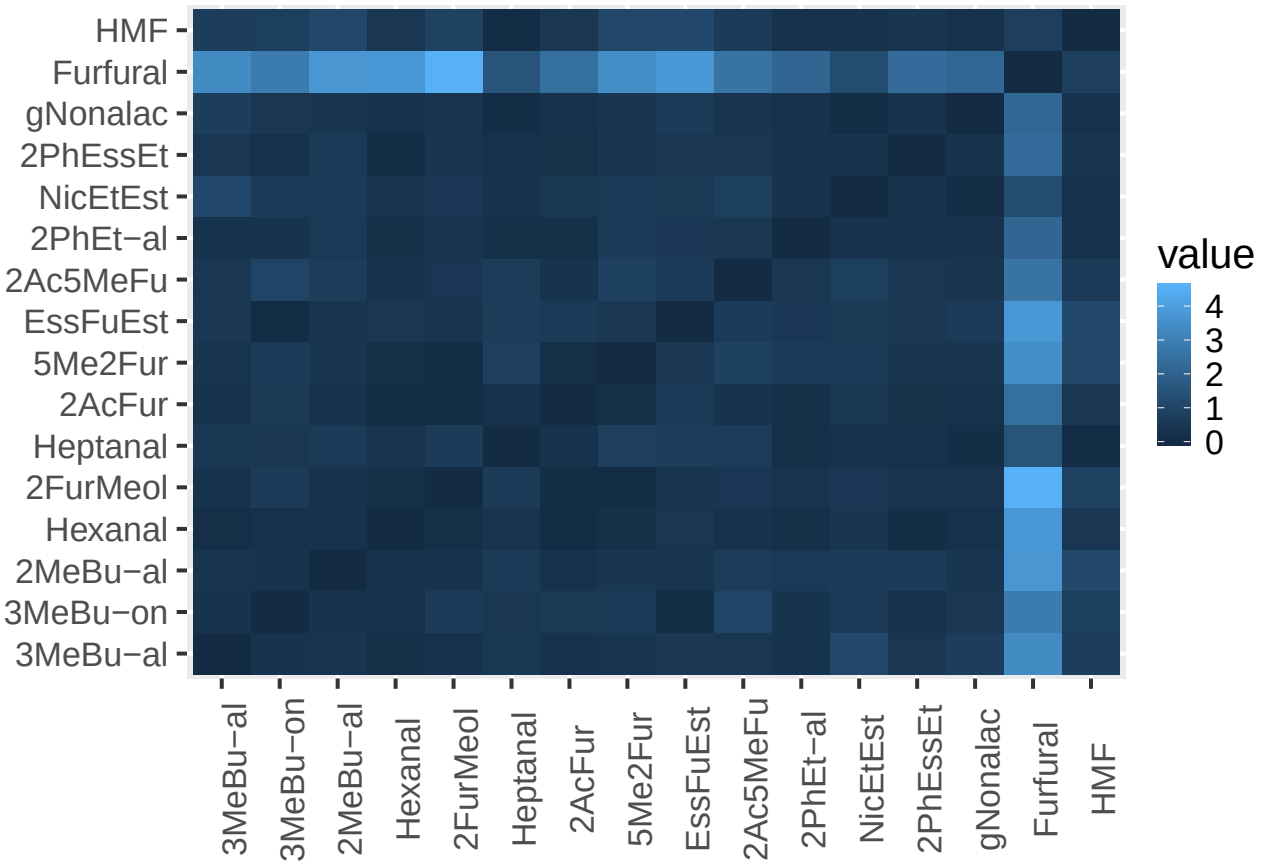
Compositions of beers - dependencies between variables

```
library("robCompositions"); library("reshape2")
## wrong way to do it:
co <- cor(beer[, 4:ncol(beer)])
ggplot(data = melt(co), aes(x = ... -->
```



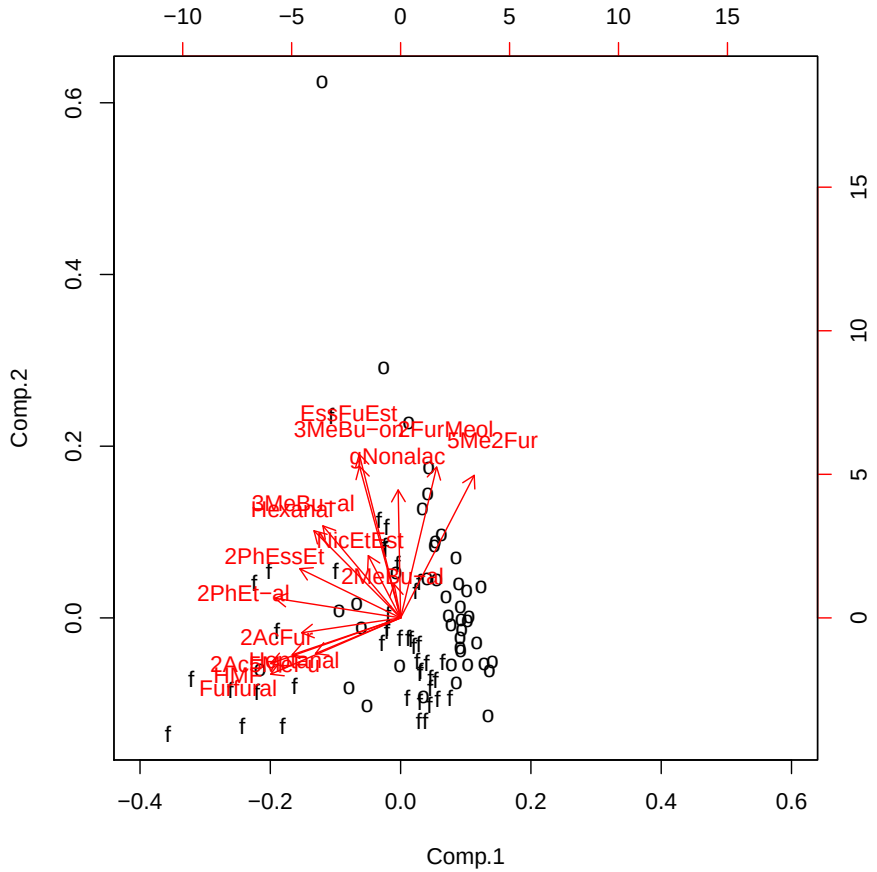
Compositions of beers - dependencies between variables

```
## compositional analysis
co_coda <- robCompositions::variation(beer[,4:ncol(beer)])
ggplot(data = melt(co_coda), aes(x = ... -->
```



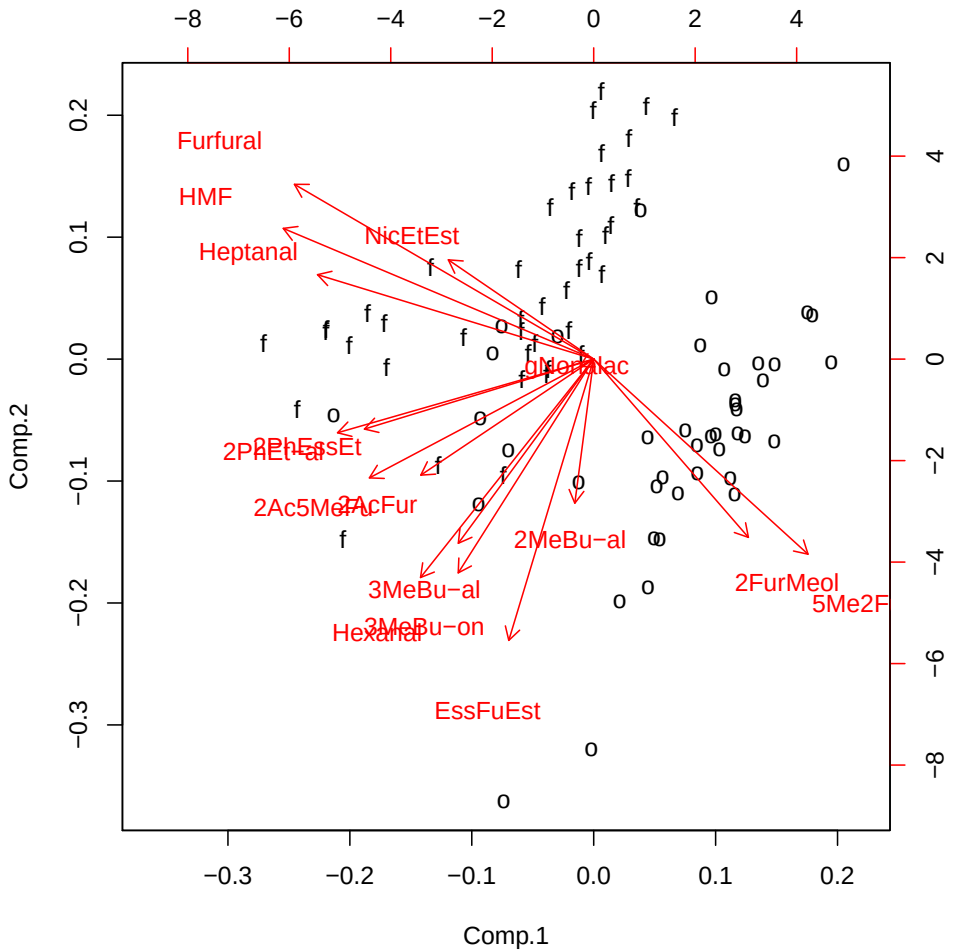
Compositions of beers - PCA on raw data

```
### PCA: this is the wrong approach:
newold <- ifelse(beer[, 3] == 2, "f","o")
biplot(princomp(beer[, 4:ncol(beer)], cor = TRUE),
       xlabs = newold, xlim = c(-0.4, 0.6))
```



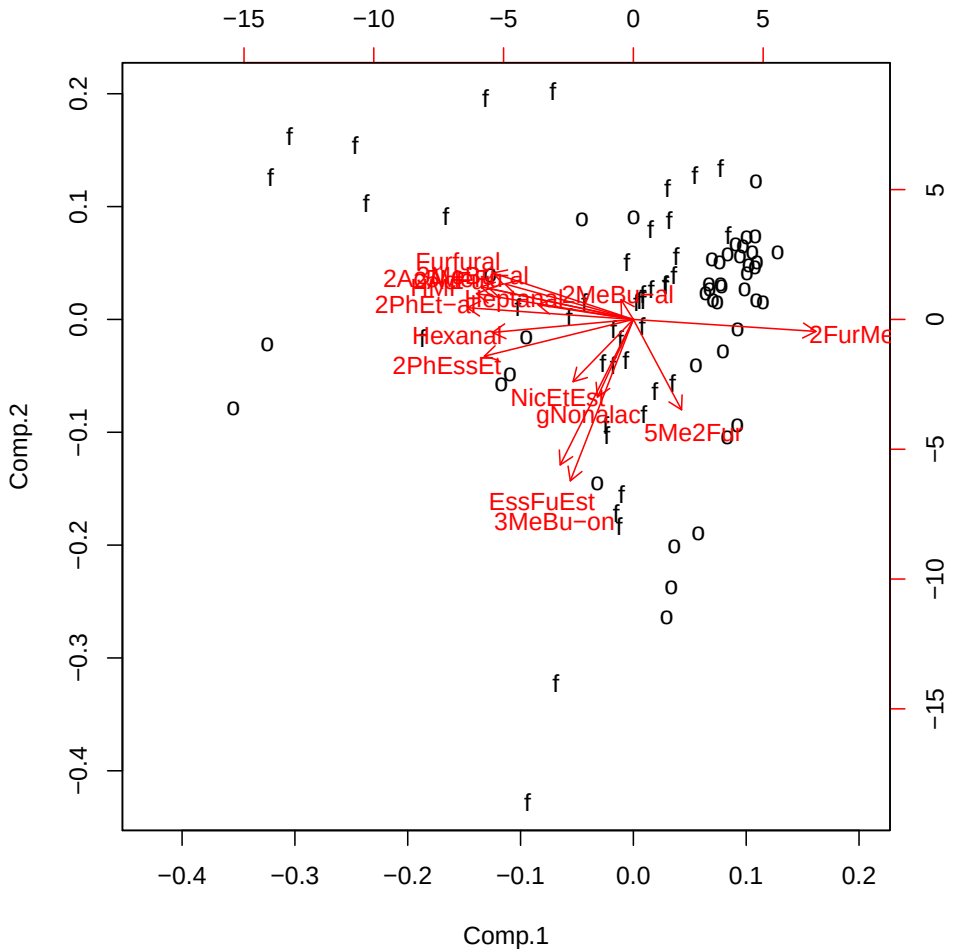
Compositions of beers - PCA on log-scaled data

```
### do it wrongly (2):
biplot(prl <- princomp(log(beer[, 4:ncol(beer)]),
                       cor = TRUE), xlabs = newold)
```



Compositions of beers - PCA on proportional data

```
### do it wrongly (3):
biplot(princomp(constSum(beer[, 4:ncol(beer)]),
                cor=TRUE), xlabs = newold)
```



Compositions of beers - PCA

So far we saw

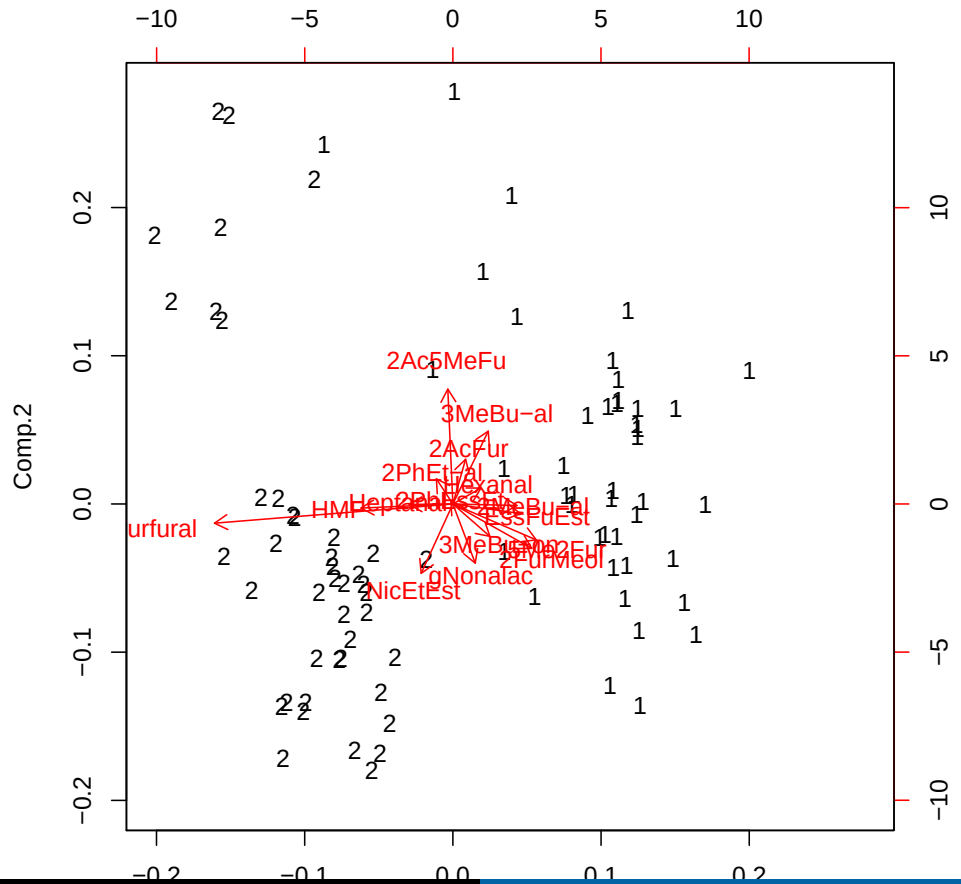
- all results look distorted
- also robust PCA will not repair this
- the results led to wrong conclusions in the original paper

The reason is that we worked in the Euclidean geometry which is not appropriate for compositional data.

$\rightarrow$ let's work in coordinates from now on

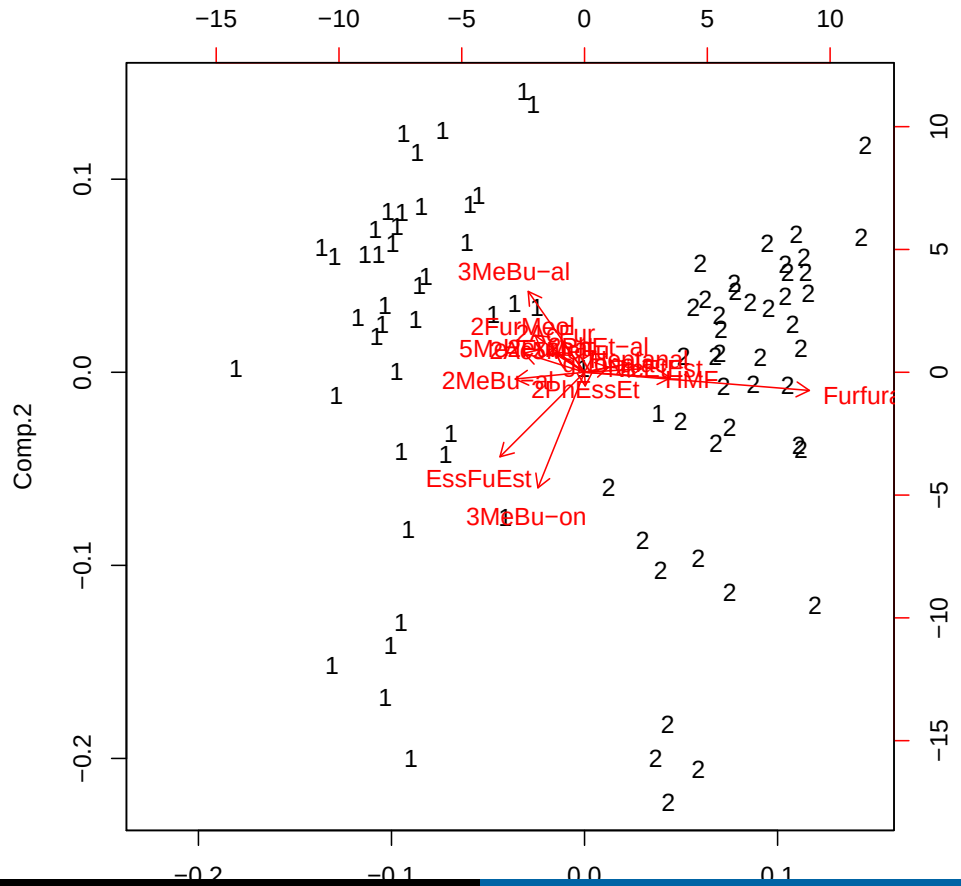
Compositions of beers - PCA compositional

```
# pivot coordianates +
# projection to centred log-ratio coordinates.
biplot(pcaCoDa(beer[, 4:ncol(beer)], method =
  "standard")$prin,  xlabs = as.numeric(beer$newold))
```

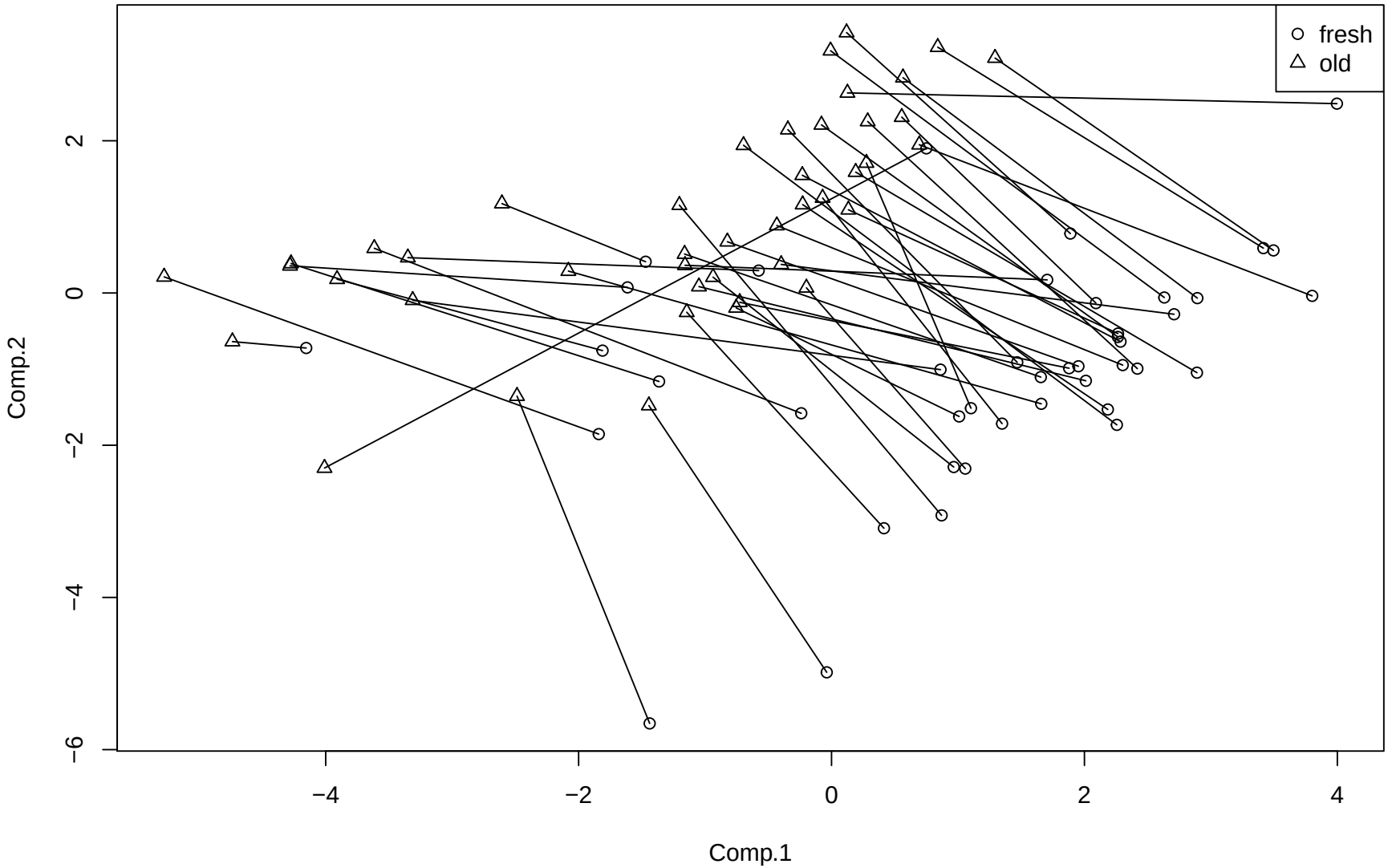


Compositions of beers - PCA compositional robust

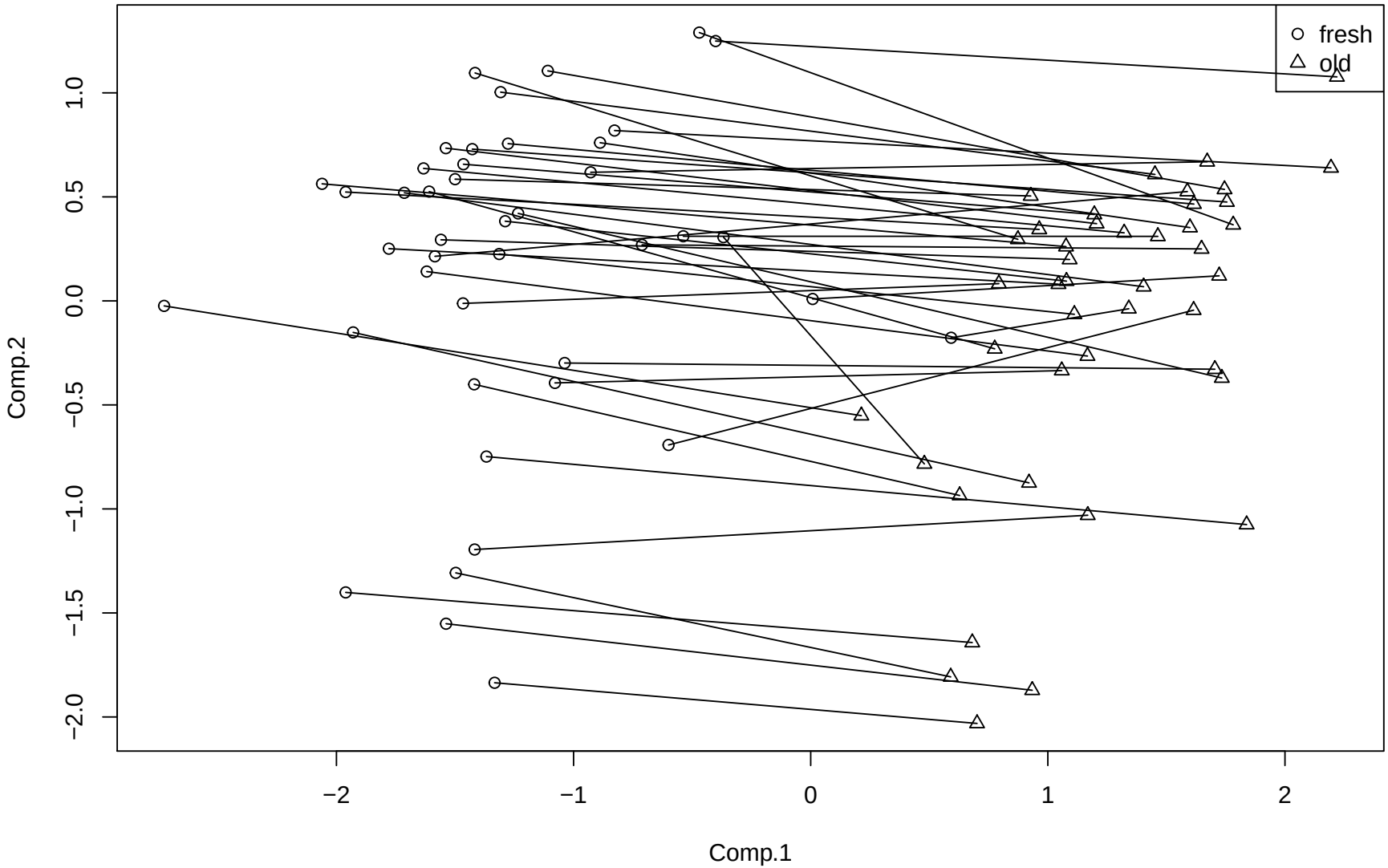
```
# pivot coordianates +
# projection to centred log-ratio coordinates.
biplot(pcaCoDa(beer[, 4:ncol(beer)], method =
  "robust")$prin,  xlabs = as.numeric(beer$newold))
```



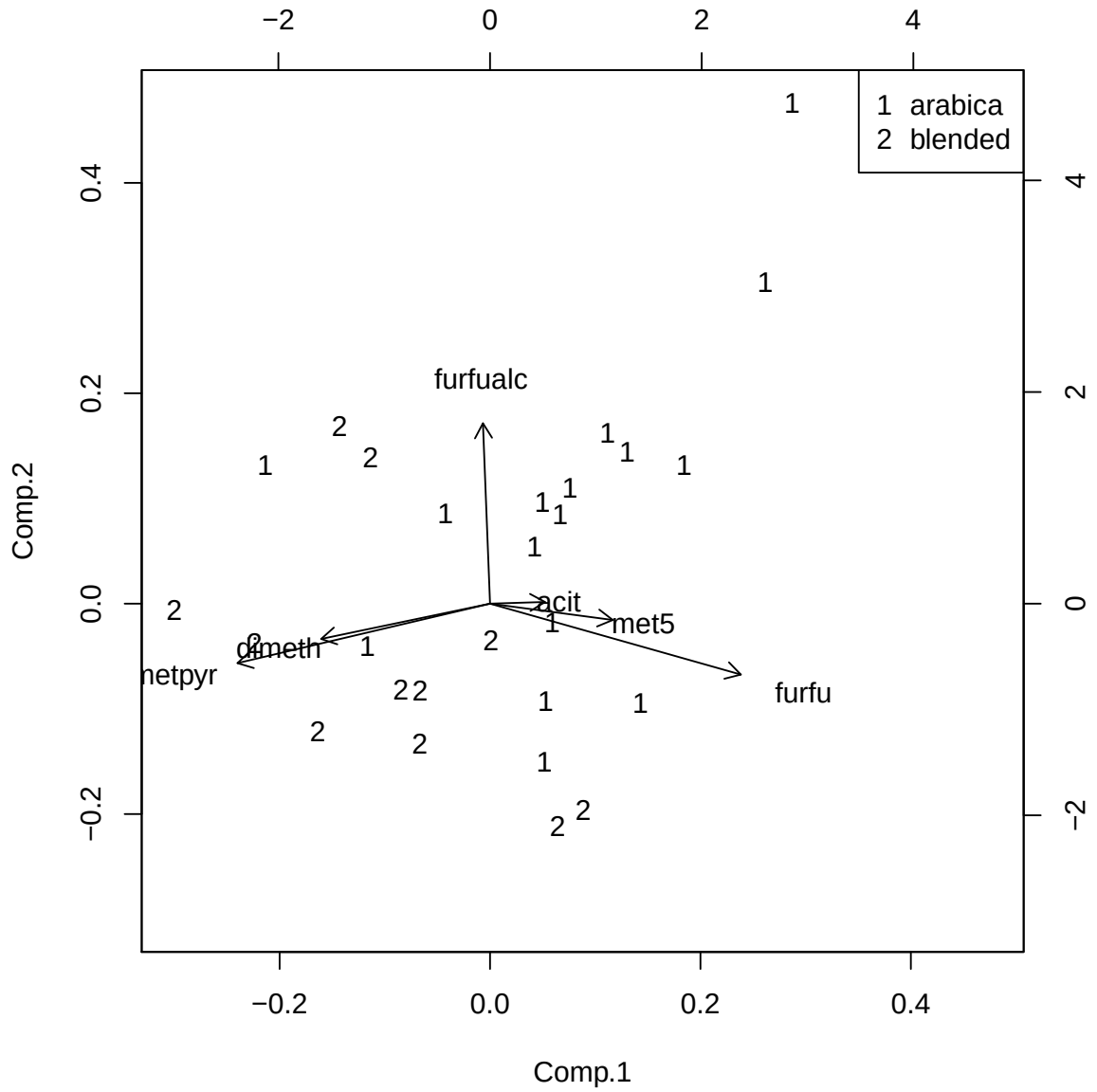
Compositions of beers - PCA on raw data, separation



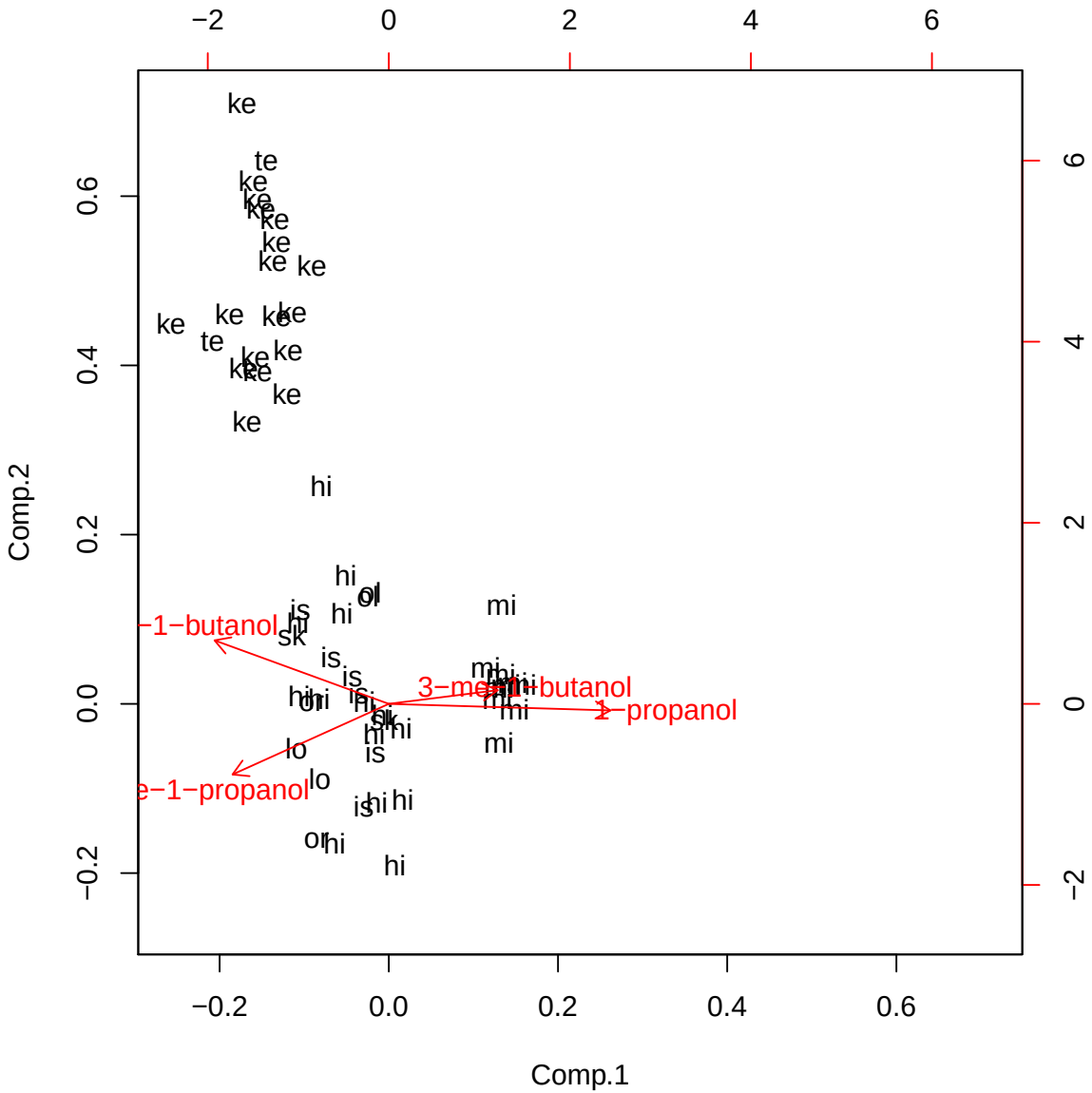
Compositions of beers - PCA compositional, separation



Compositions of coffee



Compositions of whisky's



Software and Book

- Springer book by Filzmoser, Hron, Templ (2018): **Applied Compositional Data Analysis** is available very soon
  - five chapters explaining the concept of coda
  - cluster analysis, pca, discriminant analysis, regression analysis, high-dimensional data, tables, outliers, missing values, zeros.
- Package **robCompositions** includes all these (robustified) methods from the book

Conclusion

Main conclusions for the talk - this should be taken serios at the eRum conference:

- **don't let the beer get warm** !

And

- Arabica coffee's may not as healthy than the blended ones
- there is quite a difference from which destination a Whisky comes from

Off topic: www.ajs.or.at is open-access and without fees

Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 45, Number 1, 2016

Special Issue on R

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT