Compositional Data Analysis with R

Matevž Bren[1] and Vladimir Batagelj[2]

[1]University of Maribor, Slovenia

[2]University of Ljubljana, Slovenia

CoDaWork

Girona, October 15 - 17, 2003

Outline

R a free statistical language and environment

R (`http://www.r-project.org/`) is a free language and environment for statistical computing and graphics. R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering...).

The term *environment* is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

… R a free statistical language and environment

It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The current version of the R library for compositional data analysis is available at **http://vlado.fmf.uni-lj.si/pub/mixture/**

Aitchison's Household budget survey

from the Aitchison's book *The Statistical Analysis of Compositional Data*:

Sample survey of single persons living alone in a 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.

H – housing, including fuel and light,
F – foodstuffs, including alcohol and tobacco,
O – other goods, including clothing, footwear. . . ,
S – services, including transport and vehicle.

We consider only the expenditure proportions, not the values – *compositional data.*

Aitchison's Household budget survey

	H	F	O	S
M1	497	591	153	291
M2	839	942	302	365
M3	798	1308	668	584
M4	892	842	287	395
M5	1585	781	2476	1740
M6	755	764	428	438
M7	388	655	153	233
M8	617	879	757	719
M9	248	438	22	65
M10	1641	440	6471	2063
M11	1180	1243	768	813
M12	619	684	99	204
M13	253	422	15	48
M14	661	739	71	188
M15	1981	869	1489	1032
M16	1746	746	2662	1594
M17	1865	915	5184	1767
M18	238	522	29	75
M19	1199	1095	261	344
M20	1524	964	1739	1410

	H	F	O	S
W1	820	114	183	154
W2	184	74	6	20
W3	921	66	1686	455
W4	488	80	103	115
W5	721	83	176	104
W6	614	55	441	193
W7	801	56	357	214
W8	396	59	61	80
W9	864	65	1618	352
W10	845	64	1935	414
W11	404	97	33	47
W12	781	47	1906	452
W13	457	103	136	108
W14	1029	71	244	189
W15	1047	90	653	298
W16	552	91	185	158
W17	718	104	583	304
W18	495	114	65	74
W19	382	77	230	147
W20	1090	59	313	177

The 'mixture' class in R

```
Household budget survey
        H       F       O       S
M1      497     591     153     291
M2      839     942     302     365
M3      798     1308    668     584
M4      892     842     287     395
M5      1585    781     2476    1740
M6      755     764     428     438
M7      388     655     153     233
M8      617     879     757     719
M9      248     438     22      65
M10     1641    440     6471    2063
M11     1180    1243    768     813
M12     619     684     99      204
M13     253     422     15      48
M14     661     739     71      188
M15     1981    869     1489    1032
M16     1746    746     2662    1594
M17     1865    915     5184    1767
M18     238     522     29      75
M19     1199    1095    261     344
M20     1524    964     1739    1410
W1      820     114     183     154
W2      184     74      6       20
W3      921     66      1686    455
W4      488     80      103     115
W5      721     83      176     104
W6      614     55      441     193
W7      801     56      357     214
W8      396     59      61      80
W9      864     65      1618    352
W10     845     64      1935    414
W11     404     97      33      47
W12     781     47      1906    452
W13     457     103     136     108
W14     1029    71      244     189
W15     1047    90      653     298
W16     552     91      185     158
W17     718     104     583     304
W18     495     114     65      74
W19     382     77      230     147
W20     1090    59      313     177
```

The *input mixture data* consist of a *data matrix* preceeded by a *title*. In R we represent them as a structure m (m`$tit`, m`$mat`, m`$sum`, m`$sta`)

m`$sum` is the *row sum* and

m`$sta` is a *status* with values:

- -2 – matrix contains negative elements
- -1 – zero sum row exists
- 0 – rows with different row sum(s)
- 1 – mixture with constant row sum
- 2 – normalized mixture

The 'mix' procedures in R

We started to develope a library **MixeR** of functions in R to support the analysis of mixtures.

```
mix.Read(file, eps=1e-6)
```

Reads a mixture data from the *file* and returns it as a mixture structure. If $|m\$sum - 1| < eps$ it sets $m\$sta = 2$.

```
mix.Check(m, eps=1e-6)
```

Determines the $m\$sum$ and $m\$sta$ of a given mixture structure m.

```
mix.Normalize(m)
```

Normalizes a given mixture structure m if $m\$sta \geq 0$.

```
mix.Random(nr, nc, s=1)
```

Generates a random mixture structure with nr rows, nc columns and row sum s.

... The 'mix' procedures in R

`mix.Matrix(a, t)`

Converts a matrix a with title t to a mixture structure.

`mix.Ternary(m, lcex=1, add=FALSE, ord=1:3, ...)`

Produces a ternary display of a given mixture structure m.

`mix.Sub(m, k)`

Returns a mixture structure obtained from m by extracting colomns from the list k.

`mix.Quad2Net(fnet, m)`

Transforms a 4 column mixture m quadrays into 3d XYZ coordinates and writes them as a **Pajek** file. **Pajek** is available at

http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Compositional data sample space

Compositions (compounds, mixtures, alloy ...,) can be represented with vectors of the portions of individual components. The portions are nonnegative and they have constant sum.

A suitable (one of) sample space for compositional data

$$\mathbf{w} = (w_1, \ldots, w_D), \quad w_k \geq 0, \, k = 1, \ldots, D,$$

$$w_1 + \cdots + w_D = \text{const.}$$

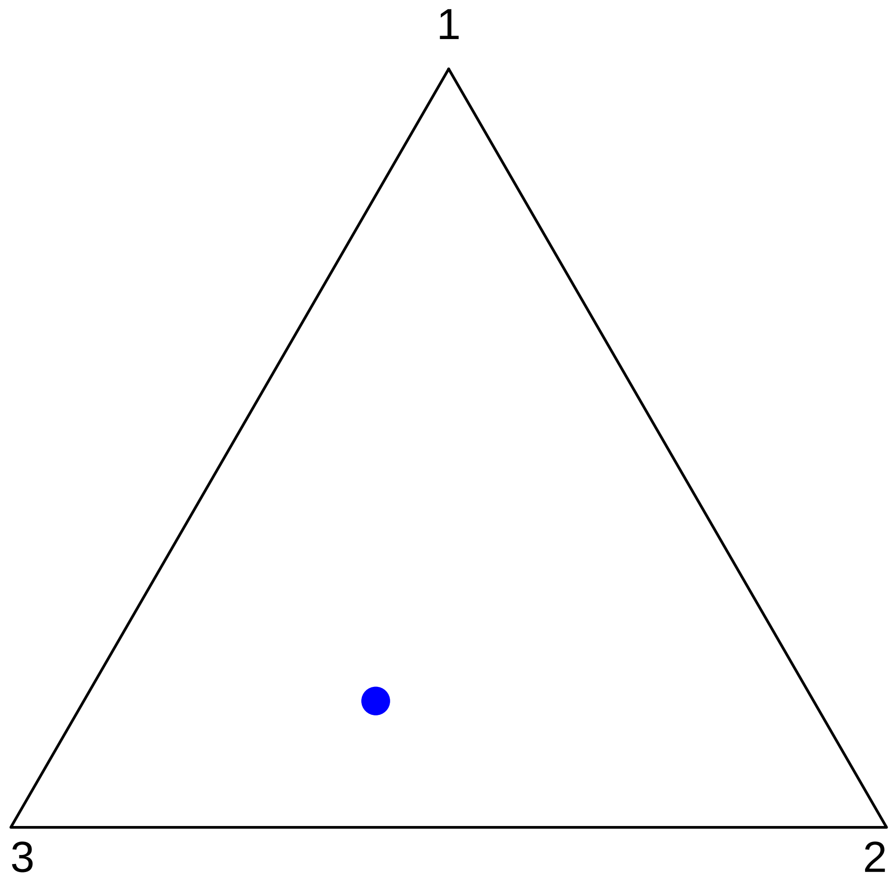
is the d - dimensional *unit simplex* $(d := D - 1)$

$$\mathcal{S}^d := \{\mathbf{x} = (x_1, \ldots, x_D); \, x_k > 0, \, k = 1, \ldots, D \wedge x_1 + \cdots + x_D = 1\}$$

Any vector of positive components $\mathbf{w} \in \mathbb{R}_+^D$ can be projected onto the simplex by the *closure operation*

$$\mathcal{C}(\mathbf{w}) = \left(\frac{w_1}{\sum w_k}, \ldots, \frac{w_D}{\sum w_k} \right) \in \mathcal{S}^d.$$

Ternary Diagram



Graphical representation of three part compositions $\mathbf{x} = (0.17, 0.33, 0.50)$.

Perturbations

The *perturbation operation*

$$\mathbf{x} \circ \mathbf{y} = \mathcal{C}\left(x_1\, y_1,\, \ldots,\, x_D\, y_D\right) \quad \text{defined on } \mathcal{S}^d \times \mathcal{S}^d$$

and the *scalar (power) multiplication*

$$\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^{\alpha},\, \ldots,\, x_D^{\alpha}) \quad \text{defined on } \mathbb{R} \times \mathcal{S}^d$$

induce a vector space structure in to the unit simplex.

$\left(\mathcal{S}^d, \circ, \diamond\right)$ **is a vector space.**

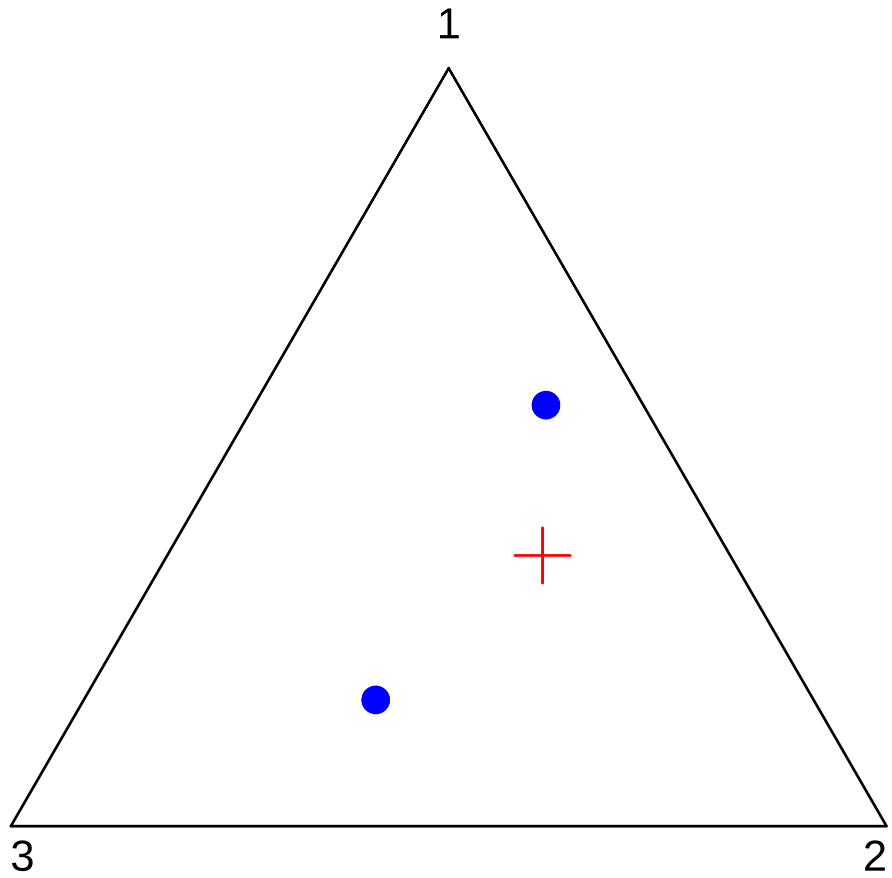
The *neutral element* of this vector space is the *barycenter*

$$\mathbf{e}_D := \left(\frac{1}{D}, \cdots, \frac{1}{D}\right) = \mathcal{C}\left(1,\, \ldots,\, 1\right)$$

and the *inverse element* of a composition $\mathbf{x} \in \mathcal{S}^d$ is

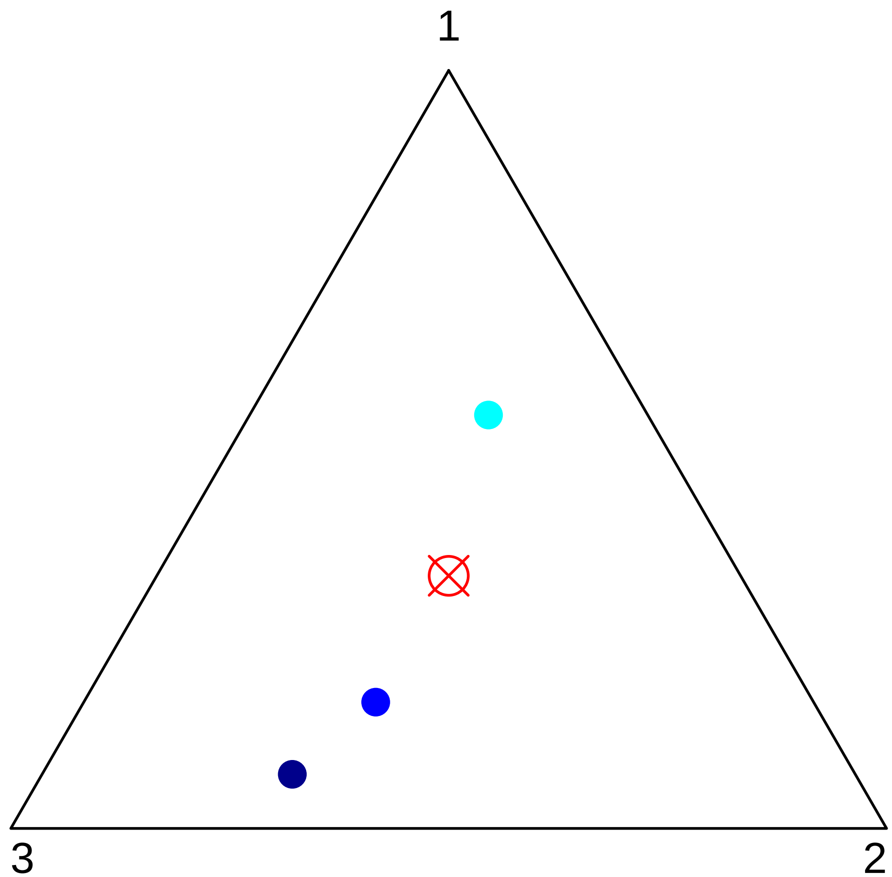
$$\mathbf{x}' := \mathcal{C}\left(\frac{1}{x_1}, \cdots, \frac{1}{x_D}\right) = -1 \diamond \mathbf{x}.$$

Perturbations



Perturbation operation of compositions $\mathbf{x} = (0.17, 0.33, 0.50)$ and $\mathbf{y} = (0.56, 0.33, 0.11)$ is $\mathbf{x} \circ \mathbf{y} = (0.36, 0.43, 0.21)$.

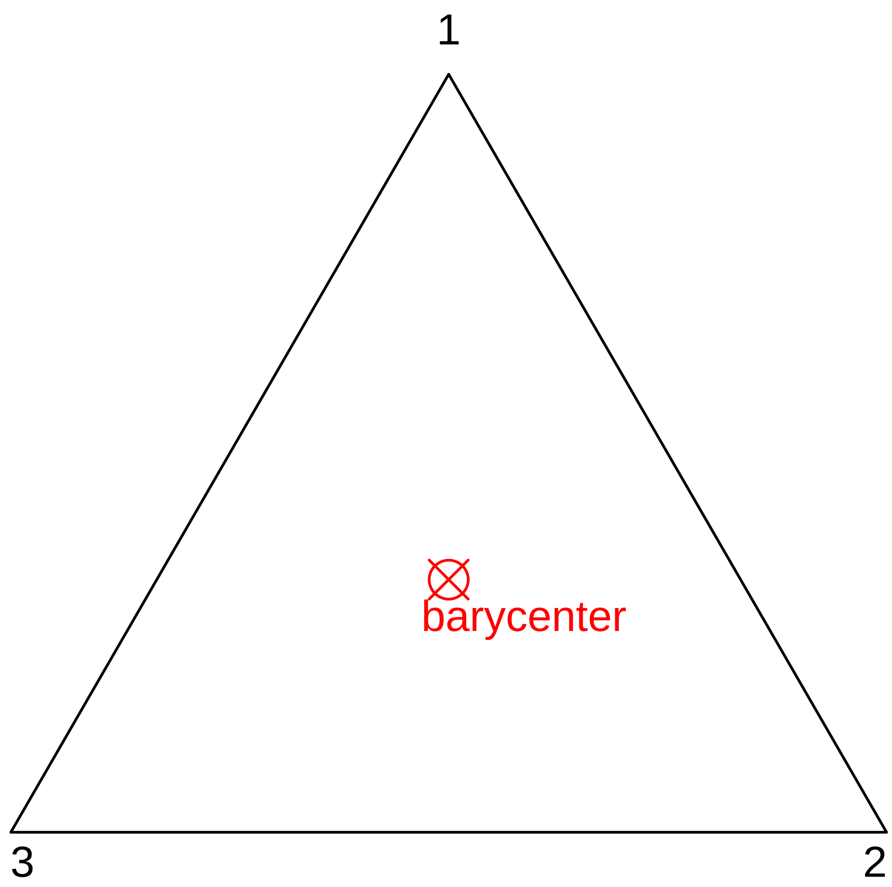
The scalar (power) multiplication



The scalar (power) multiplication of composition $\mathbf{x}$:

$$2 \diamond \mathbf{x}, \ -1 \diamond \mathbf{x}.$$

Barycenter



Ternary Diagram with the barycenter **e**.

Subcomposition

If we are interested only in some of measured properties – only in some part of the composition

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

we just skip the no more observed components and in order to keep the unit sum constraint we divide with the new sum:

For the $S \subset \{1, 2, \dots, D\}$ and $s := |S|$ we get the mapping

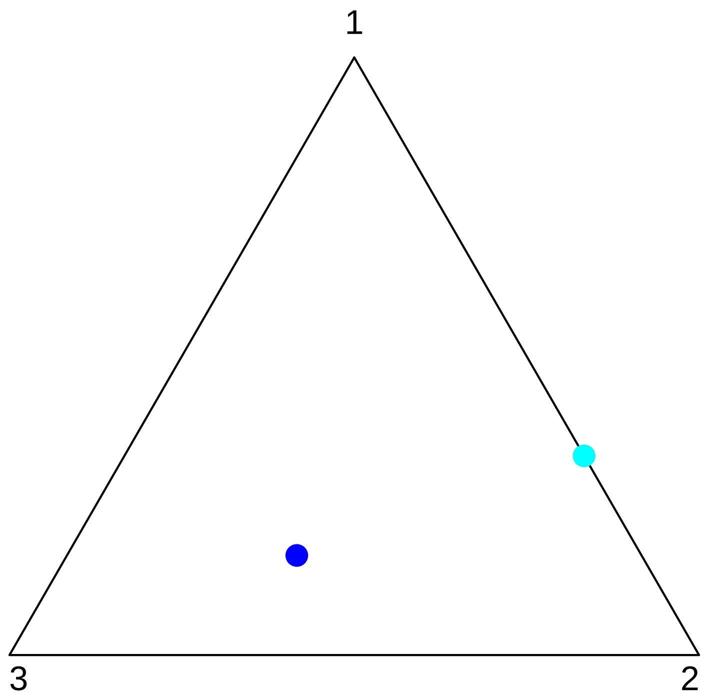
$$\mathbf{x} \in \mathcal{S}^d \longrightarrow \mathbf{x}_S \in \mathcal{S}^s$$

defined with

$$\mathbf{x}_S := \frac{1}{\sum_{i \in S} x_i} (x_{i_1}, \cdots, x_{i_s})$$

and we call $\mathbf{x}_S$ the *subcomposition* of the composition $\mathbf{x}$.

. . . Subcomposition



Two part subcomposition.

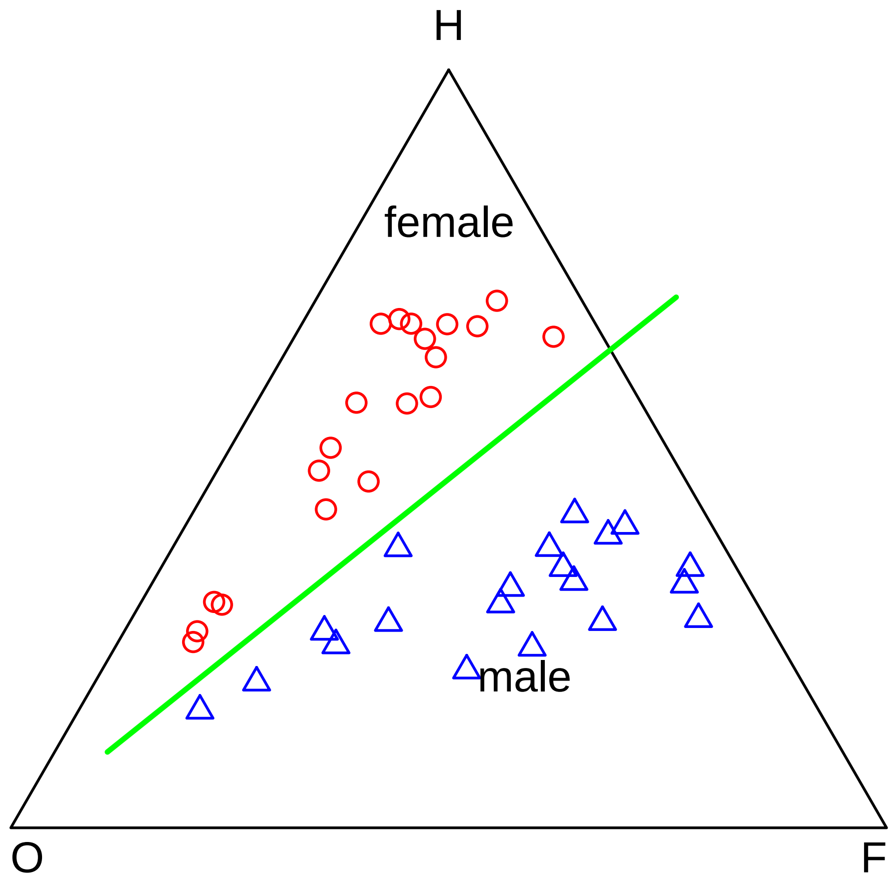
. . . Subcomposition

>h <- mix.Read("house.dat"); h

```
$tit
[1] "Household budget survey"
$sum
[1] NA
$sta
[1] 0
$mat
       H    F    O    S
M1   497  591  153  291
M2   839  942  302  365
.................
W19  382   77  230  147
W20 1090   59  313  177
attr(,"class")
[1] "mixture"
```

>mix.Ternary(h)

. . . Subcomposition



The three part subcomposition of Household data.

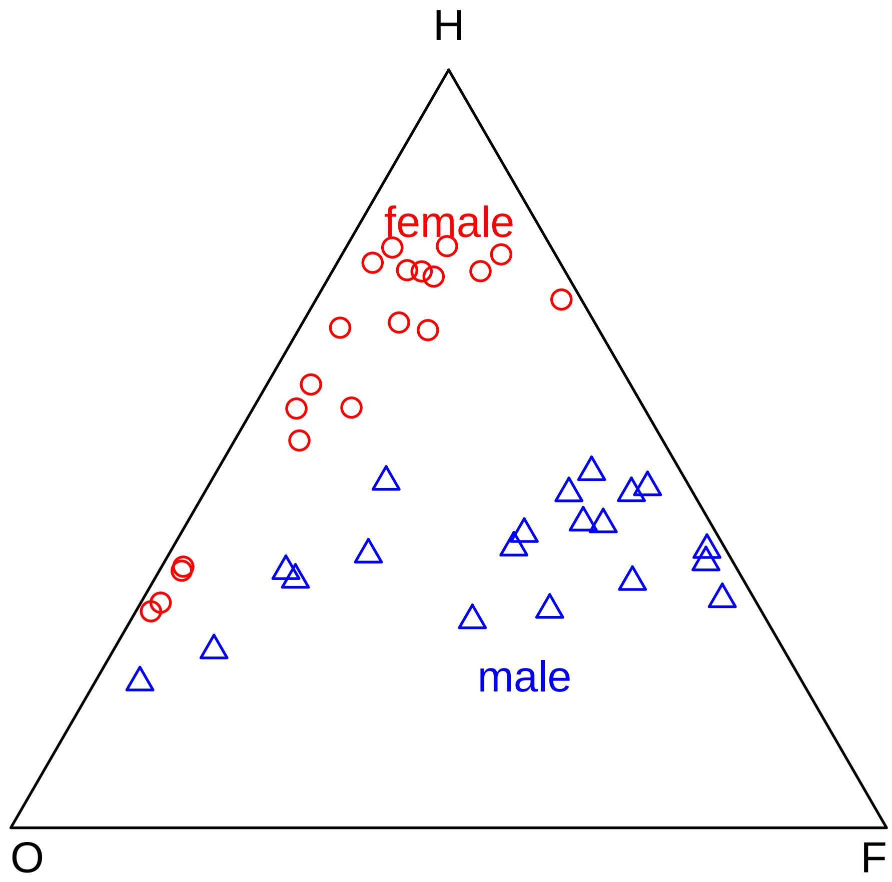
...Subcomposition

```
> h4 <- mix.Sub(h,4) ; h4
```

```
$tit
[1] "Household budget survey"
$sum
[1] 1
$sta
[1] 2
$mat
            H          F          O
M1  0.4004835 0.47622885 0.12328767
M2  0.4027844 0.45223236 0.14498320
..........................
W19 0.5544267 0.11175617 0.33381713
W20 0.7455540 0.04035568 0.21409029
attr(,"class")
[1] "mixture"
```

```
> mix.Ternary(h4)
```

. . . Subcomposition



The three part HOF subcomposition of Household data.

Centered data set

The *geometric mean* of the set of compositions

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}^d$$

is defined

$$G(X) := \mathcal{C}\left(g_1, \ldots, g_D\right) \quad \text{where} \quad g_k := \left(\prod_{j=1}^{N} x_{jk}\right)^{1/N}$$

is the geometric mean of the components $k = 1, \ldots, D$.

Geometric mean is the adequate measure of central tendency for compositional data:

- $G(\mathbf{y} \circ X) = \mathbf{y} \circ G(X)$ for all $\mathbf{y} \in \mathcal{S}^d$,
- $G(\lambda \diamond X) = \lambda \diamond G(X)$ for all $\lambda \in \mathbb{R}$.

. . . Centered data set

In case that the data set $\mathbf{X}$ is near to the corner – this happens when one of the components of the data set is near to 1 it is very difficult to establish if there are differences between the points.

If we perturb the data set $\mathbf{X}$ by the $-1 \diamond G(\mathbf{X})$ the result data set is centered, i.e. the center of the set $-1 \diamond G(\mathbf{X}) \circ \mathbf{X}$ is the barycenter of the simplex

$$G(-1 \diamond G(\mathbf{X}) \circ \mathbf{X}) = \mathbf{e}.$$

Now we can observe the real pattern of the data (in Aitchison's geometry!).

EXAMPLE: Household budget survey

> mix.Gmean(h4)

```
$tit
[1] "Geometric mean of the data"
$sum
[1] 1
$sta
[1] 2
$mat
             H         F         O
[1,] 0.5656061 0.1909976 0.2433962
$class
[1] "mixture"
```

> mix.InvGmean(h4)

```
$tit
[1] "Inverse geometric mean of the data"
$sum
[1] 1
$sta
[1] 2
$mat
             H         F         O
[1,] 0.1591056 0.4711635 0.3697309
$class
[1] "mixture"
```

> pert(mix.Gmean(h4)$mat,mix.InvGmean(h4)$mat)

```
[1] 0.3333333 0.3333333 0.3333333
```

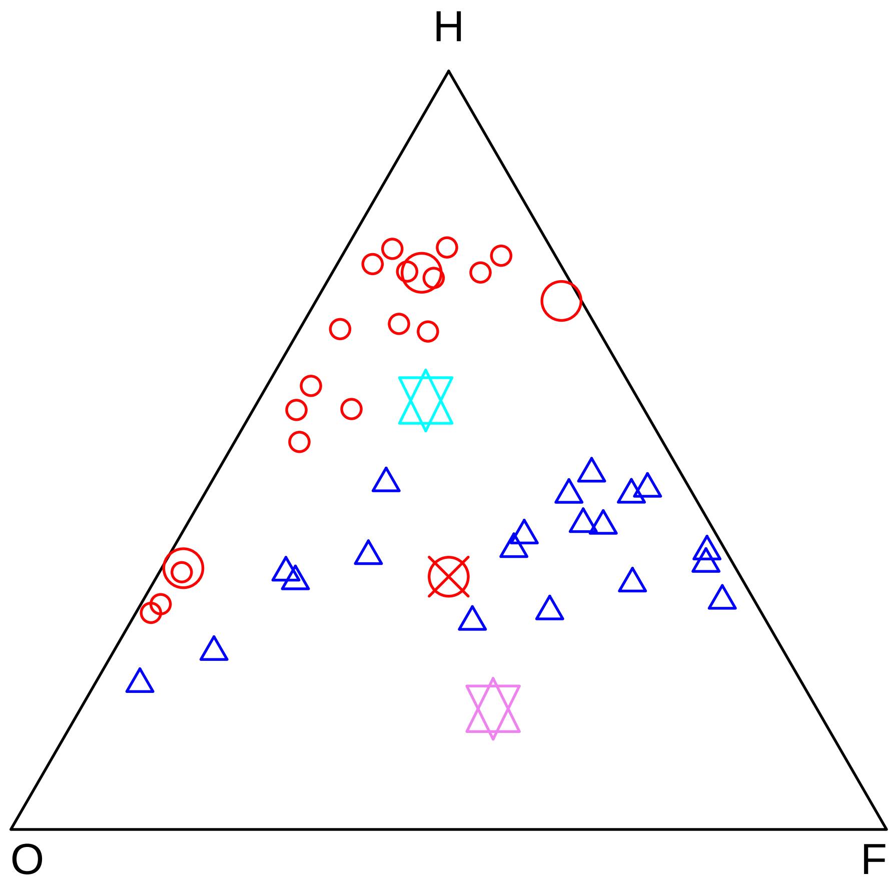
EXAMPLE: Household budget survey

```
> G <- mix.Matrix(rbind(mix.Unitn(3)$mat, mix.Gmean(h4)$mat,
> + mix.InvGmean(h4)$mat, h4$mat), "Barycenter, Gmean, InvGmean
```

```
$tit
[1] "Barycenter, Gmean, InvGmean, HOF Data"
$sum
[1] 1
$sta
[1] 2
$mat
            H          F          O
    0.3333333 0.33333333 0.33333333
    0.5656061 0.19099763 0.24339624
    0.1591056 0.47116349 0.36973090
M1  0.4004835 0.47622885 0.12328767
M2  0.4027844 0.45223236 0.14498320
................................
W19 0.5544267 0.11175617 0.33381713
W20 0.7455540 0.04035568 0.21409029
$class
[1] "mixture"
```

```
> t <- c(rep(1,20),rep(2,20))
> spol <- c("blue", "red")
> liki <- c(22,2)
> mix.Ternary(G,col=c("red","cyan","violet",spol[t]),
> + pch=c(13,11,11,liki[t]), cex=c(rep(2,3),rep(1,20)))
```

The geometric mean and it’s inverse



The geometric mean and it’s inverse of the three part subcomposition of the House data.

Centered Data

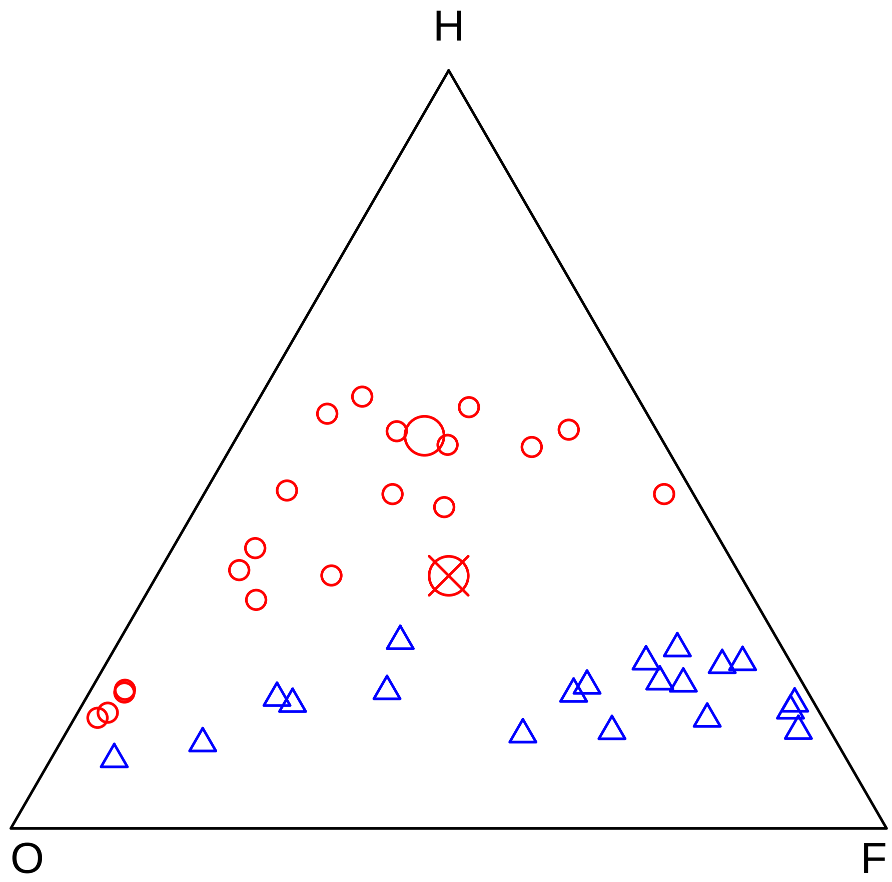
```
> mix.Center(h4)
```

```
$tit
[1] "Centered Data"
$sum
[1] 1
$sta
[1] 2
$mat
             H          F          O
M1  0.19095658 0.67243738 0.13660604
M2  0.19374839 0.64418881 0.16206280
..........................
W19 0.33377082 0.19923327 0.46699592
W20 0.54716950 0.08770685 0.36512365
$class
[1] "mixture"
```

```
> G1 <- mix.Matrix(rbind(mix.Unitn(3)$mat, mix.Center(h4)$mat),
> + "Barycenter and Centered Data")
```

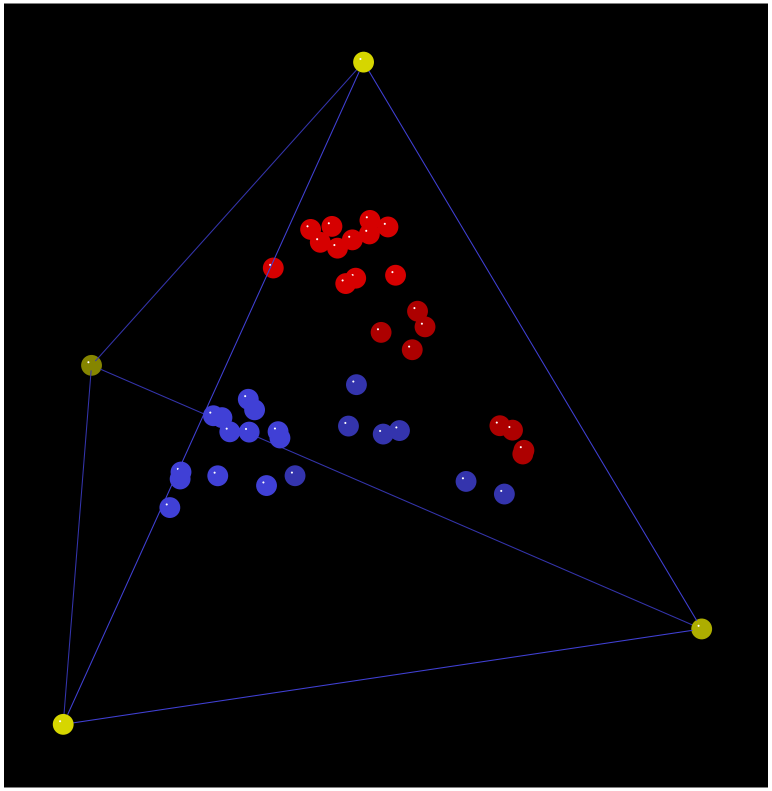
```
> mix.Ternary(G1,col= c("red",spol[t]),
> + pch=c(13,liki[t]),cex=c(rep(2,1),rep(1,20)))
```

Centered subcomposition



Centered HOF three part subcomposition of the House data with the barycenter.

Tetrahedral display



Snapshot of Kinimage view of tetrahedral display of Household budget survey

K. Urner: Quadrays and XYZ; T. Ace: Quadray formulas; Mage viewer.

Conclusions

From the abstract we resume:

We need an R library for compositional data analysis comprehending compositional concepts jet not applied originally in R. Programming in R

operations on compositions such as perturbation, power multiplication, subcomposition, distances . . .

various logratio transformations of compositions to transform compositions into real vectors that are amenable to standard multivariate statistical analysis,

compositional concepts such as complete subcompositional independence, the relation of compositions to bases, logcontrast models . . . and

graphical presentation of compositions in ternary diagrams and tetrahedrons

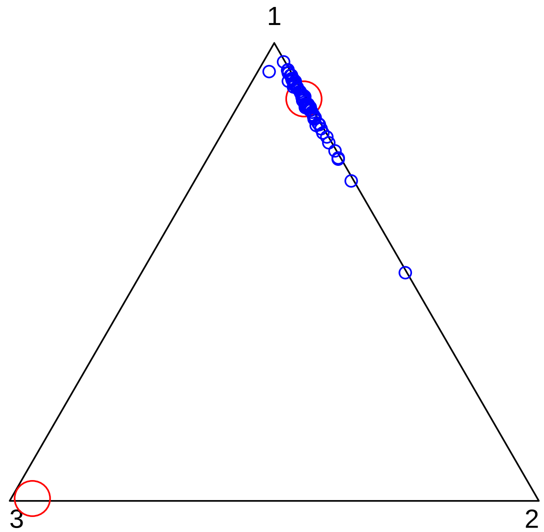
. . . Conclusions

will provide an GNU library for compositional data analysis.

And we conclude:

We have managed the first and the last item. The rest is our goal in future.

Appendix



The geometric mean and it’s inverze of the three part subcomposition of the Labour data.