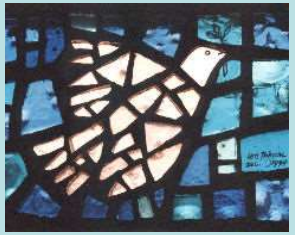
IUGS-CGGB International Workshop, 18 November 2021

Compositional Data Analysis: Graphical tools and software

Michael Greenacre
Universitat Pompeu Fabra
Barcelona
www.econ.upf.edu/~michael



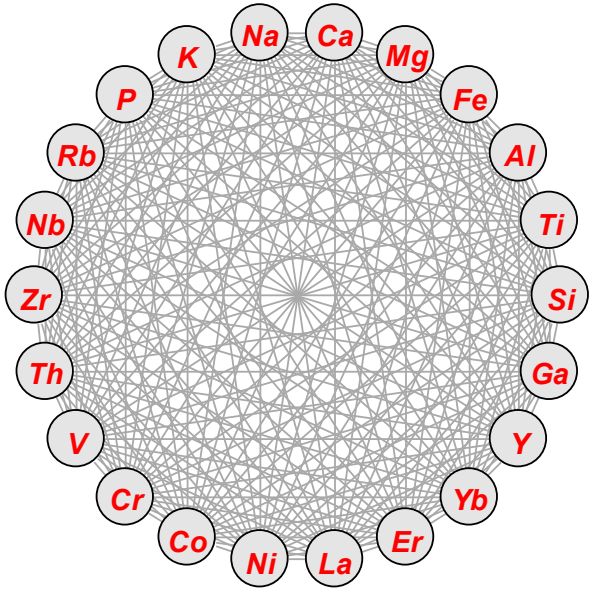
www.globalsong.net

www.multivariatestatistics.org
https://github.com/michaelgreenacre/CODAinPractice
youtube.com/StatisticalSongs

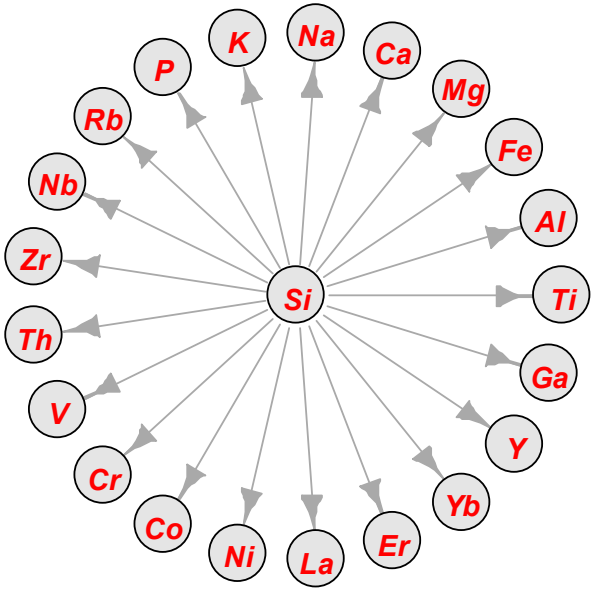
Email: michael.greenacre@upf.edu

Types of logratios
(graph representation)

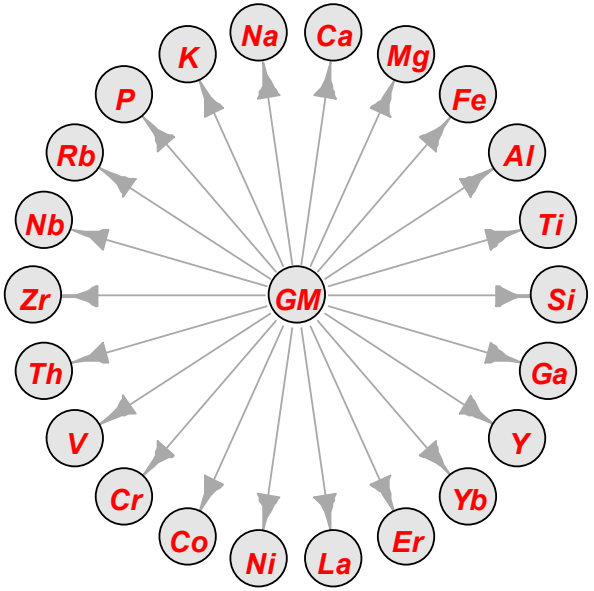
All pairwise logratios



Additive logratios



Centred logratios
w.r.t. geometric mean (GM)



`igraph package` for drawing graphs

`easyCODA package` for computing logratios

`LR()` `ALR()` `CLR()`

Ratios: univariate statistics

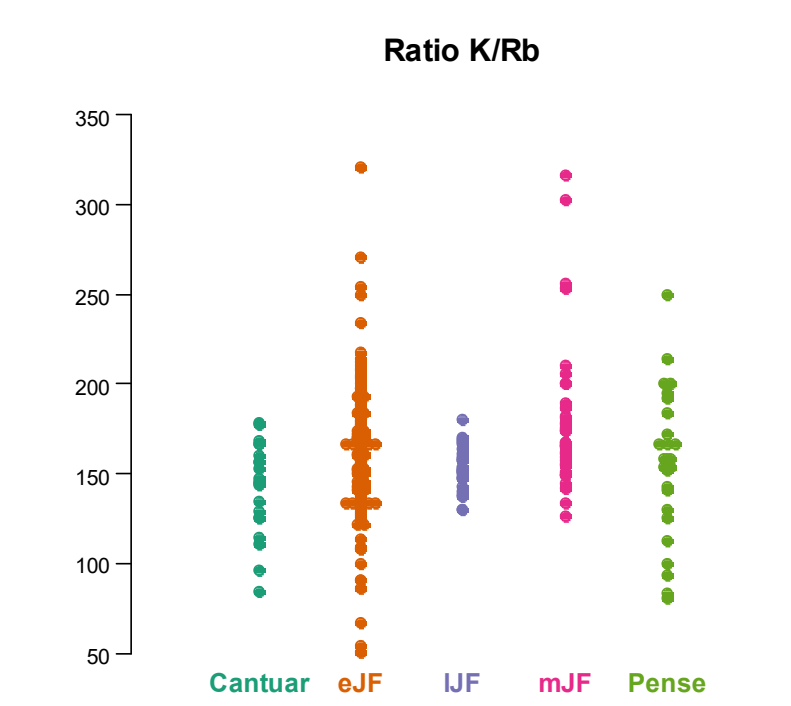
RATIO	MEDIAN	95% REFERENCE RANGE
Mg/K	199.85	(52.61 , 1107.62)
Si/La	5714.4	(1667.4 , 9902.5)
Si/Na	187.30	(51.14 , 1740.11)
Al/Ca	0.602	(0.170 , 1.278)
Al/Ni	15.556	(7.463 , 36.864)
Rb/Yb	19.891	(6.566 , 59.709)
Ti/P	4.475	(1.841 , 7.062)
Rb/La	0.160	(0.032 , 0.442)

using **quantile** function in R, after bootstrapping the median:

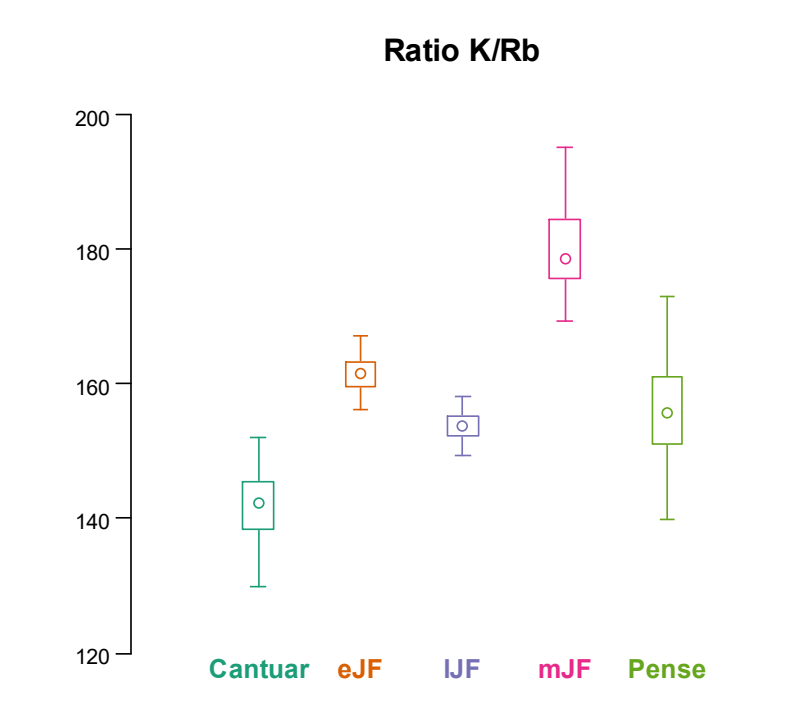
quantile(x, c(0.025, 0.975))

Showing group differences: univariate

Ranges of values



Confidence intervals for means



Left: dot plots using **`plotdot`** from **`multivariatestatistics.org`** or **`multivariatestatistics.net`** or **`DOT()`** from **`easyCODA`** package, or direct download in R

```
source("http://www.econ.upf.edu/~michael/multivariatestatistics/plotdot.R")
```

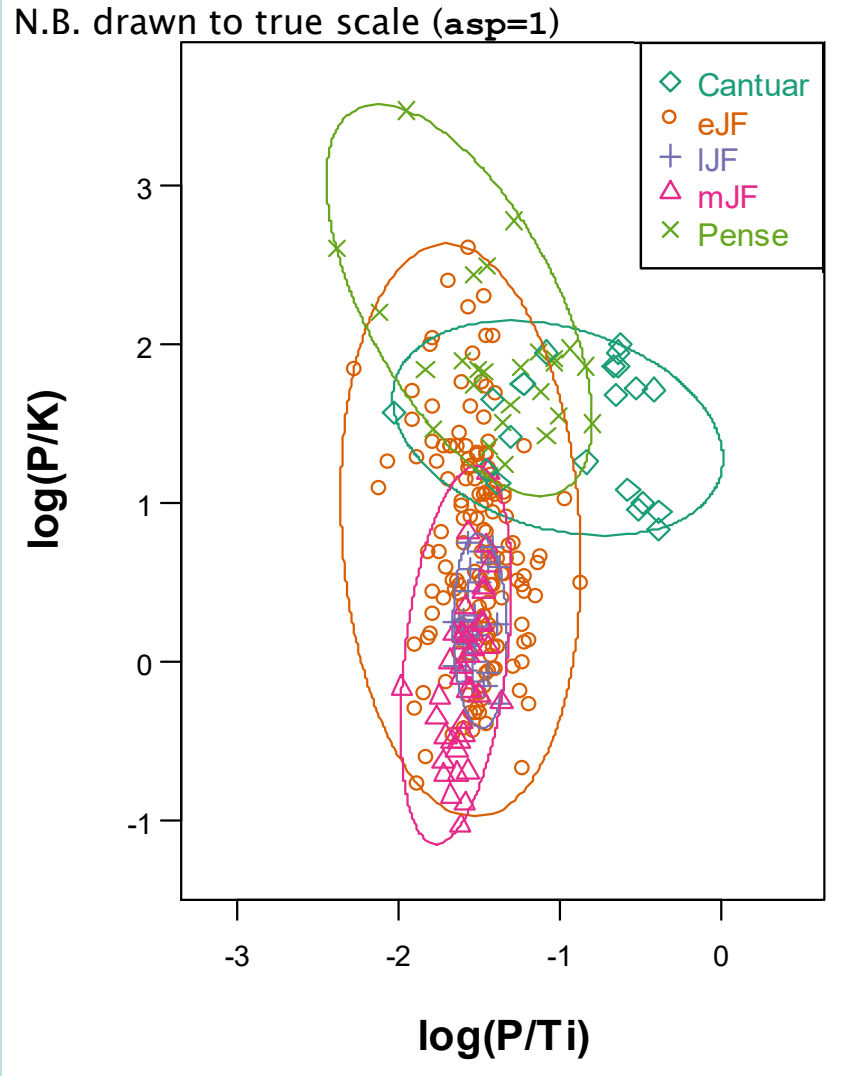
Right: confidence plots using **`CIplot_uni()`** from **`multivariatestatistics.org`** or **`multivariatestatistics.net`** or direct download in R

```
source("http://www.econ.upf.edu/~michael/multivariatestatistics /CIplot_uni.R")
```

Greenacre, M. (2016) Data reporting and visualization in ecology. *Polar Biology* 39:2189–2205

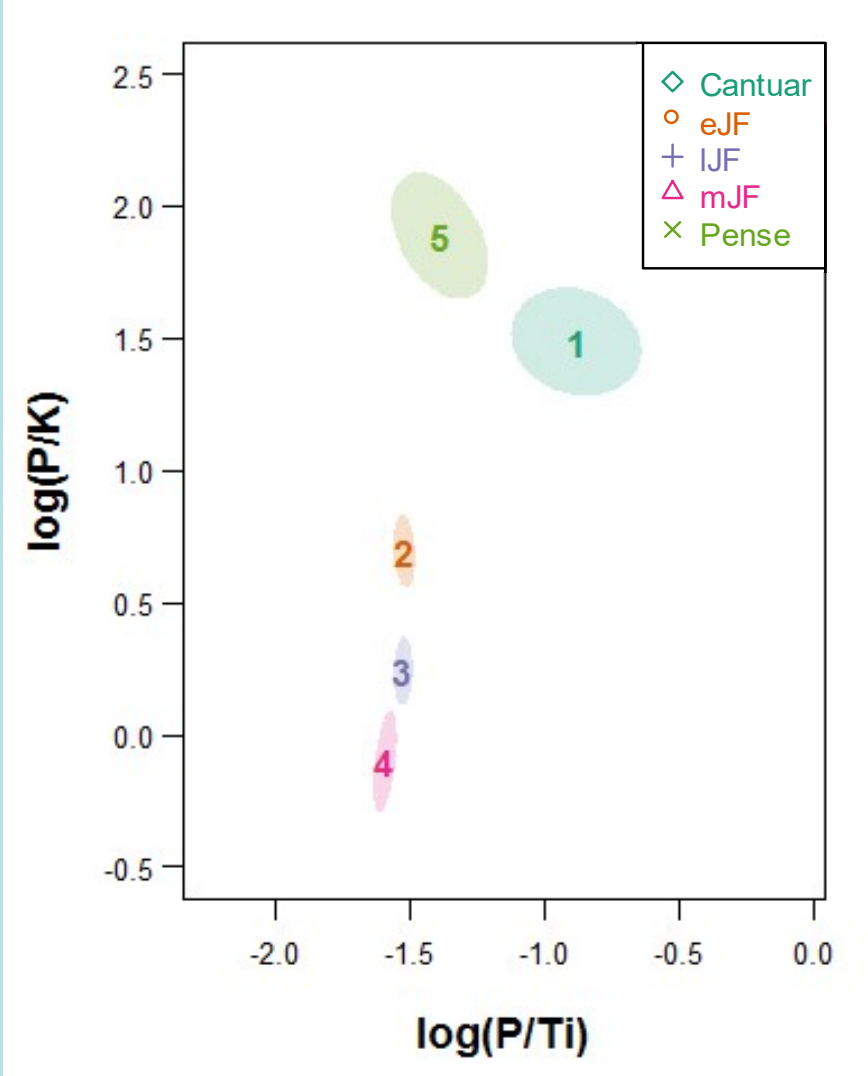
Showing group differences: bivariate

Covering ellipses



Uses `ellipsoidhull()` function
in package `cluster`

Confidence ellipses for means



Uses `CIplot_biv()` `from easyCODA`
or from `multivariatestatistics.net /.org`

```
source(" http://www.econ.upf.edu/~michael/multivariatestatistics/CIplot_biv.R")
```

Three-part (sub)composition in the simplex

P	Ti	K
0.01863	0.09596	0.01118
0.01377	0.10816	0.00393
0.02864	0.11063	0.00988
0.02399	0.11395	0.00700
0.02674	0.12135	0.00720
0.01763	0.11653	0.00490
0.02578	0.11720	0.00703
0.02364	0.11348	0.00662
0.01876	0.15723	0.00625
0.02267	0.10942	0.02169
:	:	:
:	:	:

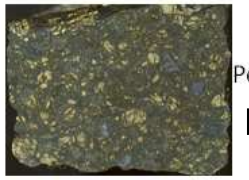
close →

P	Ti	K
0.14815	0.76296	0.08889
0.10938	0.85937	0.03125
0.19205	0.74172	0.06623
0.16552	0.78621	0.04828
0.17219	0.78146	0.04636
0.12676	0.83803	0.03521
0.17188	0.78125	0.04688
0.16447	0.78947	0.04605
0.10294	0.86275	0.03431
0.14744	0.71154	0.14103
:	:	:
:	:	:

Early Joli Fou **eJF**

Mid Joli Fou **mJF**

Late Joli Fou **lJF**

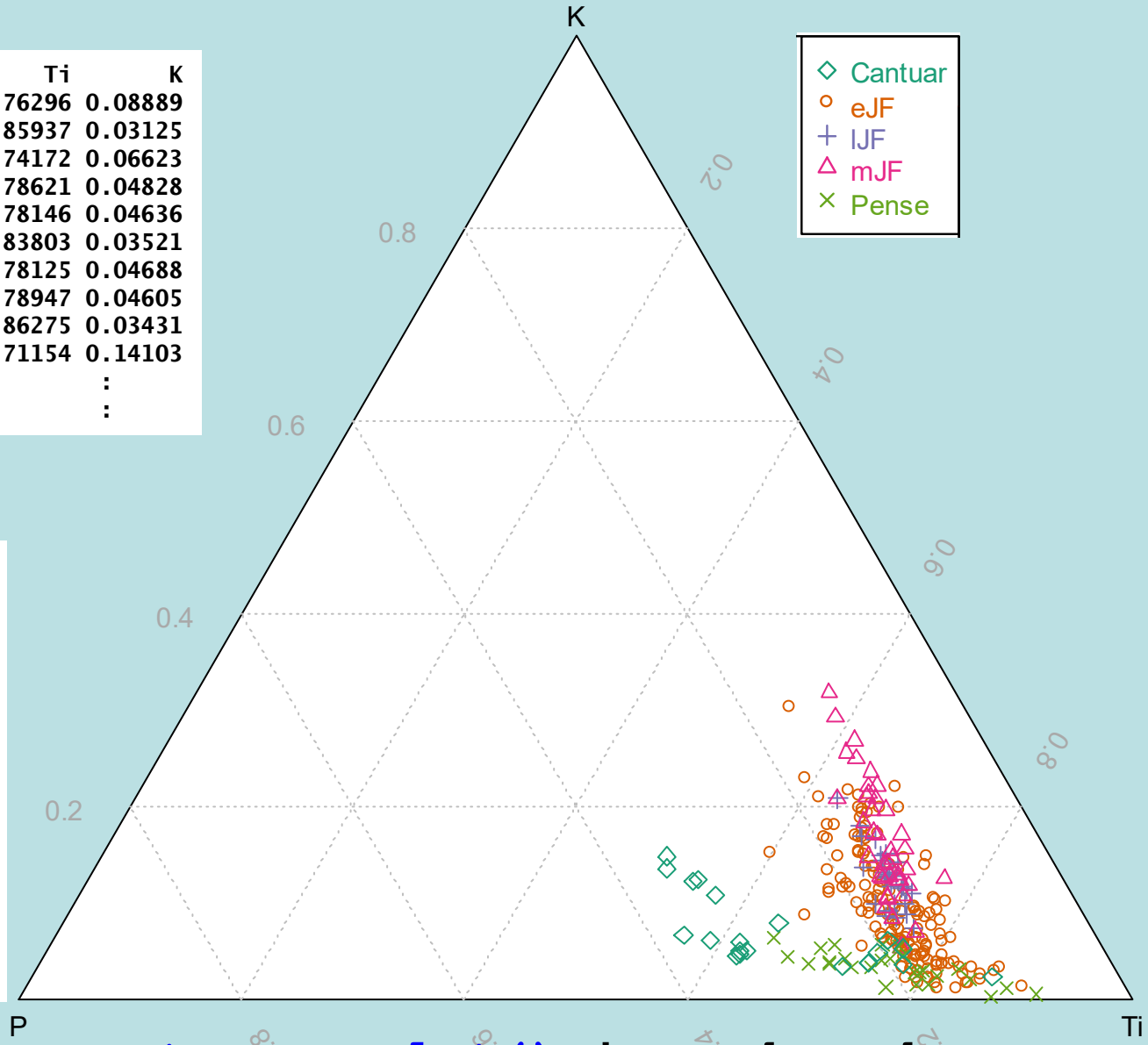


Pense

Pense

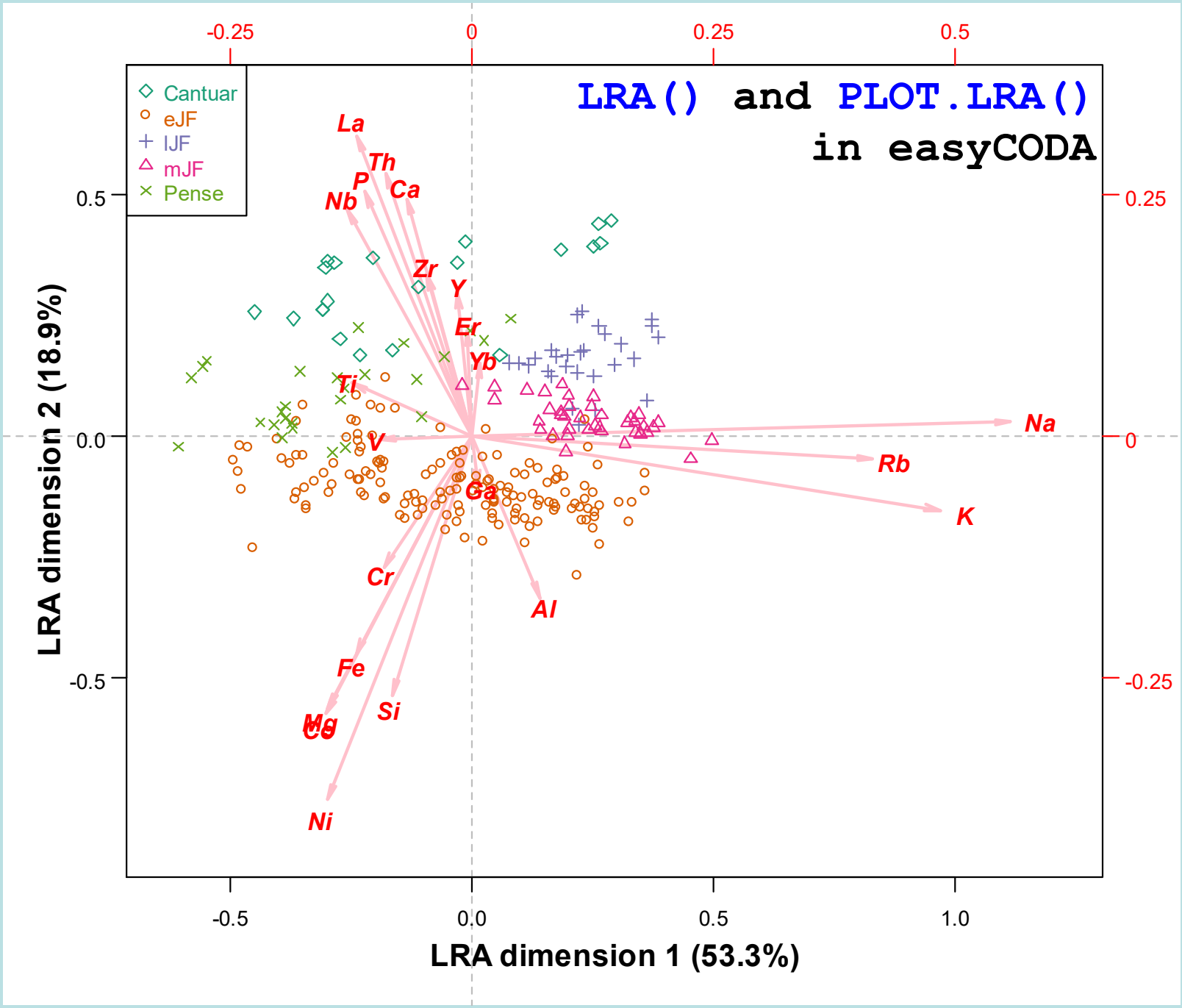
Cantuar

Cantaur



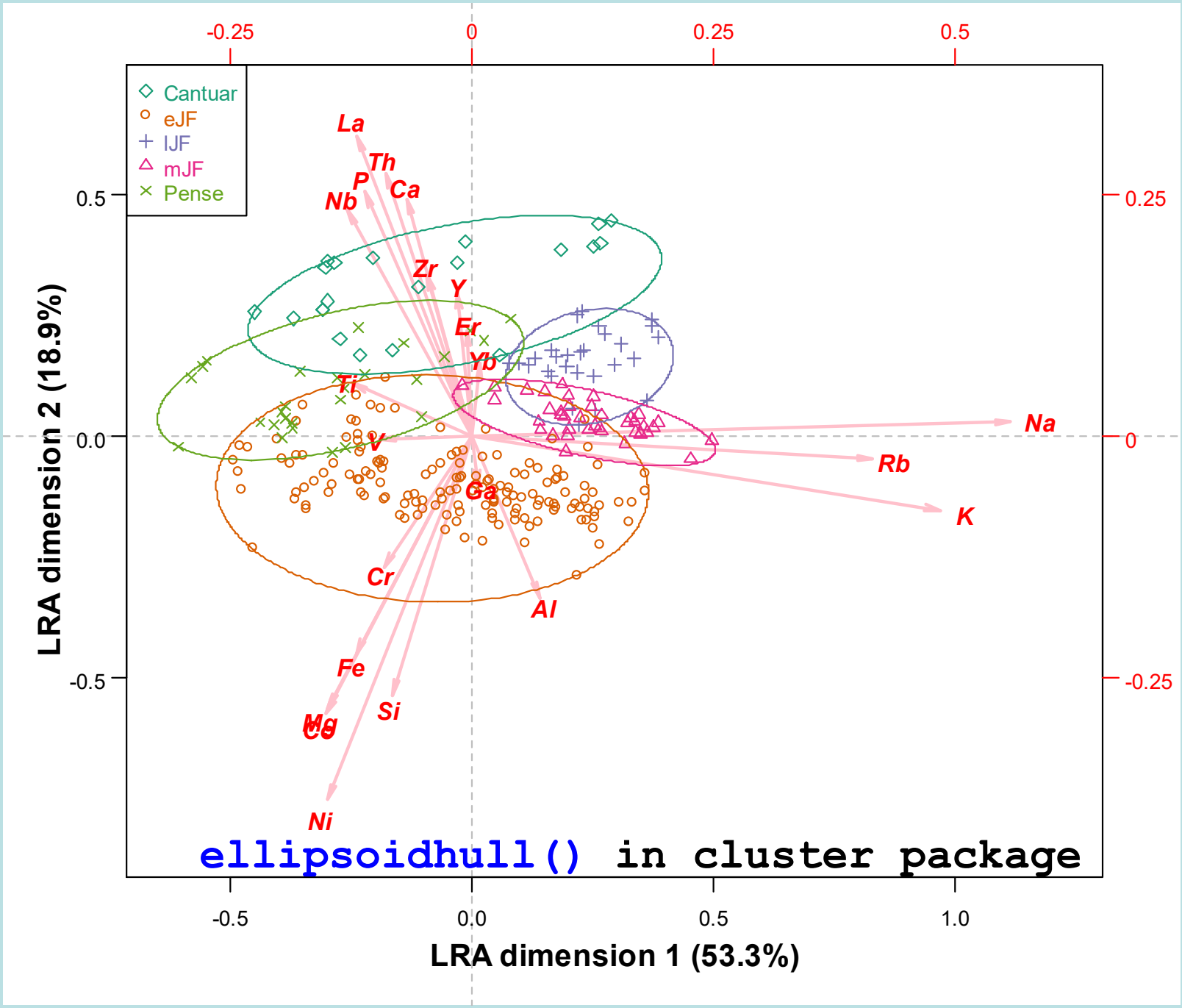
ternaryplot() in vcd package

Logratio biplot



Logratio biplot

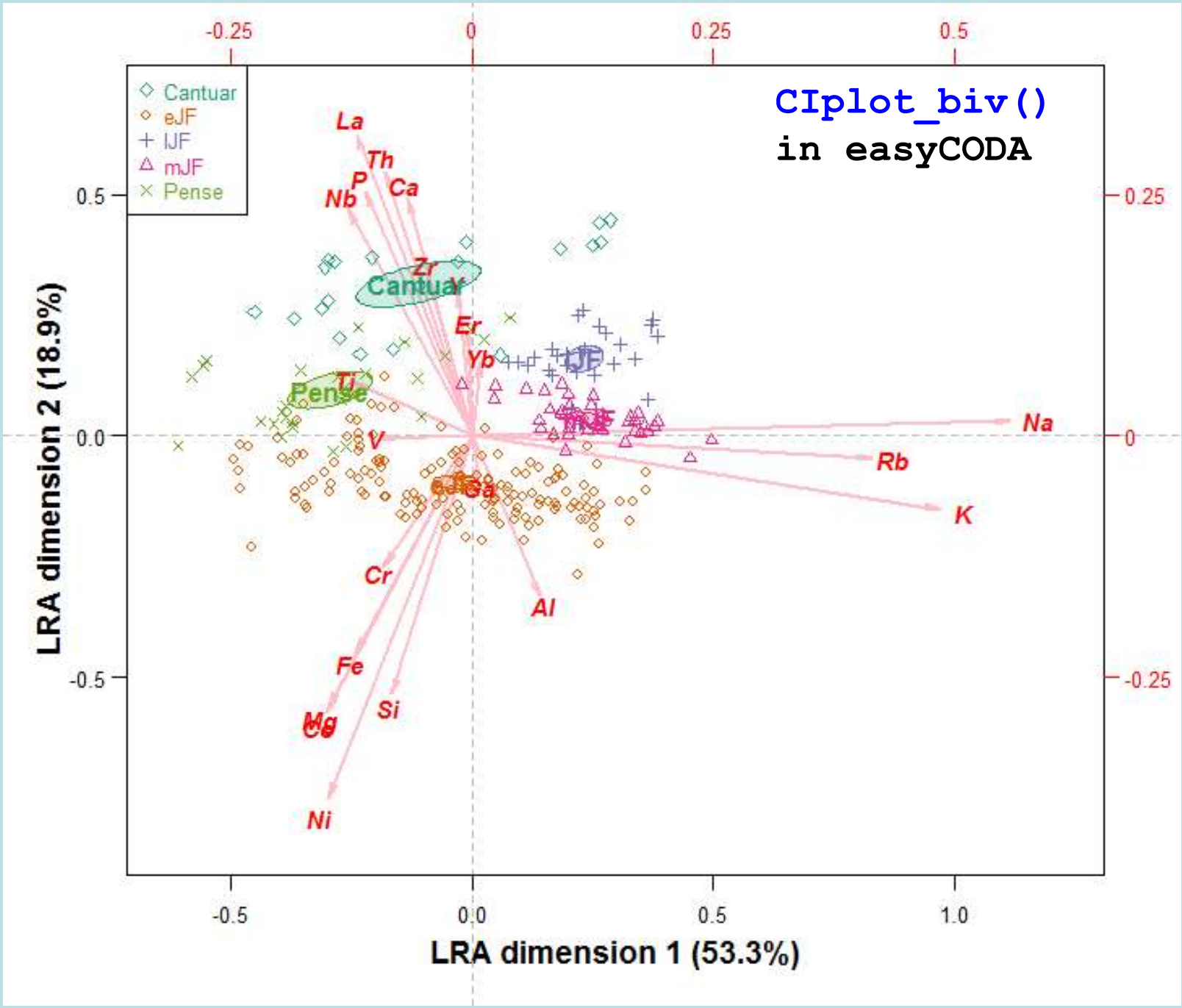
with minimum covering ellipses around groups



Between
57.9%
Within
42.1%

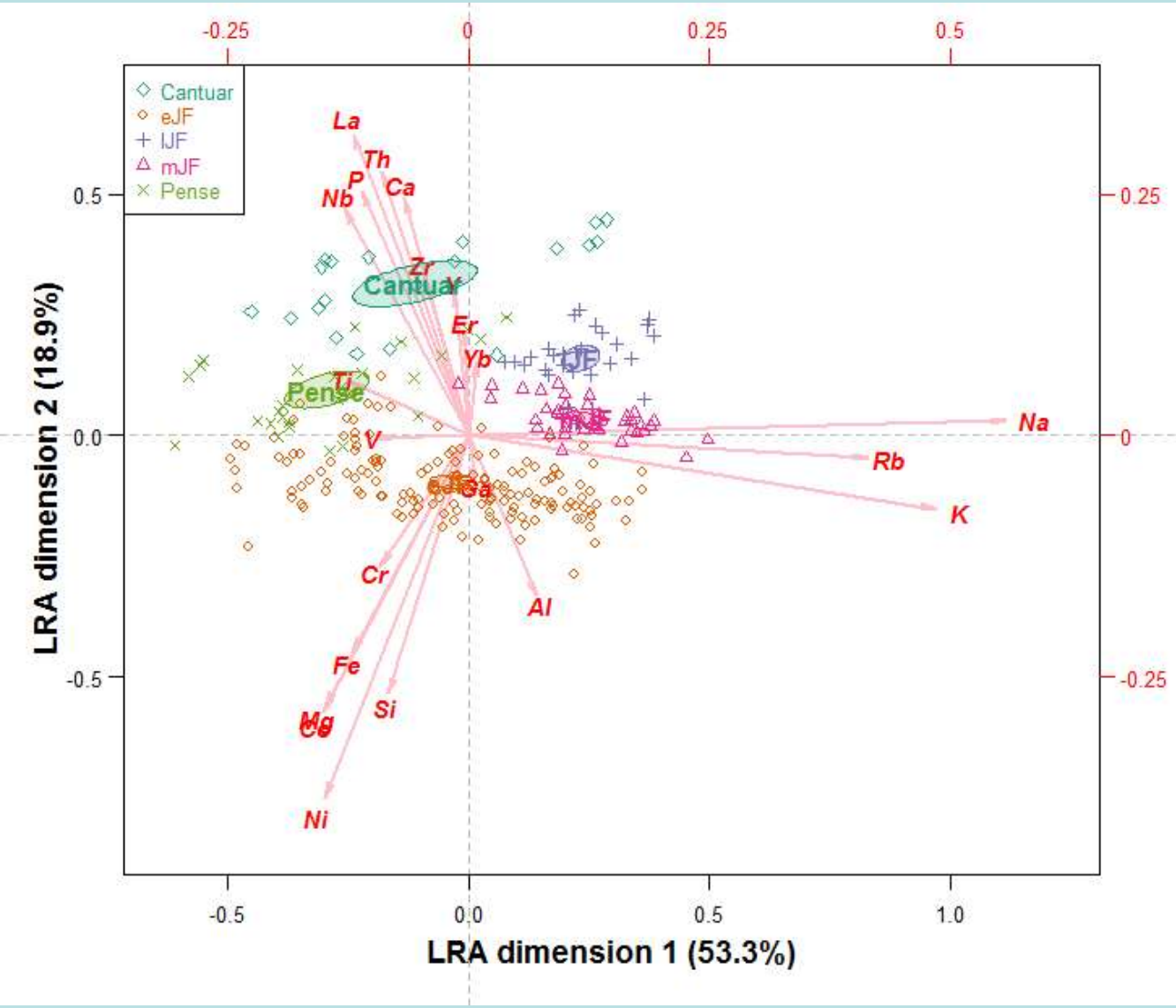
Logratio biplot

with 95% confidence ellipses
around group means

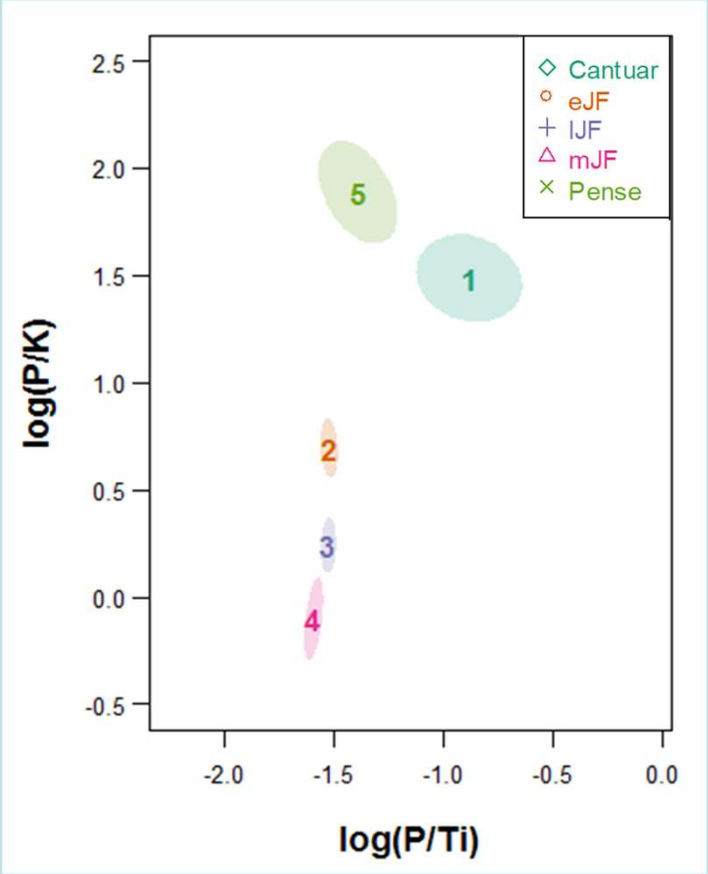


Logratio biplot

with 95% confidence ellipses around group means



compared to confidence ellipses in previous scatterplot of just two logratios



This suggests that perhaps we can get away with a few logratios to identify the essential structure in a compositional data set...

Stepwise selection of Kimberlite ratios

Logratio analysis using all parts

Dimn	eigenvalue	%	cum%	scree plot
1	0.060354	53.3	53.3	**********
2	0.021315	18.8	72.1	*****
3	0.013914	12.3	84.4	***
4	0.006617	5.8	90.2	*
5	0.003457	3.1	93.3	*
6	0.001740	1.5	94.8	
7	0.001051	0.9	95.8	
8	0.000971	0.9	96.6	
9	0.000852	0.8	97.4	
10	0.000574	0.5	97.9	
11	0.000459	0.4	98.3	
12	0.000411	0.4	98.6	
13	0.000316	0.3	98.9	
14	0.000290	0.3	99.2	
15	0.000208	0.2	99.4	
16	0.000171	0.2	99.5	
17	0.000157	0.1	99.7	
18	0.000141	0.1	99.8	
19	0.000119	0.1	99.9	
20	8.1e-050	0.1	100.0	
21	5.1e-050	0.0	100.0	
	--------	-----		
Total:	0.113247	100.0		

Decomposition of total variance along principal axes

Total variance (“inertia”)

Stepwise logratio selection

Step	Ratio	R^2	Procrustes
1	Mg/K	0.473	0.688
2	Si/La	0.666	0.729
3	Si/Na	0.826	0.811
4	Al/Ca	0.884	0.868
5	Al/Ni	0.915	0.875
6	Rb/Yb	0.930	0.875
7	Ti/P	0.943	0.879
8	Rb/La	0.955	0.879
9	P/Ni	0.962	0.881
10	Si/V	0.969	0.883
11	Ni/Er	0.975	0.883
12	Ti/Ga	0.980	0.883
13	Ca/Th	0.984	0.884
14	Zr/Cr	0.988	0.884
15	Mg/Ni	0.990	0.890
16	Zr/Th	0.992	0.890
17	Y/Ga	0.994	0.890
18	Na/Zr	0.996	0.890
19	Co/Yb	0.997	0.890
20	Fe/K	0.999	0.895
21	Al/Nb	1	0.895

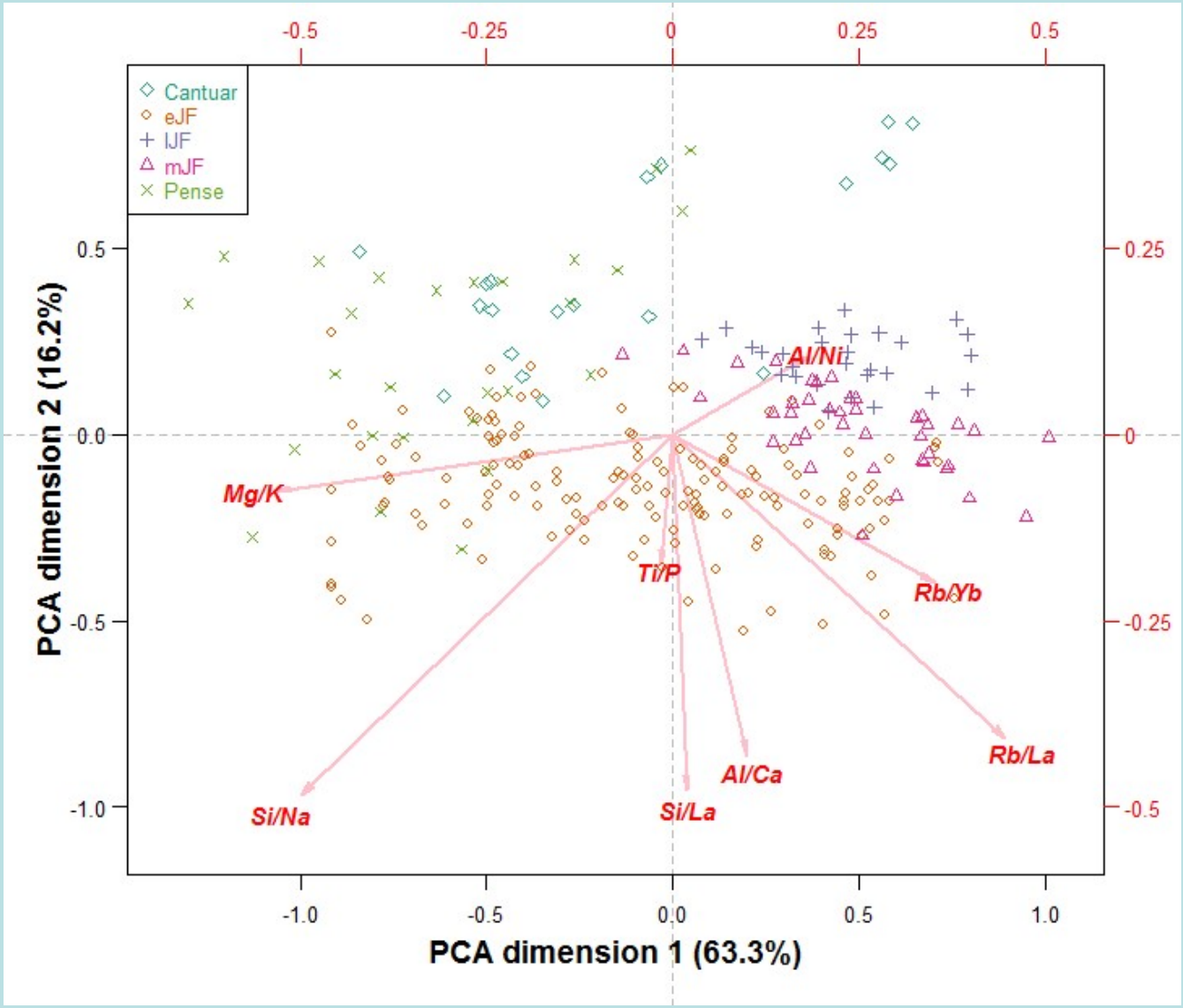
Decomposition of total variance w.r.t. optimal logratios

In easyCODA

```
kim.lra <- LRA(kim, weight=FALSE)
summary(kim.unlra)
```

```
kim.step <- STEP(kim, weight=FALSE)
 ..$rationames ..$R2max ..$procr
```

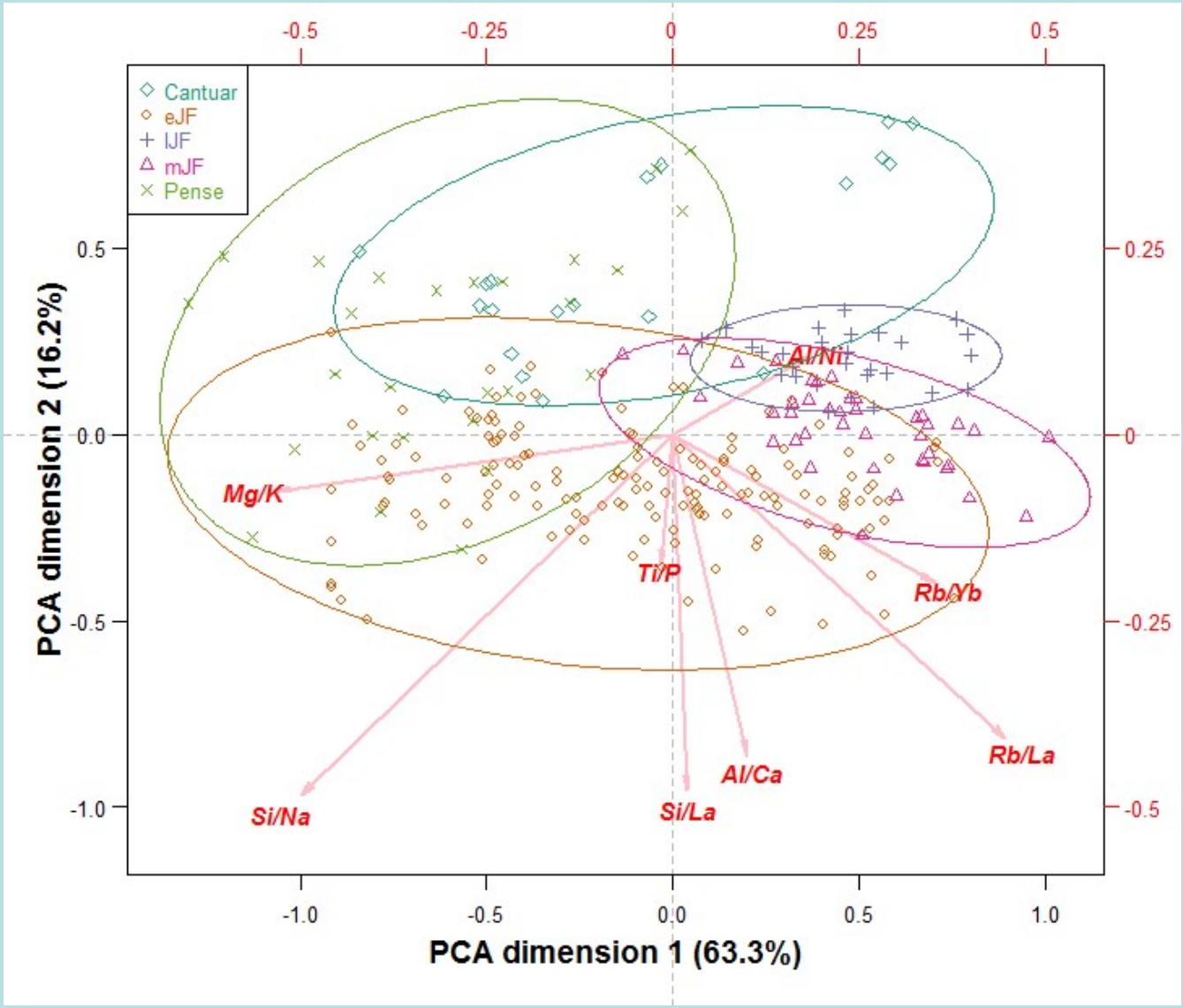
Best 8 logratios (95.5% of the variance)



Between
60.6%
Within
39.4%

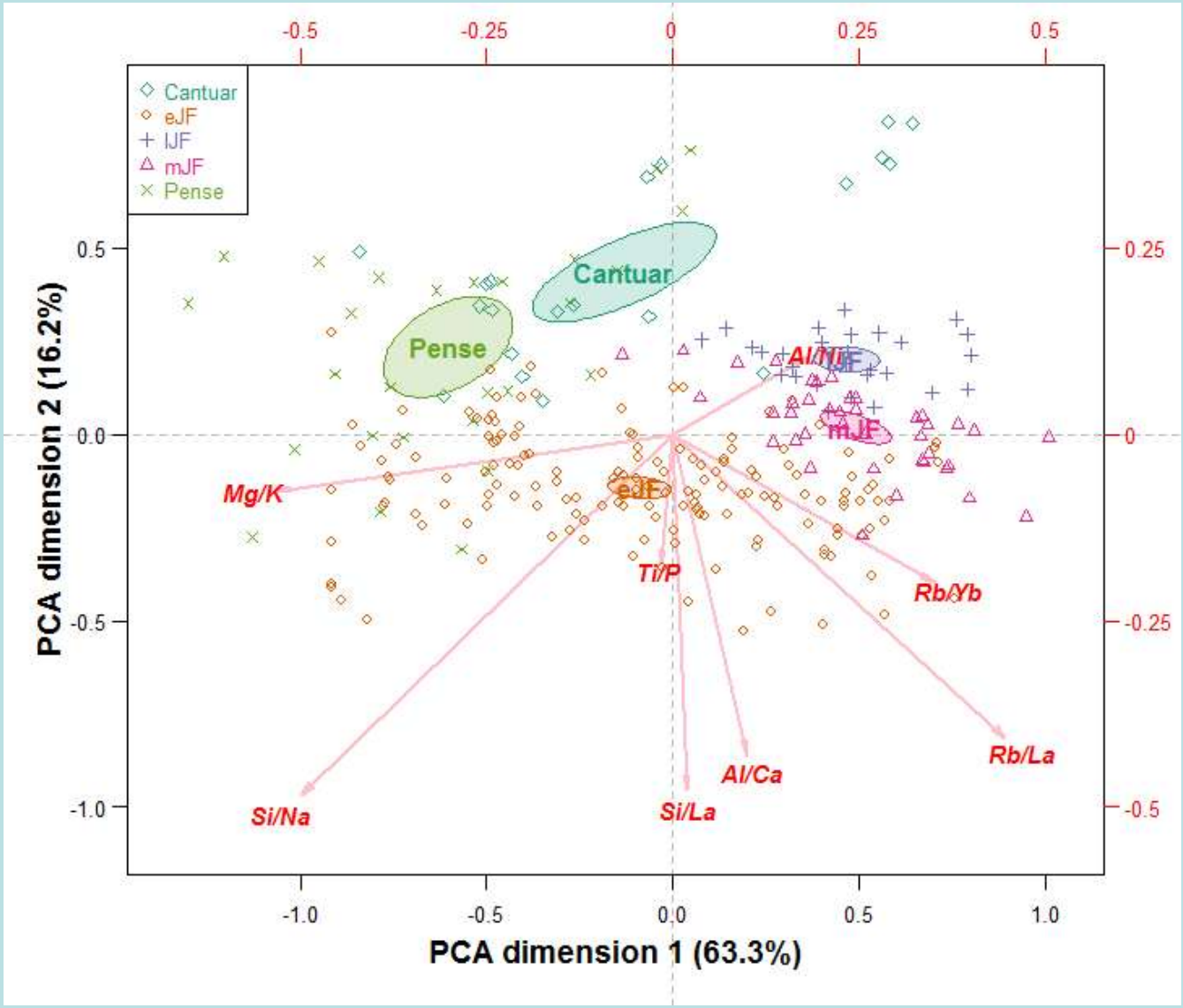
PCA() and PLOT.PCA() in easyCODA

Best 8 logratios (95.5% of the variance)



Between
60.6%
Within
39.4%

Best 8 logratios (95.5% of the variance)



Between
60.6%
Within
39.4%

Supervised selection step-by-step

- Now we look for logratios that maximally discriminate between the **groups** of samples, as opposed to the individual samples as done up to now
- This is a slight variation of the LRA/PCA where the points now become the group centroids, which I call centroid discriminant analysis.
- In addition, we do the selection step-by-step, with intervention by the specialist, in this case geologist EG. Here are the "top 15" for Step 1

STEP 1

Ratio	R2	Procr
Ca/Rb	58.2%	0.763
P/Rb	58.1%	0.762
K/P	57.6%	0.759
Ca/K	57.3%	0.757
K/V	57.1%	0.756
Rb/V	57.0%	0.755
Al/Cr	56.6%	0.752
Rb/La	56.5%	0.752
Rb/Nb	56.5%	0.751
K/Nb	56.3%	0.750
K/La	56.3%	0.750
Nb/Y	56.2%	0.750
Ti/K	55.4%	0.744
Fe/K	54.9%	0.741
Rb/Th	54.9%	0.741

chosen by EG, will be included in next step

```
STEP(kim.aggr, nsteps=1, top=15, weight=FALSE)
```

Supervised selection step-by-step

- Step 2

```
STEP 2: K/P and...

Ratio      R2     Procr
--------------------
K/Ni      90.5%  0.906
P/Ni      90.5%  0.925
Ti/Th     90.4%  0.842
Si/Nb     90.4%  0.895
Si/La     90.3%  0.908
Mg/V      90.3%  0.877
Ca/Ni     90.2%  0.935
P/Co      90.2%  0.919
K/Co      90.2%  0.894
Rb/Ni     90.2%  0.916
Mg/La     90.1%  0.917
Fe/La     90.1%  0.910
Si/V      90.1%  0.856
Si/Th     89.2%  0.905
Co/La     89.9%  0.924
```

chosen by EG, will be included in next step

```
STEP(kim.aggr, nsteps=1, top=15, weight=rep(1/ncol(kim), ncol(kim)),
     previous=log(kim.aggr[,"K"]/kim.aggr[,"P"]))
```

Supervised selection step-by-step

- Step 3

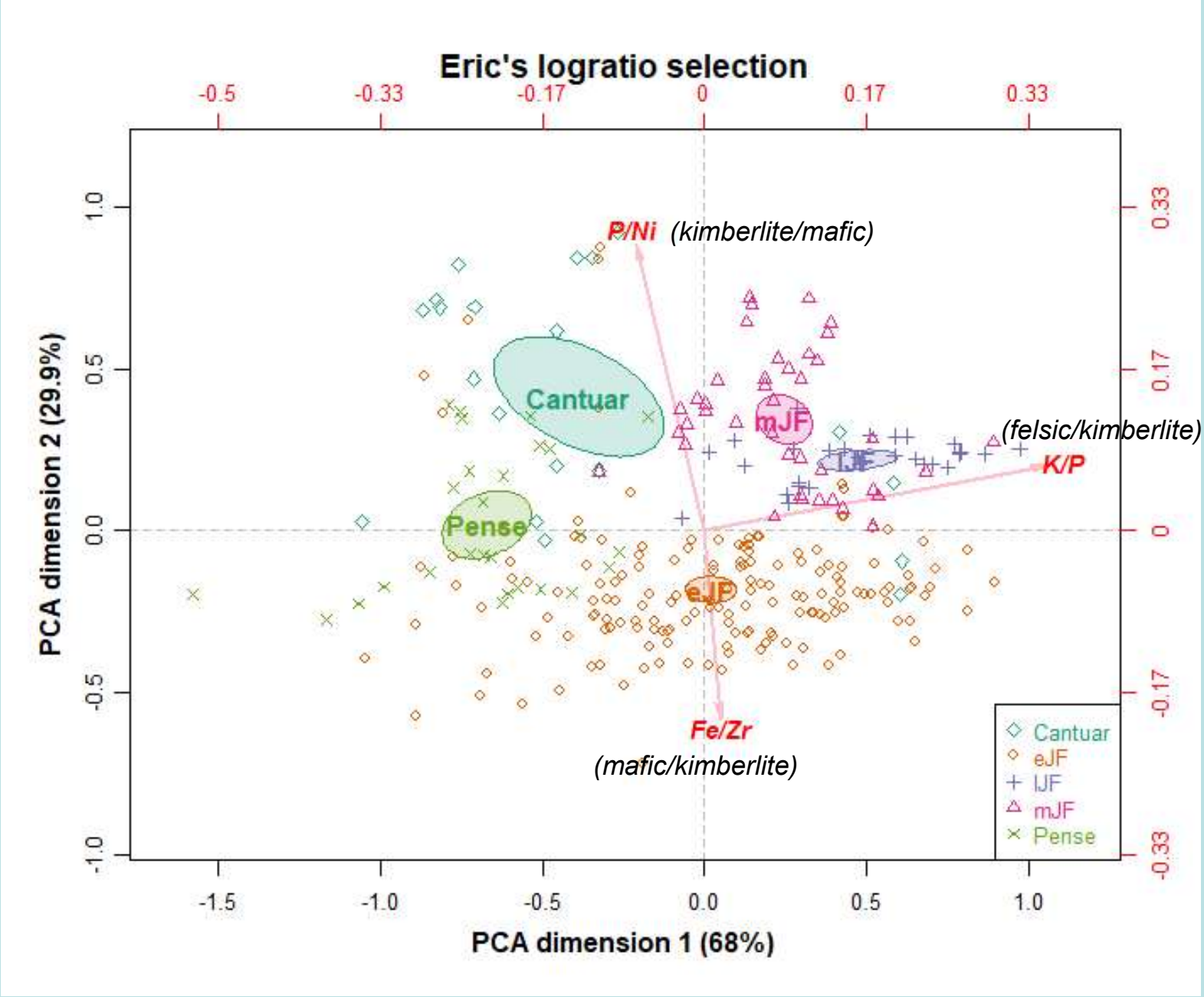
```
STEP 2: K/P, P/Ni and...

Ratio      R2     Procr
--------------------
Na/La     99.2%  0.947
Fe/Zr     99.2%  0.961
Nb/Th     99.2%  0.935
Zr/Ni     99.2%  0.969
K/Zr      99.2%  0.947
P/Zr      99.2%  0.956
P/Th      99.2%  0.959
Th/Ni     99.2%  0.961
K/Th      99.2%  0.937
Rb/Ga     99.2%  0.963
Nb/Ni     99.2%  0.955
P/Nb      99.2%  0.955
K/Nb      99.2%  0.933
Ca/Cr     99.2%  0.949
Th/Co     99.2%  0.961
```

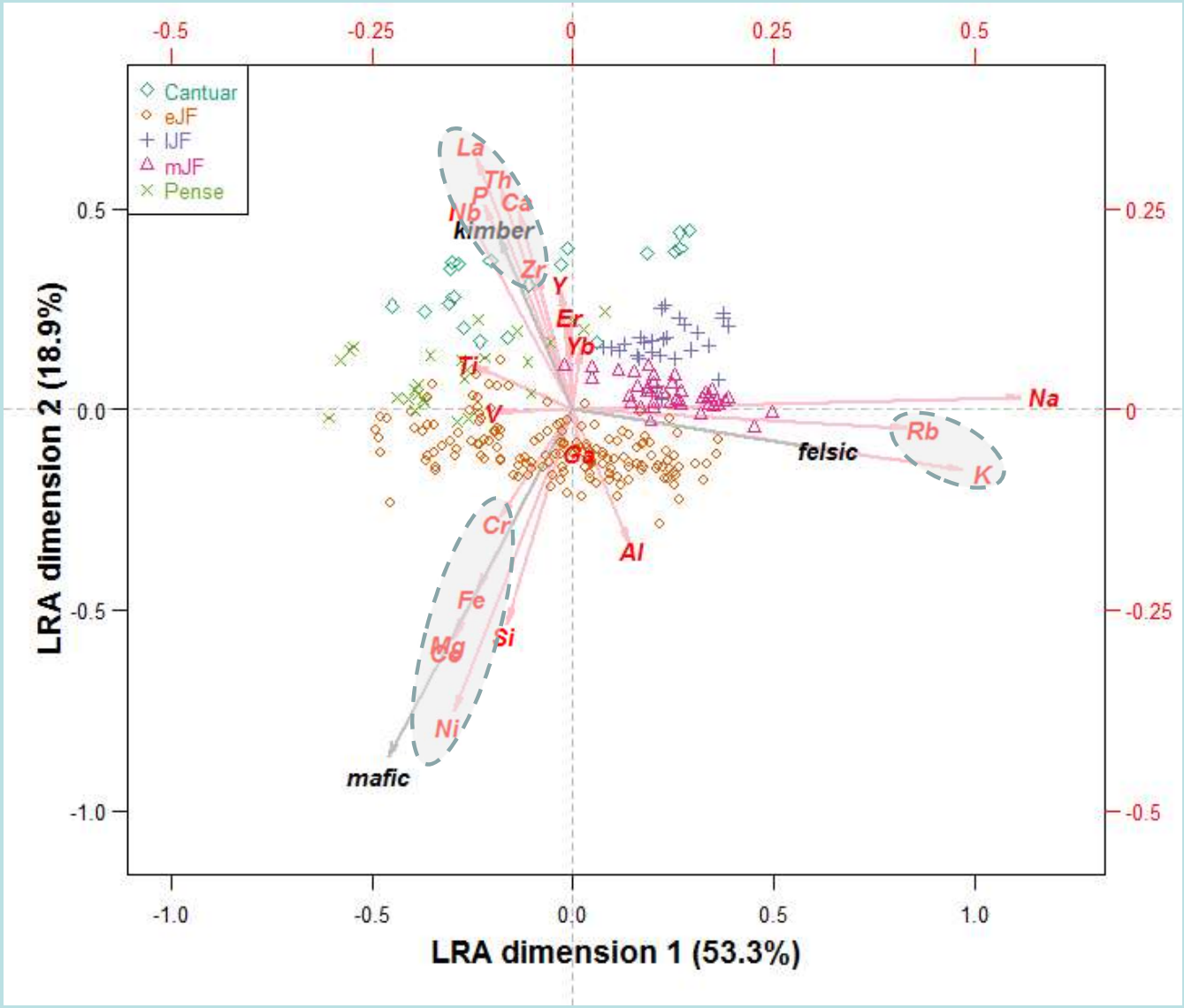

chosen by EG, will be included in next step, and so on... Actually only one more step needed to get to 100%!

```
STEP(kim.aggr, nsteps=1, top=15, weight=rep(1/ncol(kim), ncol(kim)),
     previous=cbind(log(kim.aggr[,"K"]/kim.aggr[,"P"]),
                    log(kim.aggr[,"P"]/kim.aggr[,"Ni"])))
```

Best 3 logratios chosen by EG (99.2% of between-group variance)

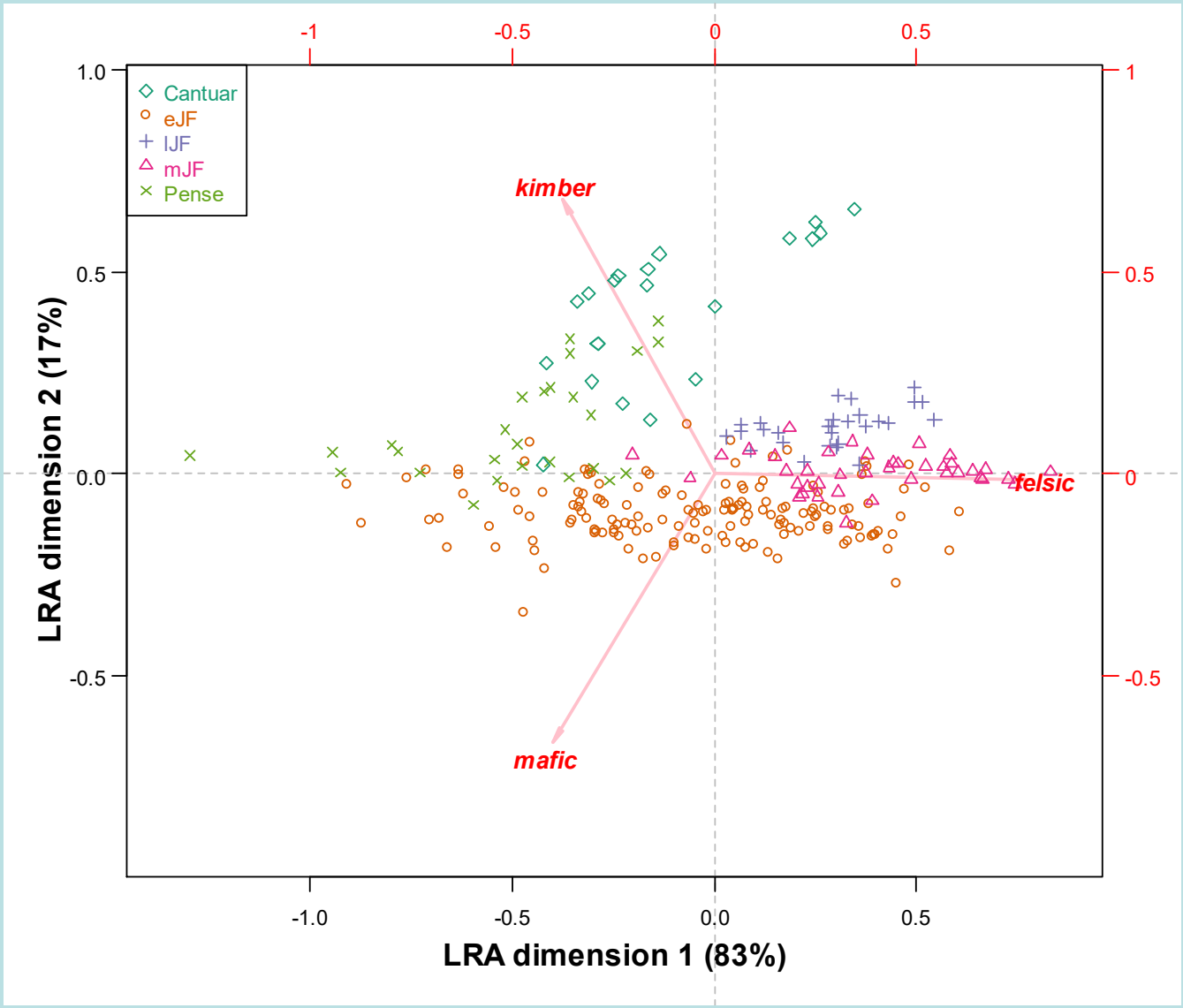


Adding supplementary variables, e.g. amalgamations



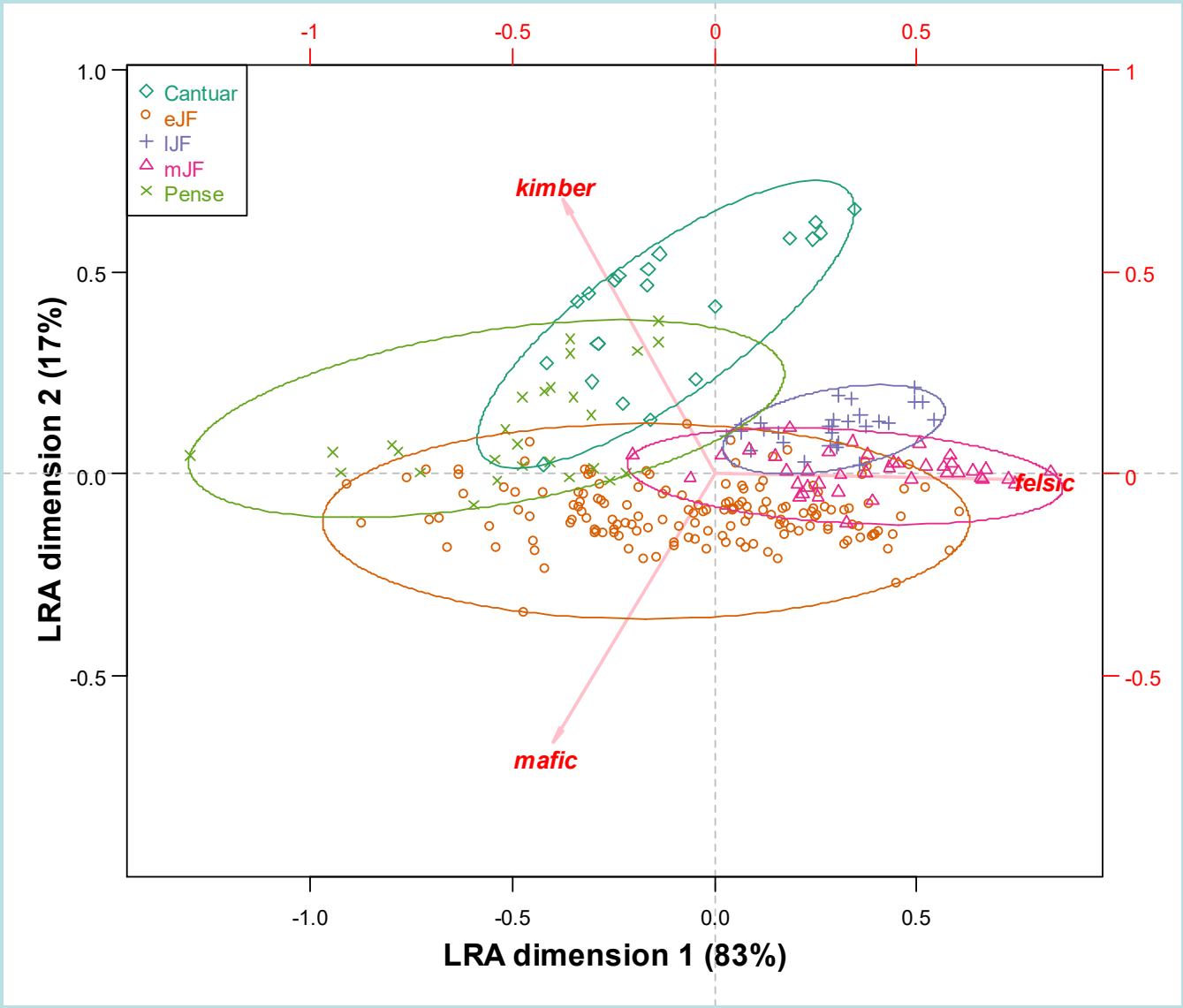

```
mafic <- kim[,"Fe"]+kim[,"Mg"]+kim[,"Co"]+kim[,"Cr"]+kim[,"Ni"]
felsic <- kim[,"K"]+kim[,"Rb"]
kimberlite <- kim[,"Nb"]+kim[,"La"]+kim[,"Th"]+kim[,"Zr"]+kim[,"P"]
```

Amalgamations defining “subcomposition”



Between
53.2%
Within
46.8%

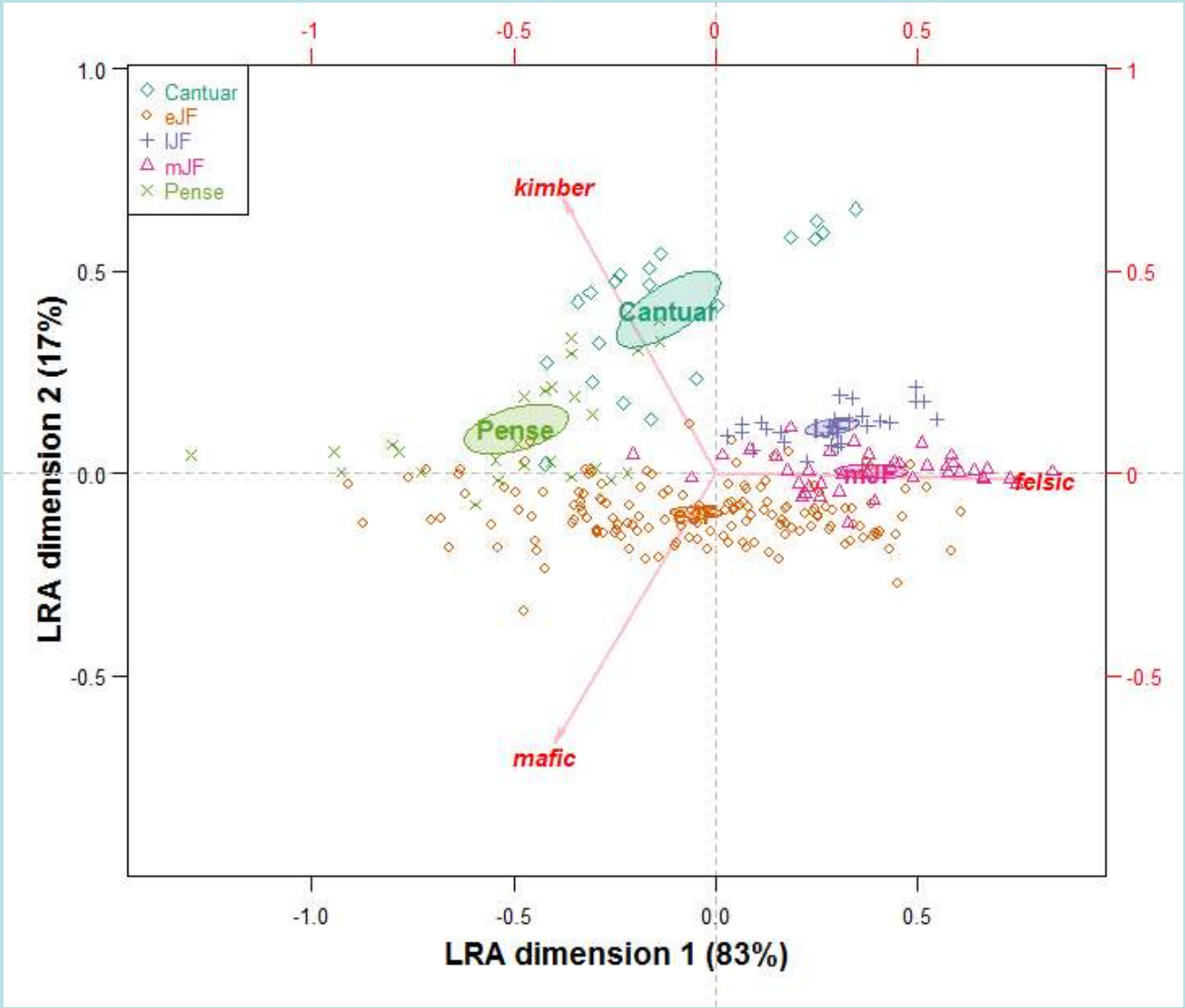
Amalgamations defining “subcomposition”



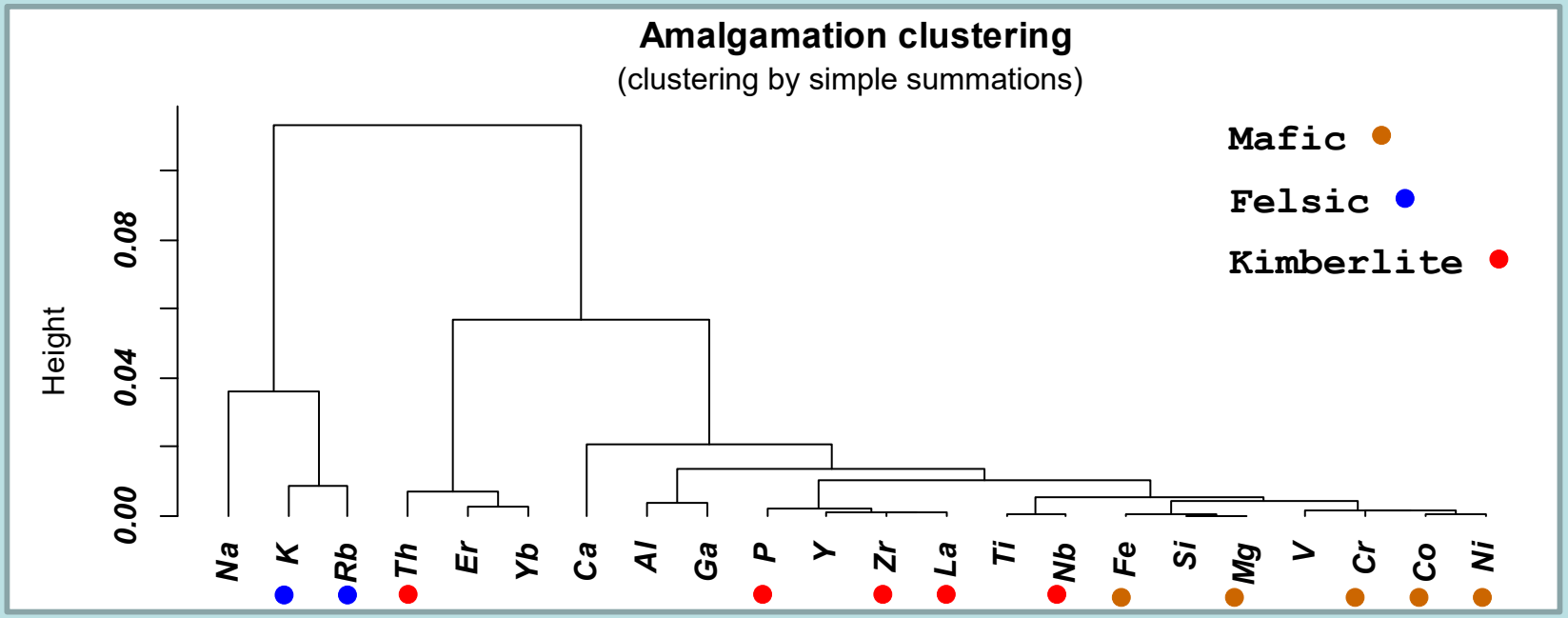
Between
53.2%
Within
46.8%

```
mafic <- kim[,"Fe"]+kim[,"Mg"]+kim[,"Co"]+kim[,"Cr"]+kim[,"Ni"]
felsic <- kim[,"K"]+kim[,"Rb"]
kimberlite <- kim[,"Nb"]+kim[,"La"]+kim[,"Th"]+kim[,"Zr"]+kim[,"P"]
```

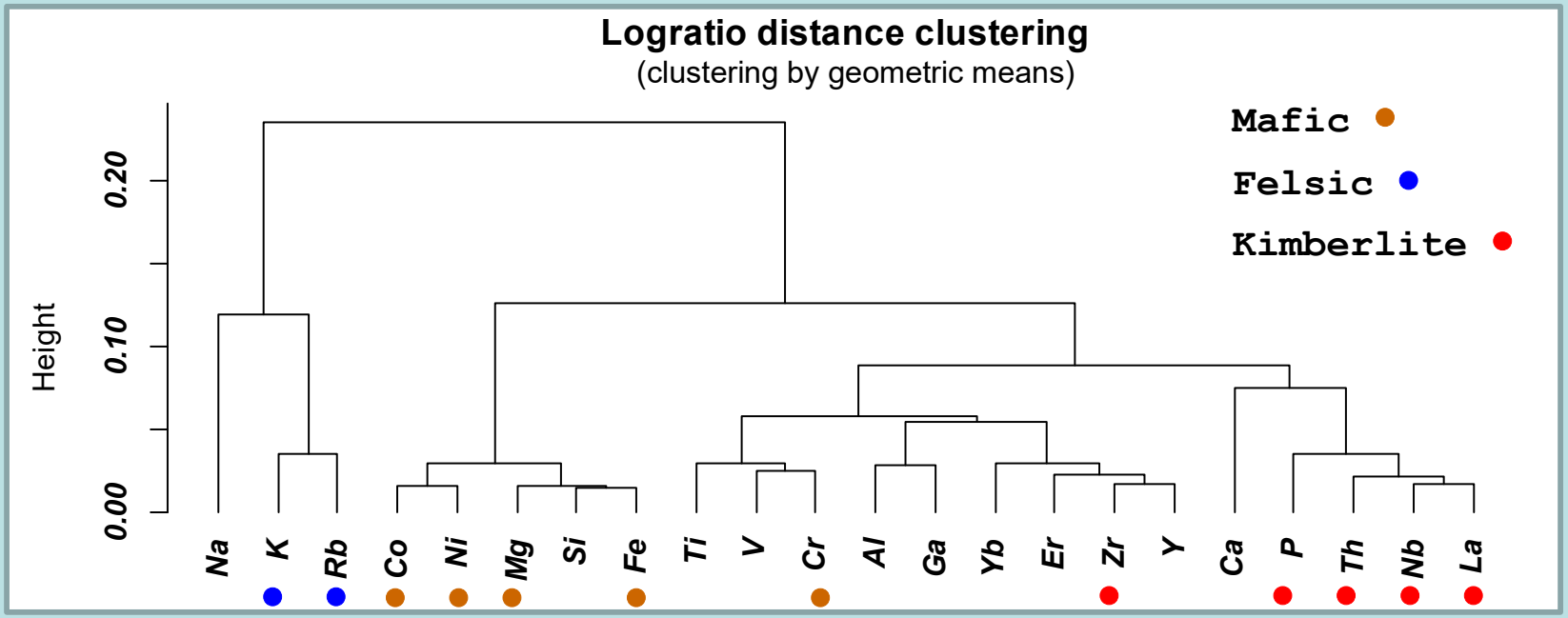
Amalgamations defining “subcomposition”



Alternative cluster analyses

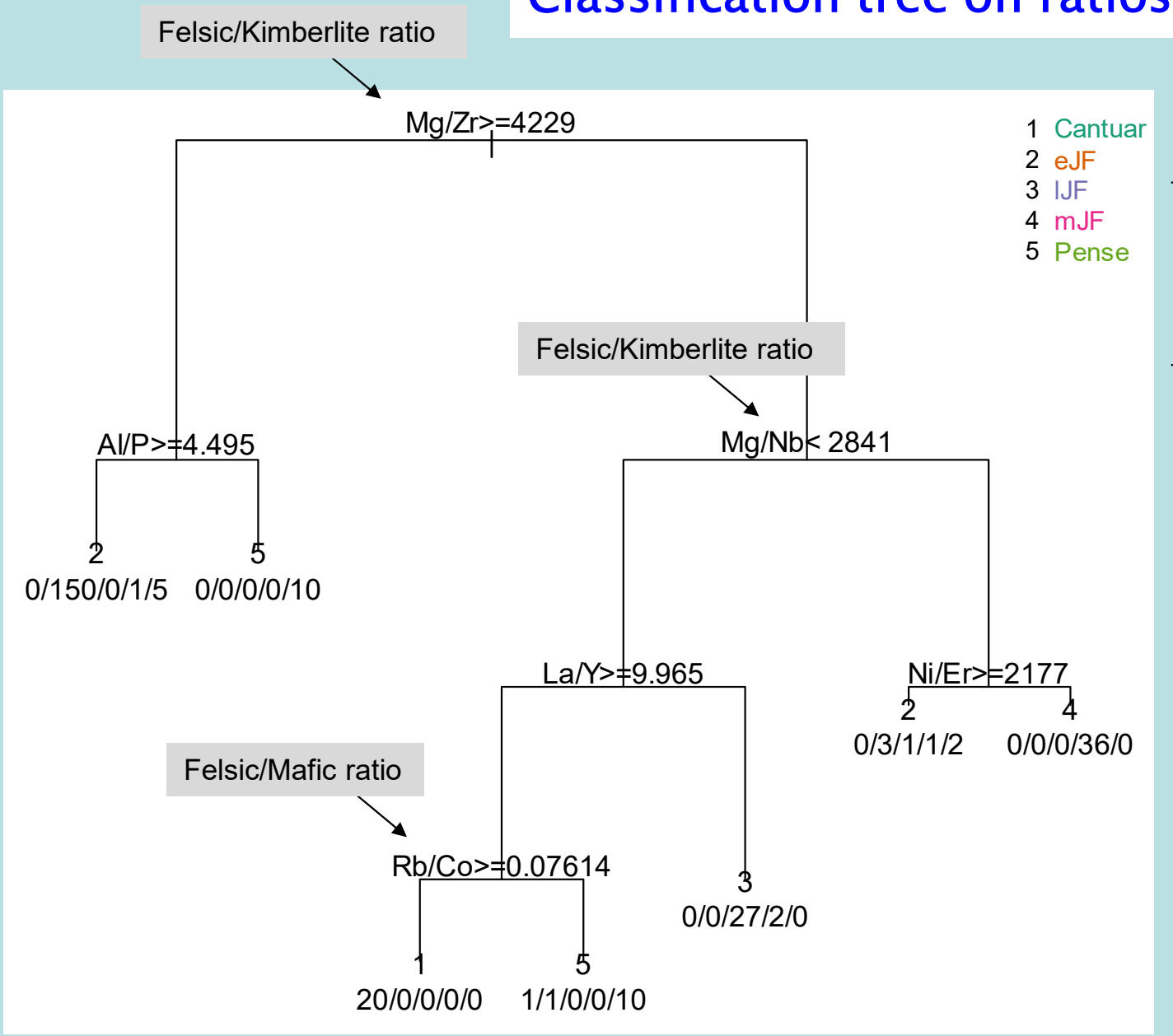


ACLUST() in easyCODA



WARD() in easyCODA (on columns not rows)

Classification tree on ratios



Classification matrix

	1	2	3	4	5
1	20	0	0	0	0
2	0	153	1	2	7
3	0	0	27	2	0
4	0	0	0	36	0
5	1	1	0	0	20

256 out of 283 correctly classified (94.8%)

e.g. predict lJF when
Mg/Zr > 4229 [39%]
& Mg/Nb < 2841 [24%]
& La/Y < 9.965 [68%]

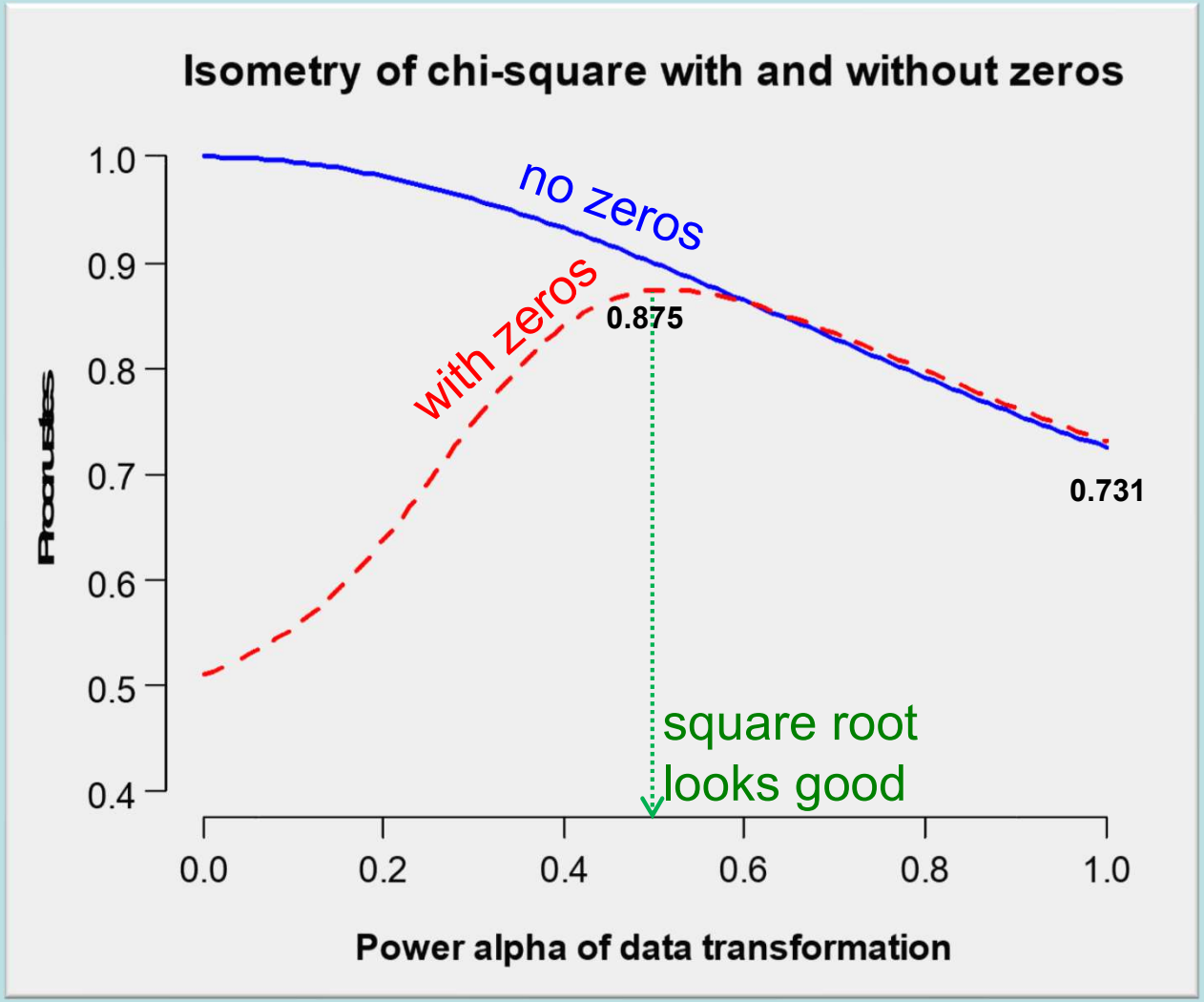
	2.5%	50%	97.5%
Mg/Zr	2222	5237	7768
Mg/Nb	1557	3743	5950
La/Y	5.96	8.85	13.7

LR() in easyCODA to compute all pairwise logratios
rpart() in rpart to compute the classification tree

Correspondence analysis as a way to deal with zeros

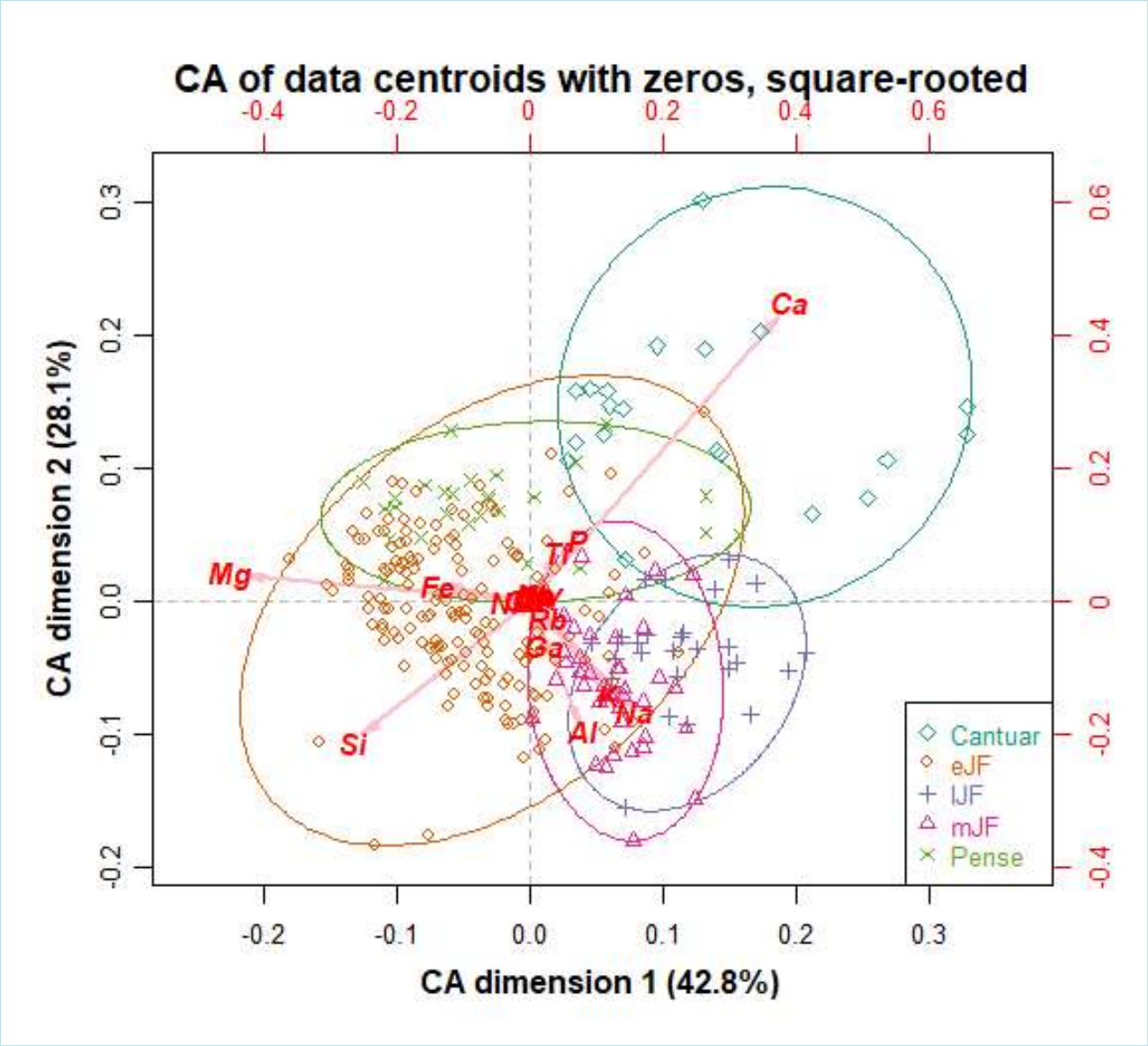
Correspondence analysis uses the chi-square distance, which has been shown to converge to the logratio distance when the data are transformed by the Box-Cox power transformation. This works for strictly positive data, but for data with zeros, the transformation can get closer to the logratio distance and then starts to break down for lower powers as the log-transformation approaches.

Here the Procrustes correlation measures how close the CA spatial configuration of samples is close to the LRA one, i.e. how isometric CA is.



Correspondence analysis of square-rooted data, with zeros

The benefit of the CA approach is that some isometry is traded off against the ability to not have to replace the zeros. CA gives more weight to the more abundant parts, but this is reduced by the root transformation.



Correspondence analysis of square-rooted data, with zeros

Simplified visualization, omitting sample points and showing only the confidence ellipses for their means. A further simplification is to remove the parts near the centre, which contribute very little to the solution.