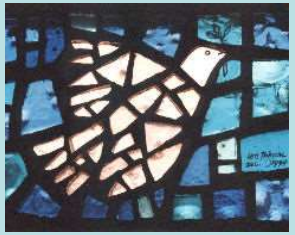
IUGS-CGGB International Workshop, 18 November 2021

Compositional Data Analysis: Multivariate Analysis

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

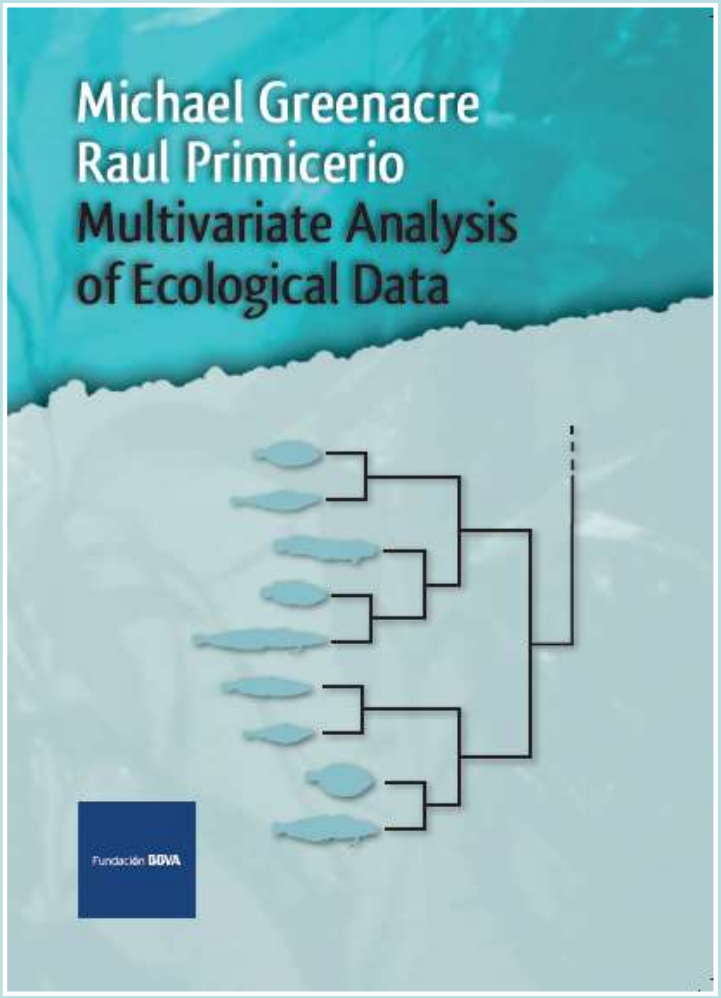


www.globalsong.net

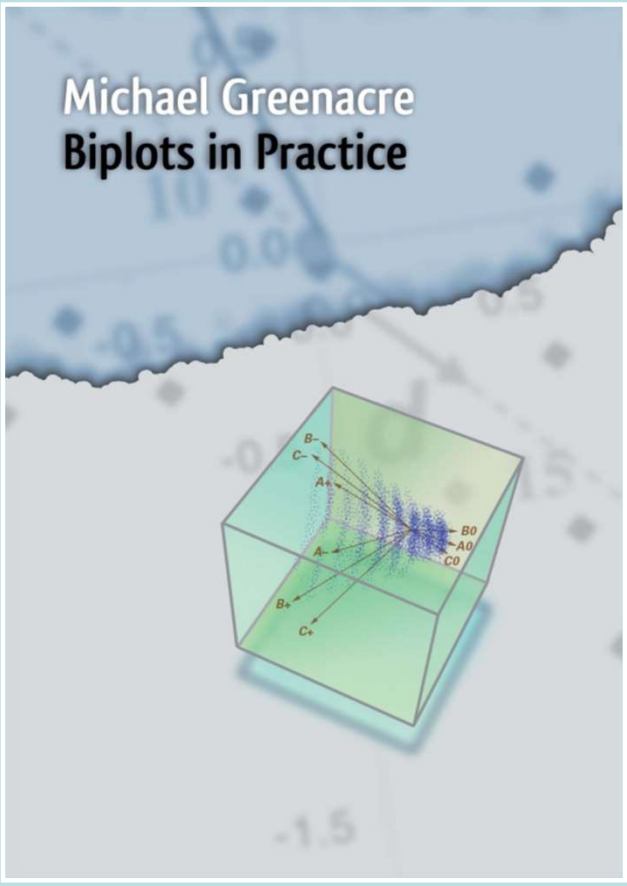
www.multivariatestatistics.org
https://github.com/michaelgreenacre/CODAinPractice

Email: michael.greenacre@upf.edu

2008



Published December 2013



2010

These books are available for
free download from
www.multivariatestatistics.org
thanks to the BBVA Foundation

Statistics

Chapman & Hall/CRC
Interdisciplinary Statistics Series

"...an essential reference for evaluating and interpreting compositional data across a broad spectrum of disciplines in the life and natural sciences for both academia and industry...takes a prescribed approach starting with the definition of compositional data, the use of logratios for dimension reduction, clustering and variable selection, along with several practical examples and a case study...using the methods described in this book will help to avoid costly mistakes made from misinterpreting compositional data."
—**Professor Eric Grunsky**, University of Waterloo, Canada

Compositional Data Analysis in Practice is a user-oriented practical guide to the analysis of data with the property of a constant sum, for example percentages adding up to 100%. Compositional data can give misleading results if regular statistical methods are applied, and are best analysed by first transforming them to logarithms of ratios. This book explains how this transformation affects the analysis, results and interpretation of this very special type of data. All aspects of compositional data analysis are considered: visualization, modelling, dimension-reduction, clustering and variable selection, with many examples in the fields of food science, archaeology, sociology and biochemistry, and a final chapter containing a complete case study using fatty acid compositions in ecology. The applicability of these methods extends to other fields such as linguistics, geochemistry, marketing, economics and finance.

Features

- **A unique didactic format**, where each chapter has exactly eight pages of study material, many illustrative figures, and an end-of-chapter summary
- **An approach aimed at students and applied researchers**, gathering the mathematical aspects in a compact theoretical appendix
- **Numerous examples** from a variety of disciplines
- **A computational appendix** that documents the **easyCODA** package for R developed by the author, making it possible for readers to reproduce the results
- **A supporting website** with data sets, R scripts and further study material

Michael Greenacre is Professor of Statistics at the Universitat Pompeu Fabra, Barcelona, Spain, where he teaches a course, amongst others, on Data Visualization. He has authored and co-edited ten books and more than a hundred journal articles and book chapters, mostly on correspondence analysis, the latest being *Correspondence Analysis in Practice (Third Edition)* in 2016. He has given short courses in fifteen countries to environmental scientists, sociologists, data scientists and marketing professionals, and has specialized in statistics in ecology and social science.

CRC Press
Taylor & Francis Group
an Informa business
www.crcpress.com

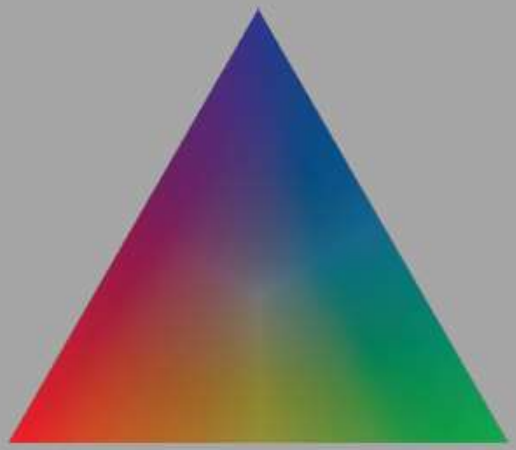
6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K391345
ISBN: 978-1-138-31643-0
9 781138 316430 90000
www.crcpress.com

COMPOSITIONAL DATA ANALYSIS IN PRACTICE Greenacre CRC Press

Chapman & Hall/CRC
Interdisciplinary Statistics Series

COMPOSITIONAL DATA ANALYSIS IN PRACTICE



Michael Greenacre

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

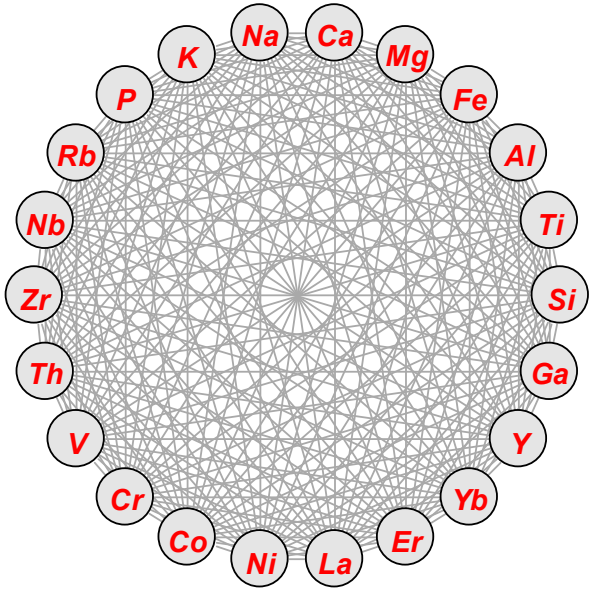
Published July 2018

You can order a copy at a **30% discount** by giving the promotion code **ASA21** when ordering.

Danie Krige receiving certificate of honorary membership of Statistical Association of South Africa by the association president at the time (that's me, folks!). About mid 1980s...

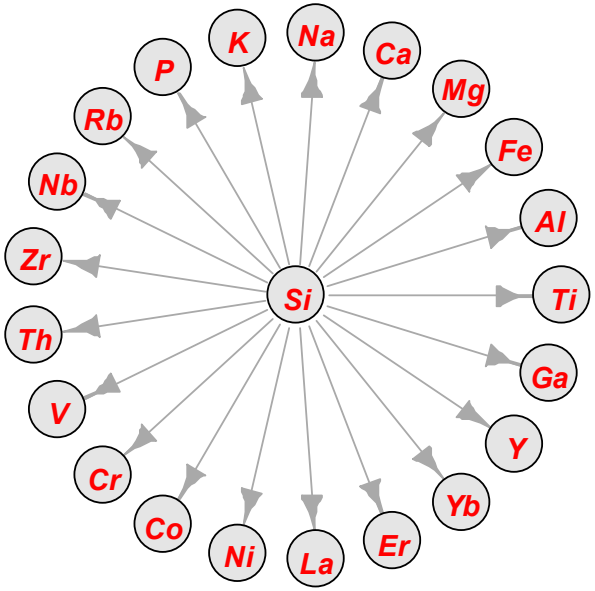
Types of logratios
(graph representation)

All pairwise logratios



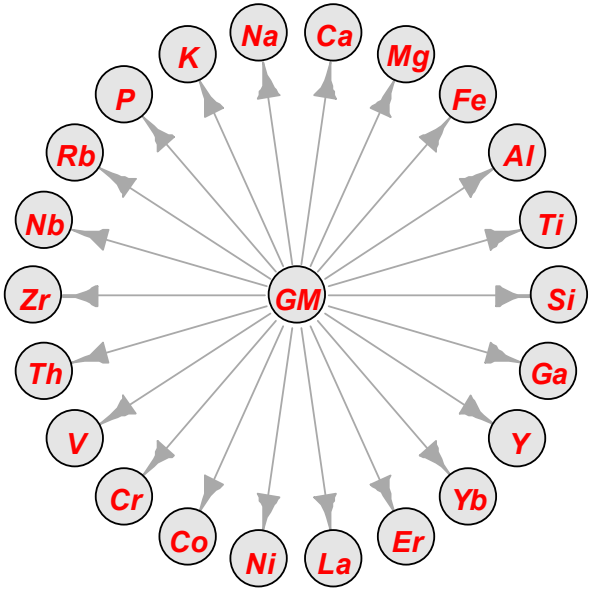
$$\#\text{ ratios} = J(J-1)/2 = 22\times 21/2 = 231$$

Additive logratios



$$\#\text{ ratios} = J-1 = 21$$

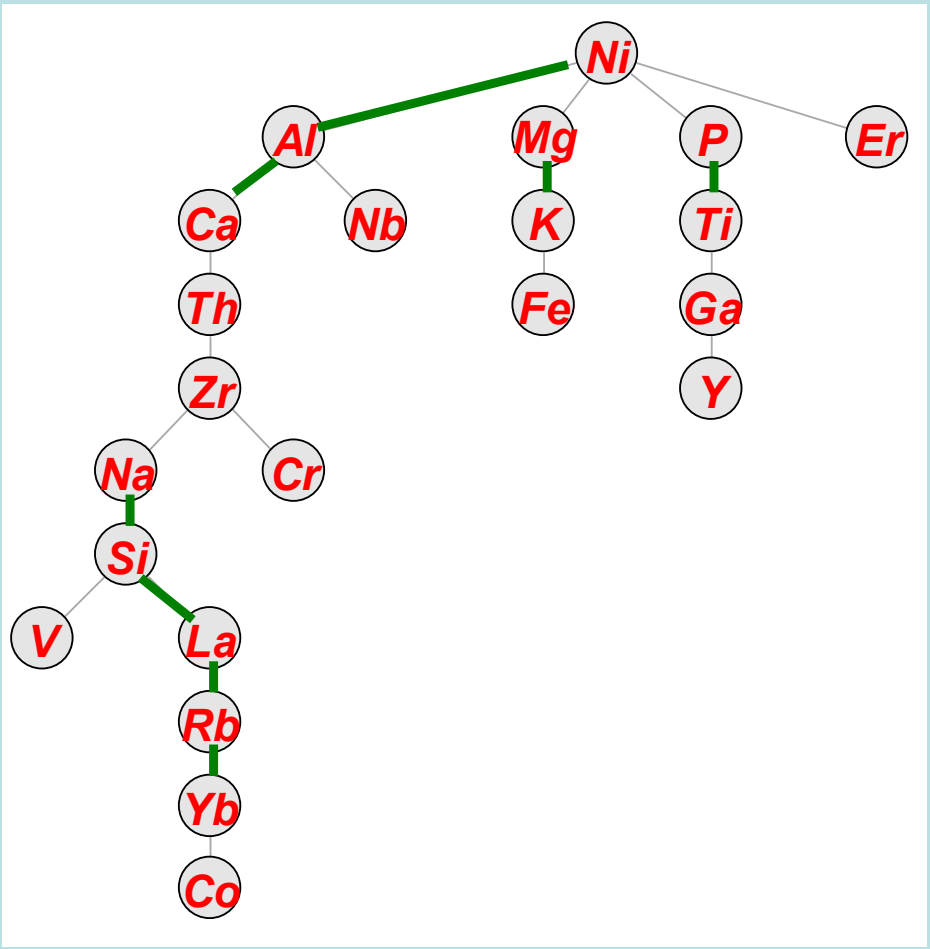
Centred logratios
w.r.t. geometric mean (GM)



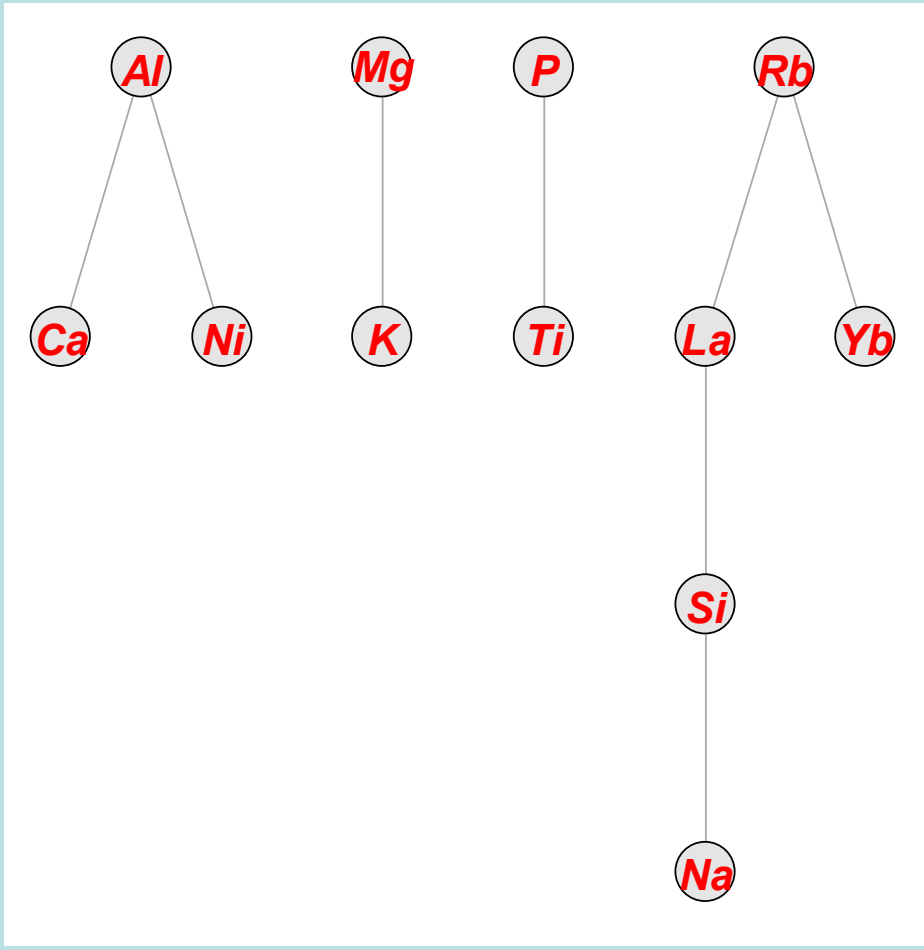
$$\#\text{ ratios} = J = 22$$

Types of logratios
(graph representation)

Acyclic connected graph



Subgraph



Acyclic connected graphs explain 100% of the data variance (21 edges, i.e. logratios, of the 22 elements). Or **spanning tree.** J^{J-2} such graphs! Needs to be chosen well.

A well-selected **subgraph** (i.e. subset of 8 of the 21 ratios on the left) explains 95% of the variance

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)

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
:	:	:
:	:	:

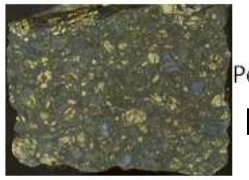
close →

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**

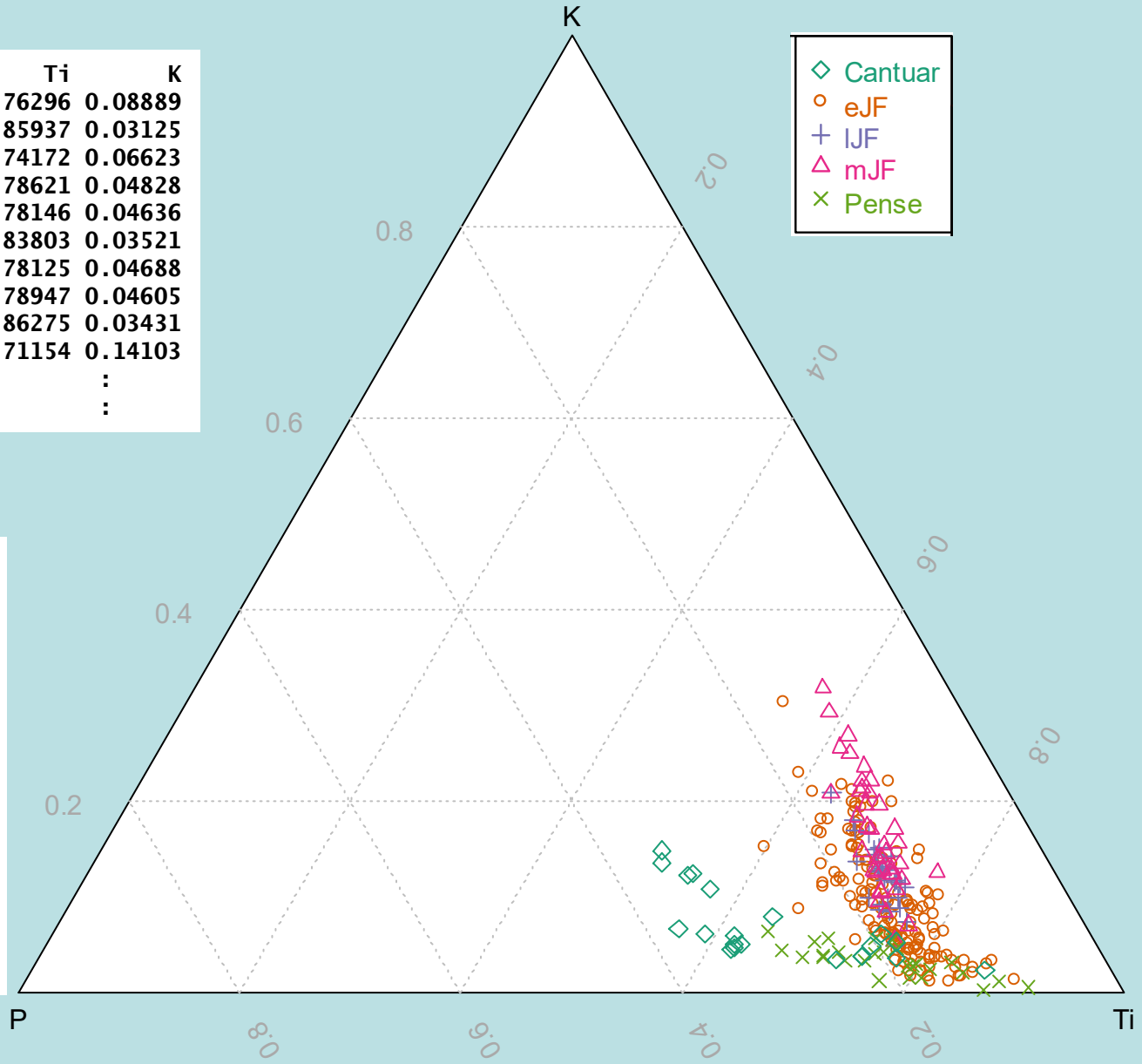


Pense

Pense

Cantuar

Cantuar

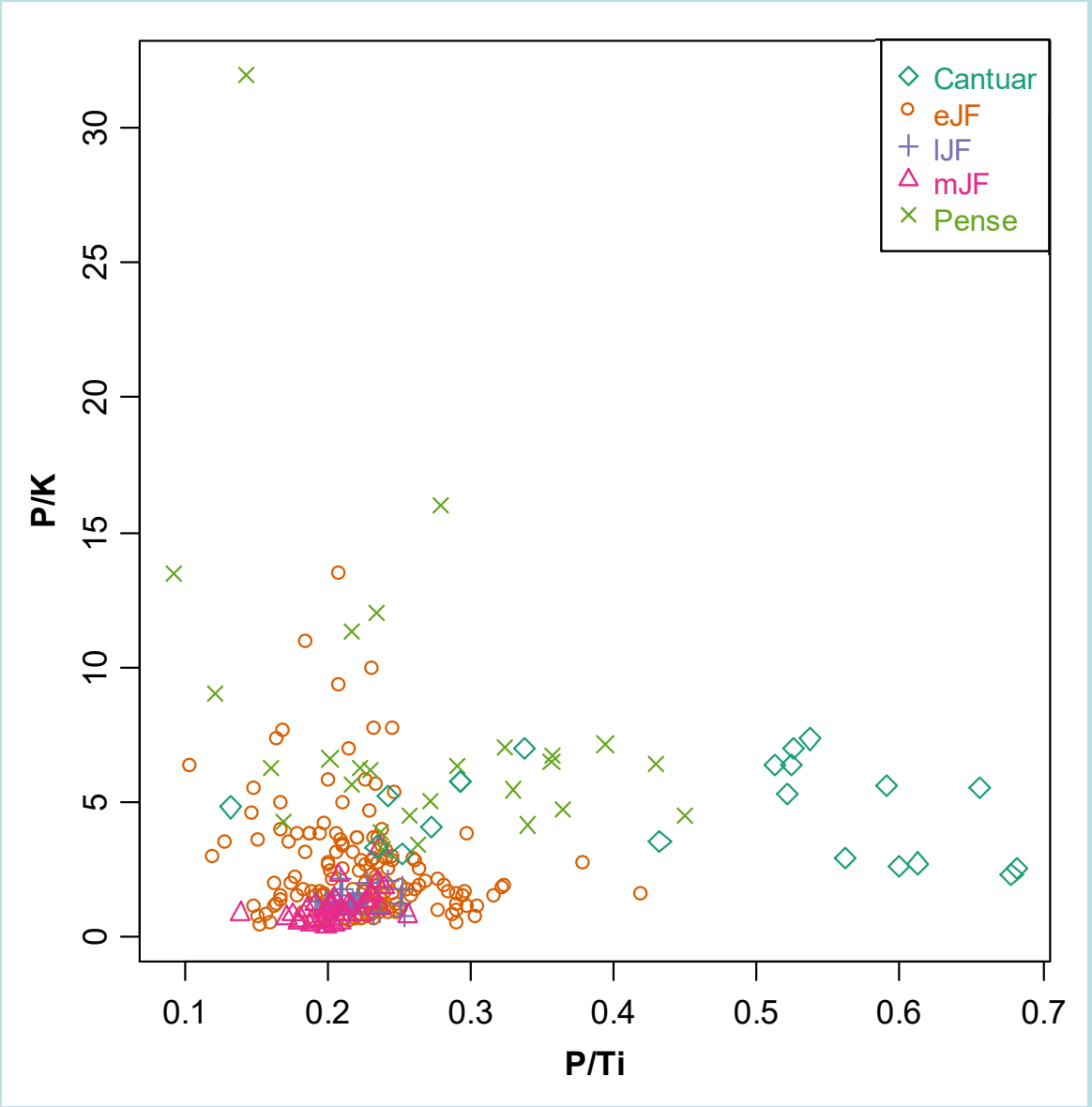


Three-part (sub)composition as ratios

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
:	:	:
:	:	:

ratio

P/Ti	P/K	Ti/K
0.1942	1.6667	8.5833
0.1273	3.5000	27.5000
0.2589	2.9000	11.2000
0.2105	3.4286	16.2857
0.2203	3.7143	16.8571
0.1513	3.6000	23.8000
0.2200	3.6667	16.6667
0.2083	3.5714	17.1429
0.1193	3.0000	25.1429
0.2072	1.0455	5.0455
:	:	:
:	:	:

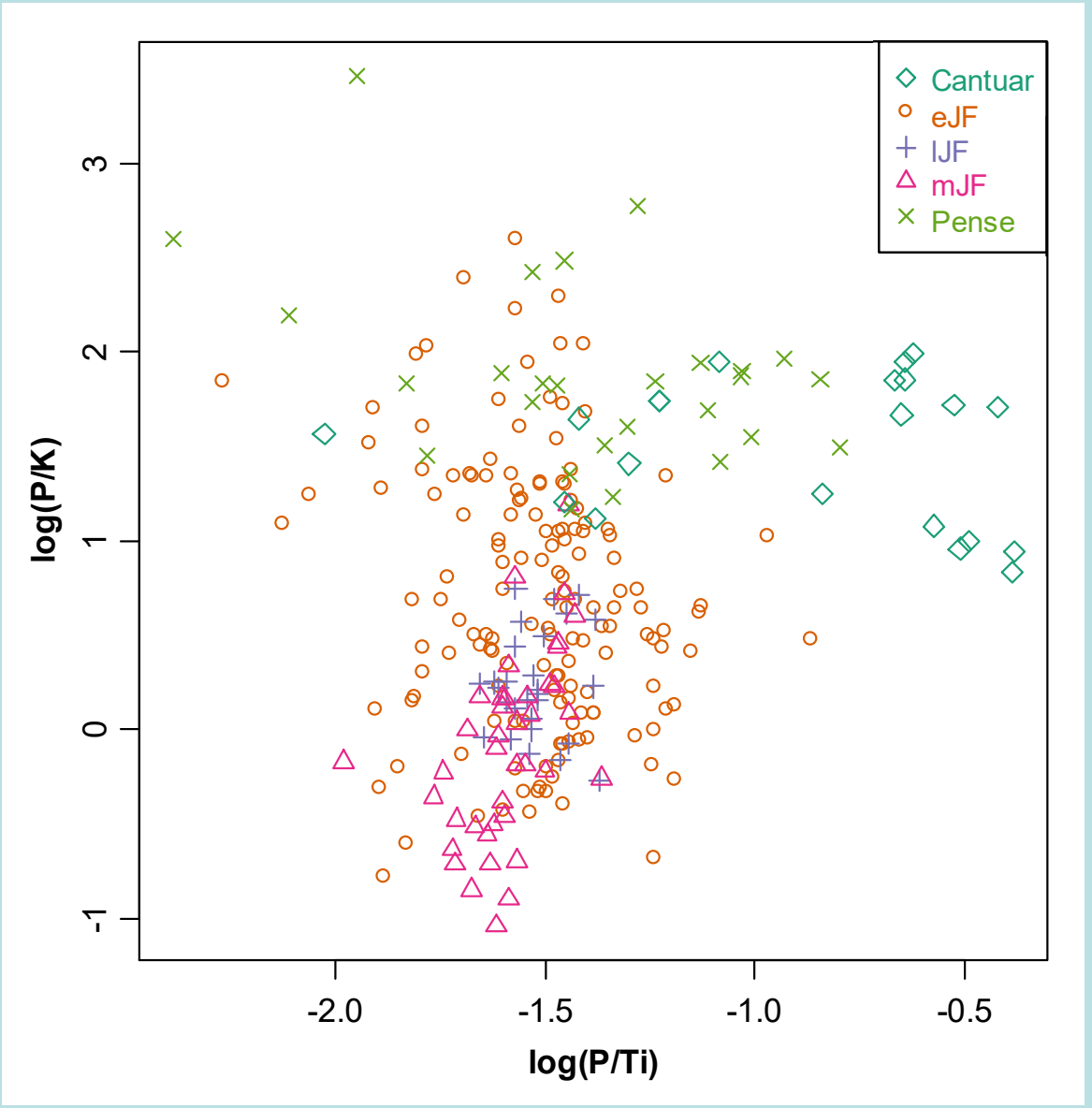


Three-part (sub)composition as logratios

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
:	:	:
:	:	:

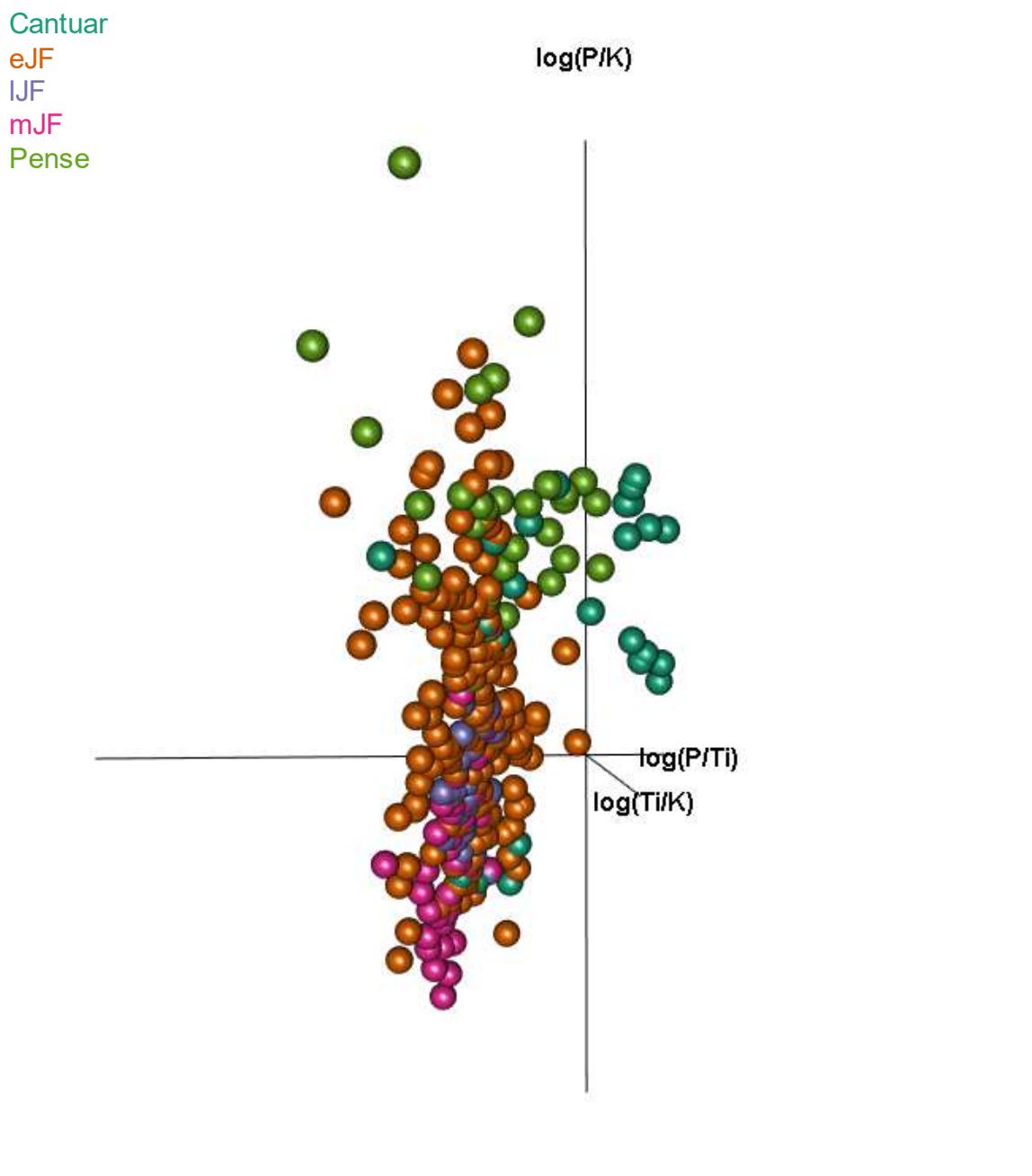
logratio

P/Ti	P/K	Ti/K
-1.63900	0.51083	2.14982
-2.06142	1.25276	3.31419
-1.35120	1.06471	2.41591
-1.55814	1.23214	2.79029
-1.51259	1.31219	2.82477
-1.88875	1.28093	3.16969
-1.51413	1.29928	2.81341
-1.56862	1.27297	2.84158
-2.12596	1.09861	3.22457
-1.57404	0.04445	1.61849
:	:	:
:	:	:



Note the aspect ratio!!

Three-part composition as logratios

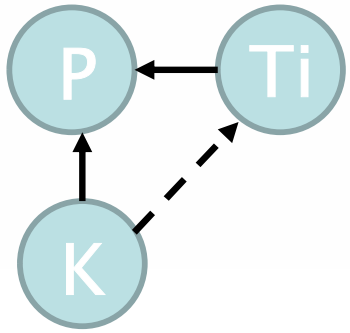


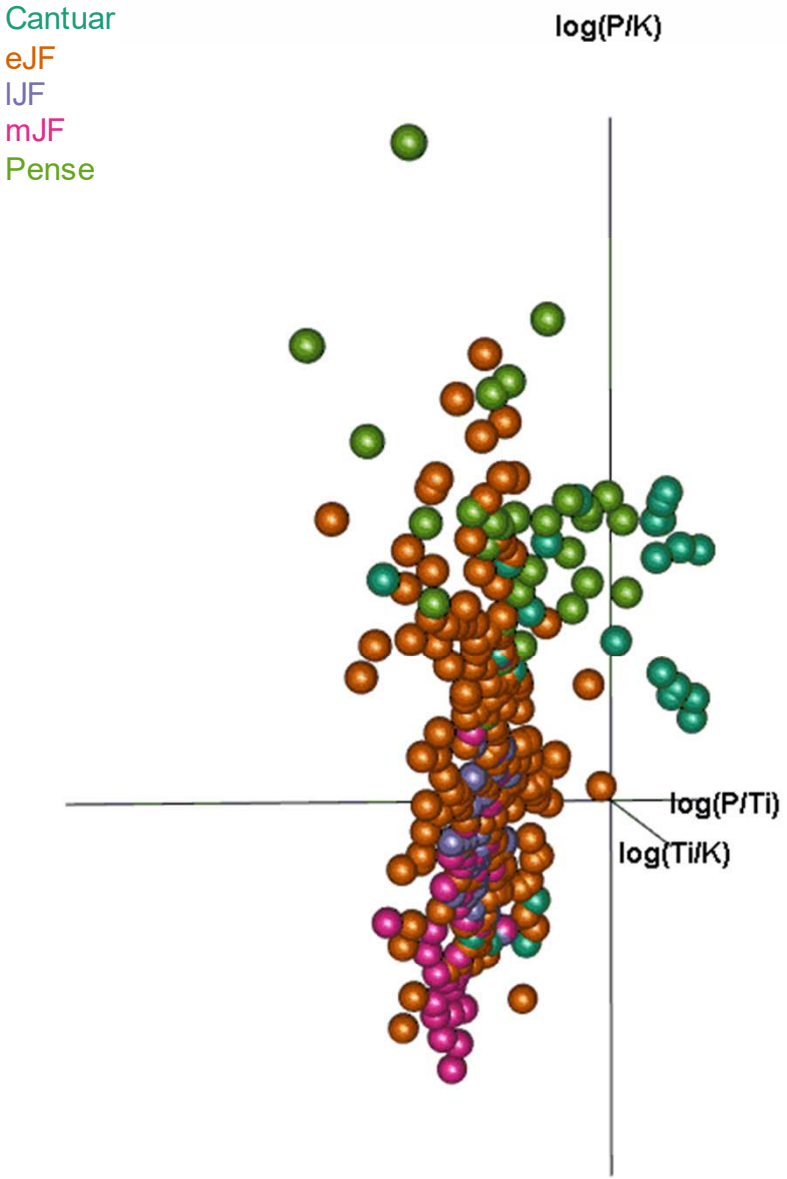
P	Ti	K
0.01863	0.09596	0.01118
0.01377	0.10816	0.00393
0.02864	0.11063	0.00988
0.02399	0.11395	0.00700
0.02674	0.12135	0.00720
0.01763	0.11653	0.00490
0.02578	0.11720	0.00703
0.02364	0.11348	0.00662
0.01876	0.15723	0.00625
0.02267	0.10942	0.02169
:	:	:
:	:	:

logratio

P/Ti	P/K	Ti/K
-1.63900	0.51083	2.14982
-2.06142	1.25276	3.31419
-1.35120	1.06471	2.41591
-1.55814	1.23214	2.79029
-1.51259	1.31219	2.82477
-1.88875	1.28093	3.16969
-1.51413	1.29928	2.81341
-1.56862	1.27297	2.84158
-2.12596	1.09861	3.22457
-1.57404	0.04445	1.61849
:	:	:
:	:	:

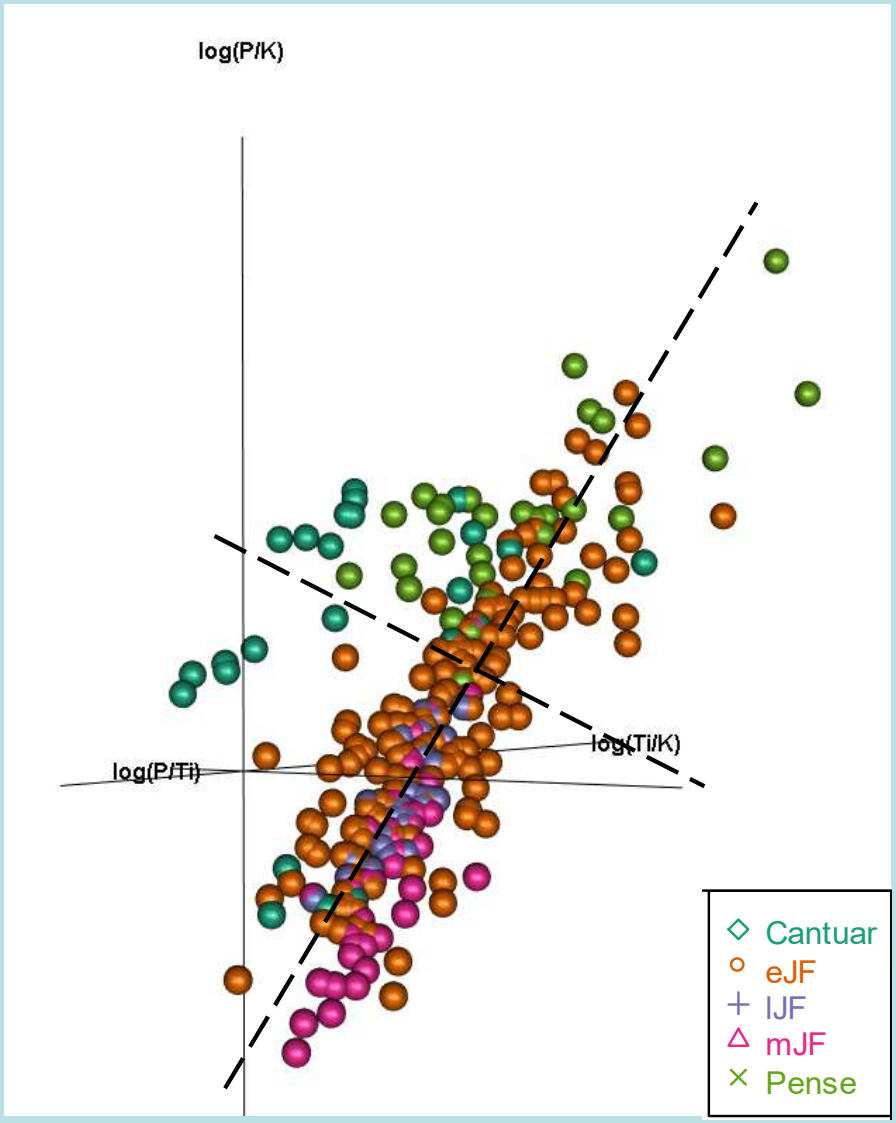
$$\log(P/K) = \log(P/Ti) + \log(Ti/K)$$
$$= \log(P) - \log(Ti) + \log(Ti) - \log(K)$$



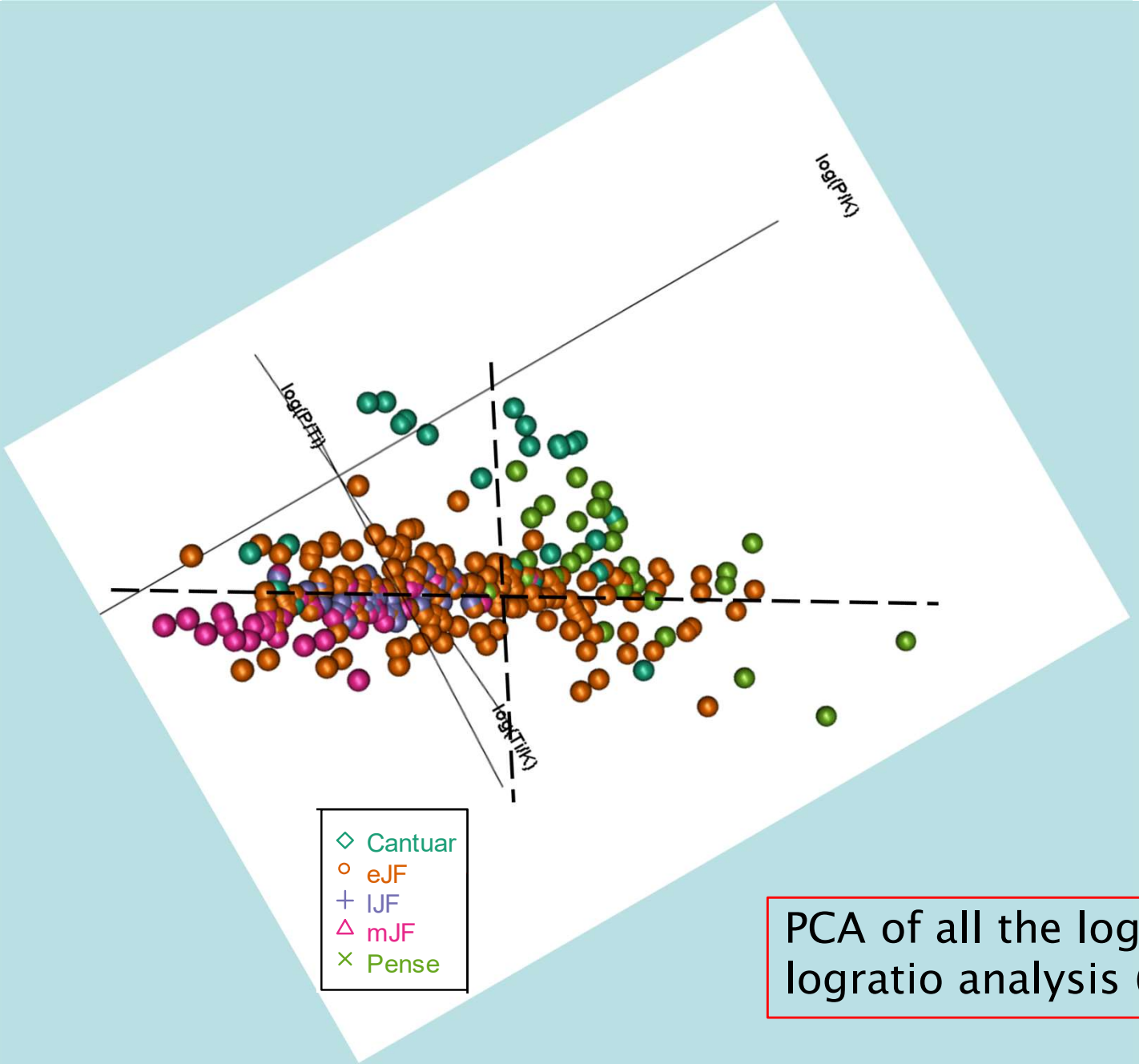


To see animation see additional PPS file

Principal component analysis (PCA)

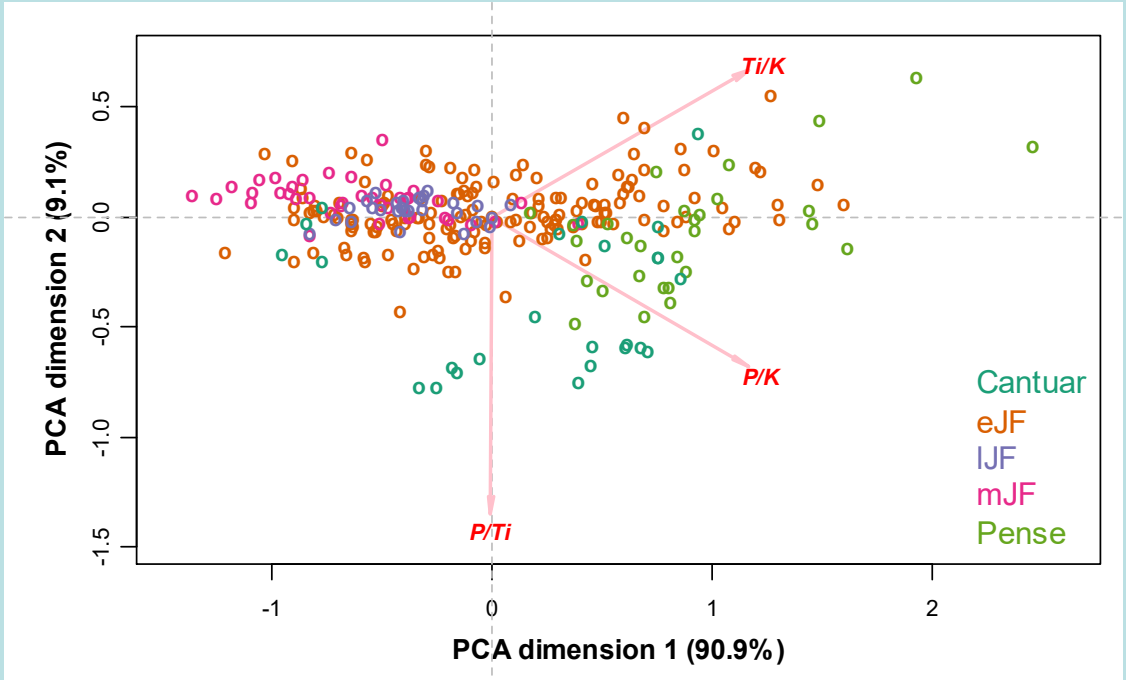


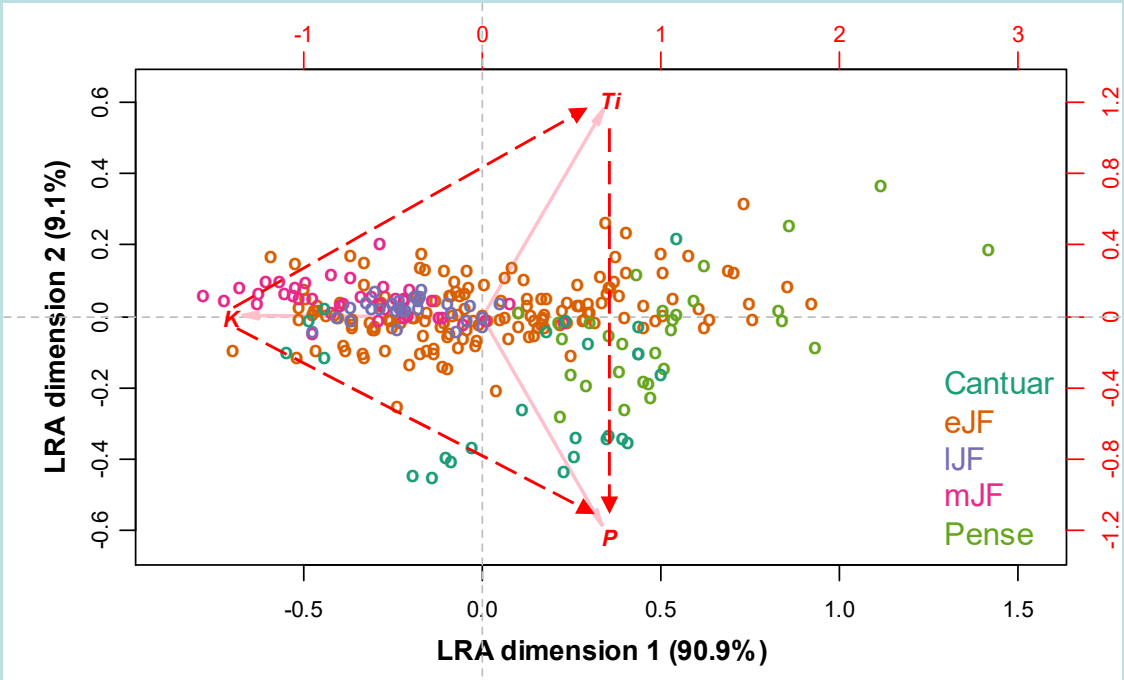
Principal component analysis (PCA)



PCA of all the logratios = logratio analysis (LRA)

PCA and LRA* of 3-part subcomposition



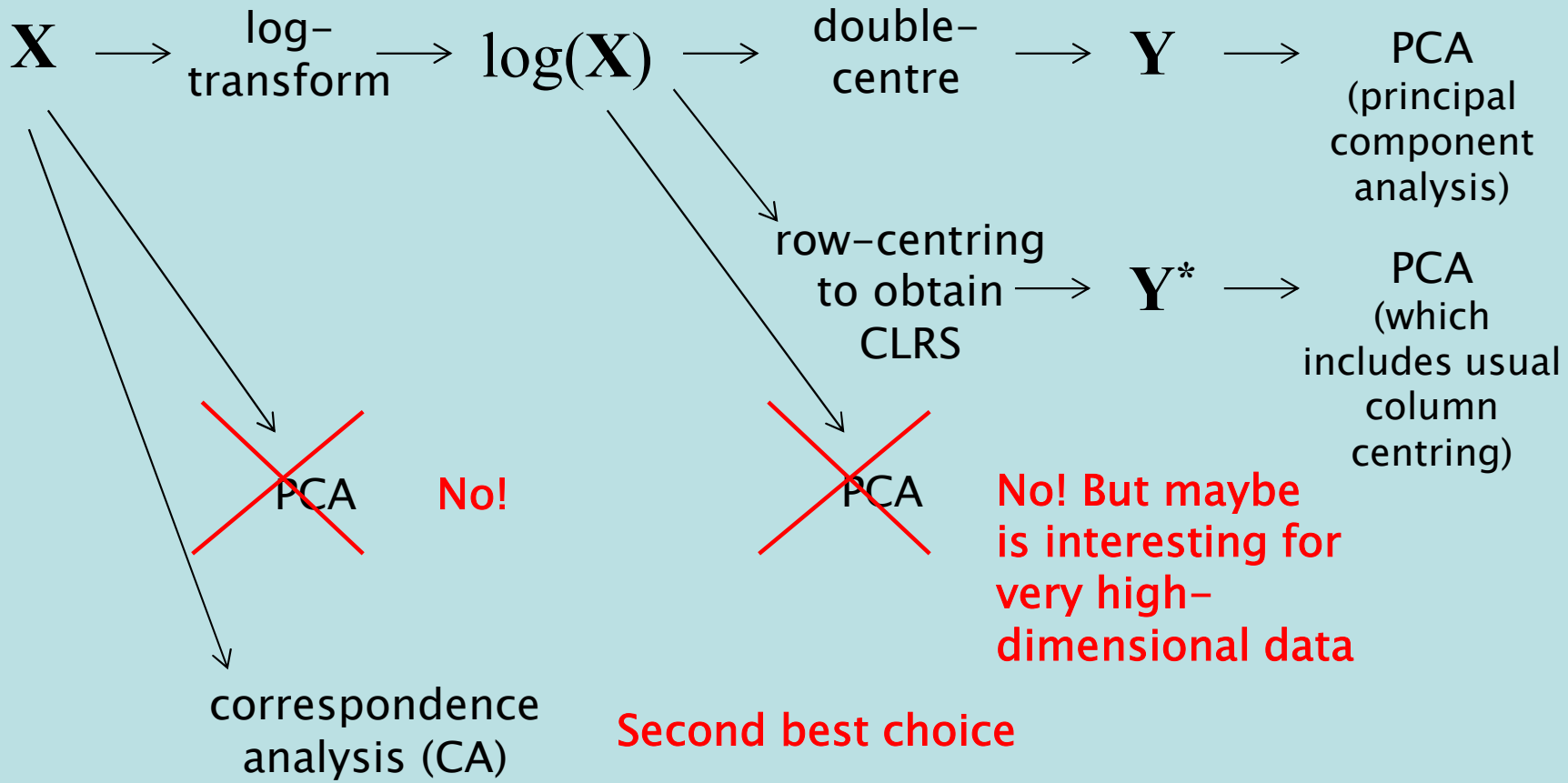


*logratio analysis

- Three-part composition has three ratios $=3\times2/2$
- The configurations of the samples are identical (up to a scaling constant)
- PCA shows the logratios, LRA shows the parts
- The way to interpret the LRA is via the links between the parts that represent the logratios
- Proximity of parts is only interesting in that they induce highly correlated logratios with other parts

Logratio analysis (LRA)

The objective of logratio analysis is to analysis all the logratios – and there are many! But there is a considerable computational short-cut.
The mathematical key is the double-centring, which allows us to analyse a matrix of the original size.



Logratio distance: subcompositionally coherent

- Samples-by-parts compositional data matrix $\mathbf{X}$ $(I \times J)$.
- ***Log-ratio distance*** between samples i and i' :

$$d_{ii'}^2 = \frac{1}{J^2} \sum\sum_{j<j'} \left[\log\left(\frac{x_{ij} x_{i'j'}}{x_{ij'} x_{i'j}} \right) \right]^2 = \frac{1}{J^2} \sum\sum_{j<j'} \left[\log\left(\frac{x_{ij}}{x_{ij'}} \right) - \log\left(\frac{x_{i'j}}{x_{i'j'}} \right) \right]^2$$

clearly coherent

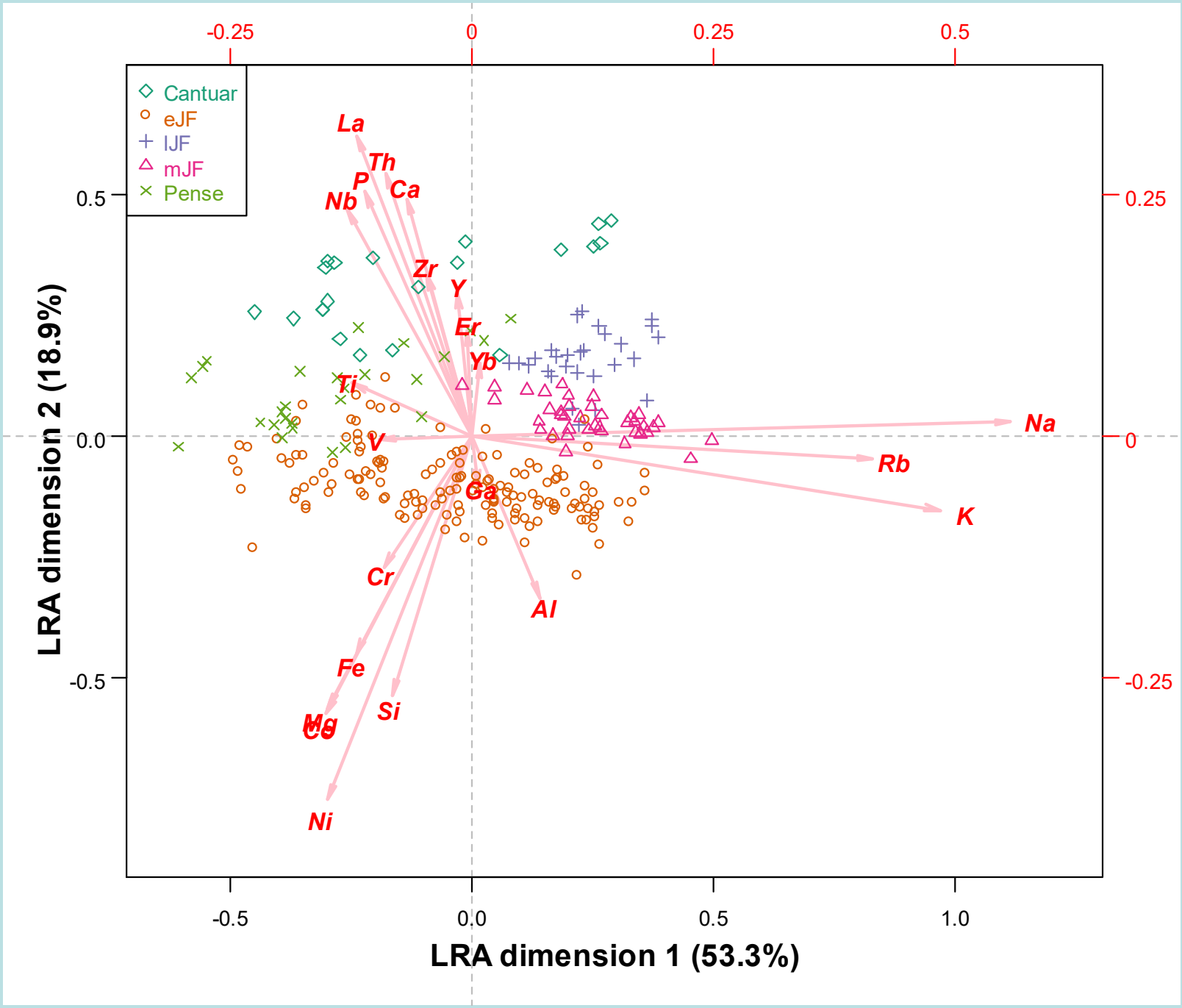
type of odds ratio

$$d_{ii'}^2 = \frac{1}{J} \sum_{i=1}^{I} \left[\log\left(\frac{x_{i'j}}{g(\mathbf{x}_i)} \right) - \log\left(\frac{x_{i'j}}{g(\mathbf{x}_{i'})} \right) \right]^2$$

(compact form, using CLRs)

- For distances between parts, just interchange i and j

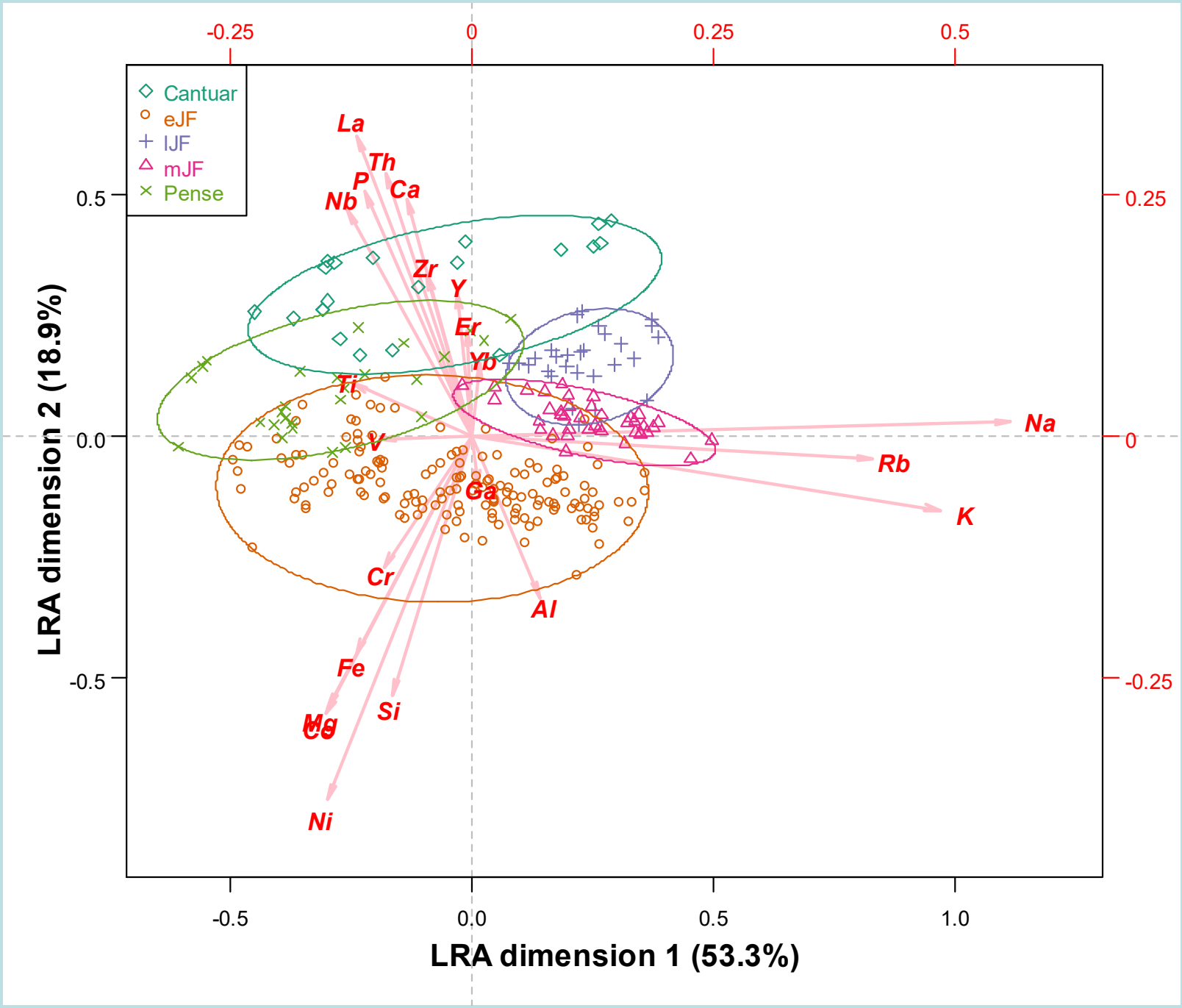
Logratio biplot



Between
57.9%
Within
42.1%

Logratio biplot

with minimum covering ellipses around groups



Between
57.9%
Within
42.1%

Total variance and its explanation

- In the study of a compositional data set, the amount of information contained in the data is measured by the (total) logratio variance.
- Although there are $J(J-1)/2$ logratios, only $J-1$ independent ones are needed to explain all the variance. All the others would be linearly related to this reduced set.
- An optimal set of logratios can be selected in a stepwise fashion to partially or completely explain the logratio variance*.
- This selection can be guided by expert knowledge, since at any step, many logratios will be competing to be selected with almost the same level of explained variance.
- Otherwise, the process can be automatic, where the statistically optimal logratio is selected at each step.
- Logratios that explain the individual total variance (unsupervised) or the between-group variance (supervised) are possible.

* Greenacre, M (2019) Variable selection in compositional data analysis, using pairwise logratios. *Mathematical Geosciences*,

Stepwise selection of Kimberlite ratios

Logratio analysis using all the 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		

Total variance (inertia)

Decomposition of total variance along principal axes

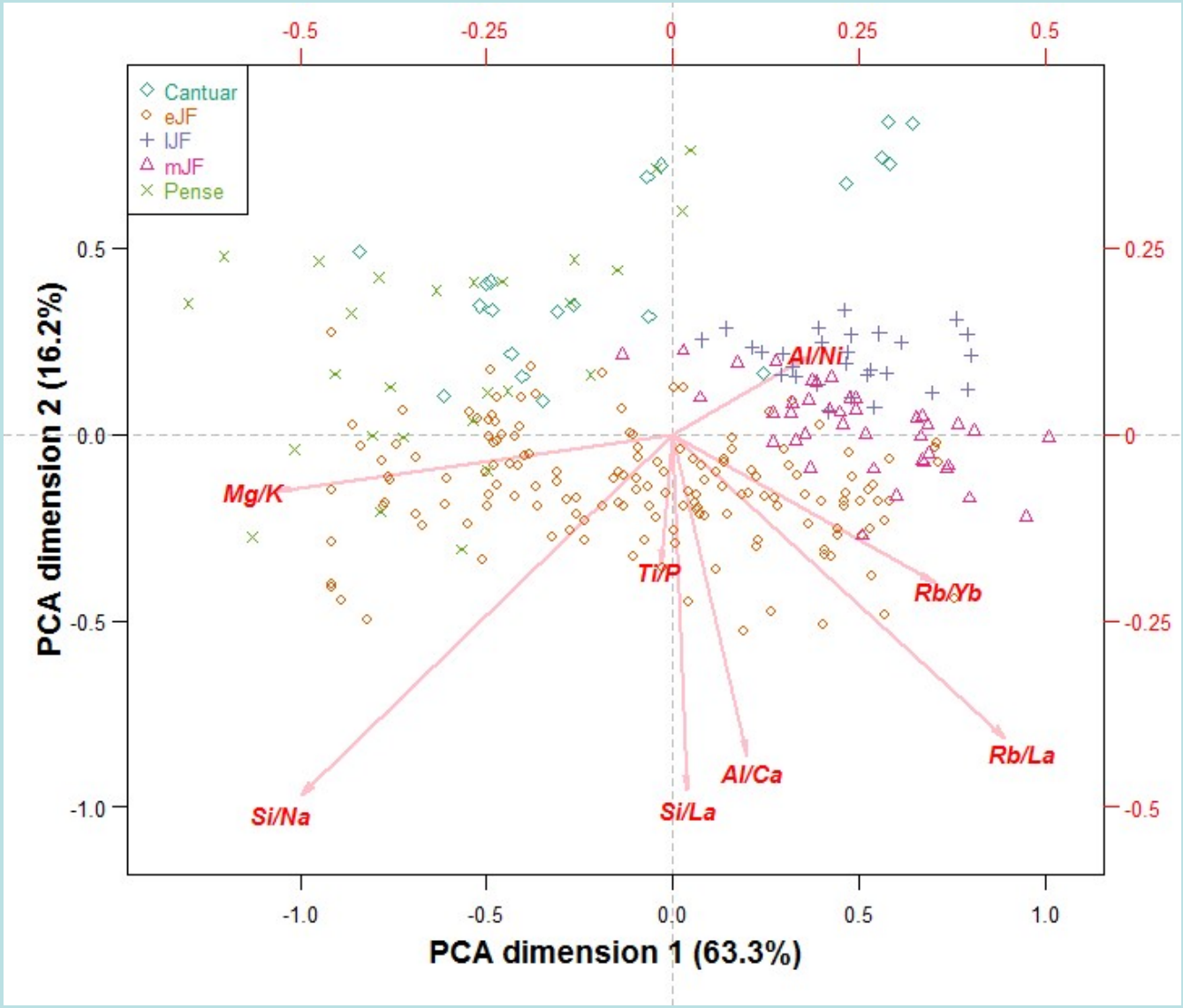
Stepwise logratio selection

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	1.000	0.895

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

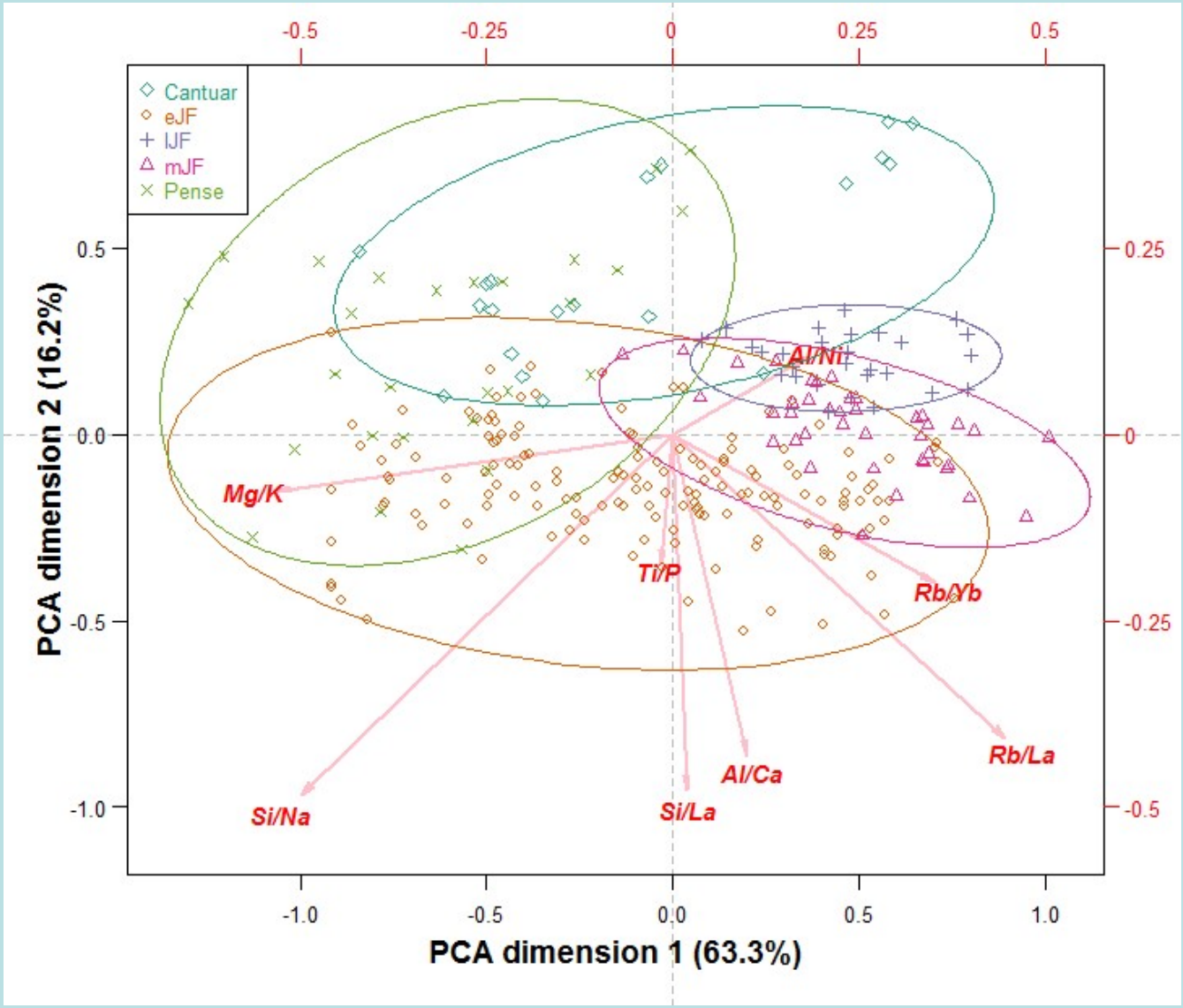
Procrustes correlation:
measures the similarity between the geometry of all 231 pairwise logratios and the geometry of the subset of pairwise logratios. A value of 1 means the geometries are isometric, i.e. the same.

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

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

Supervised selection step-by-step

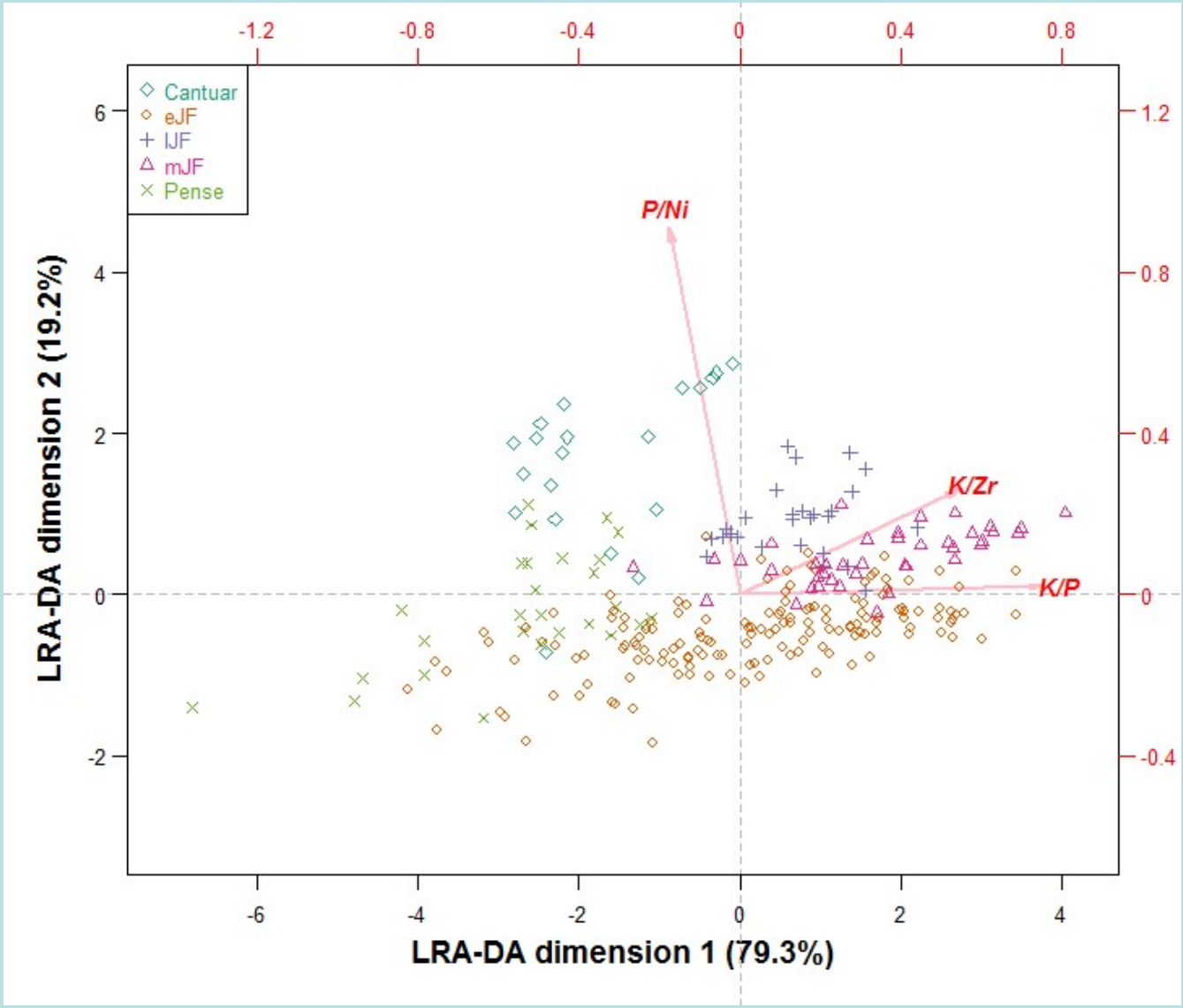
- Step 3

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%!

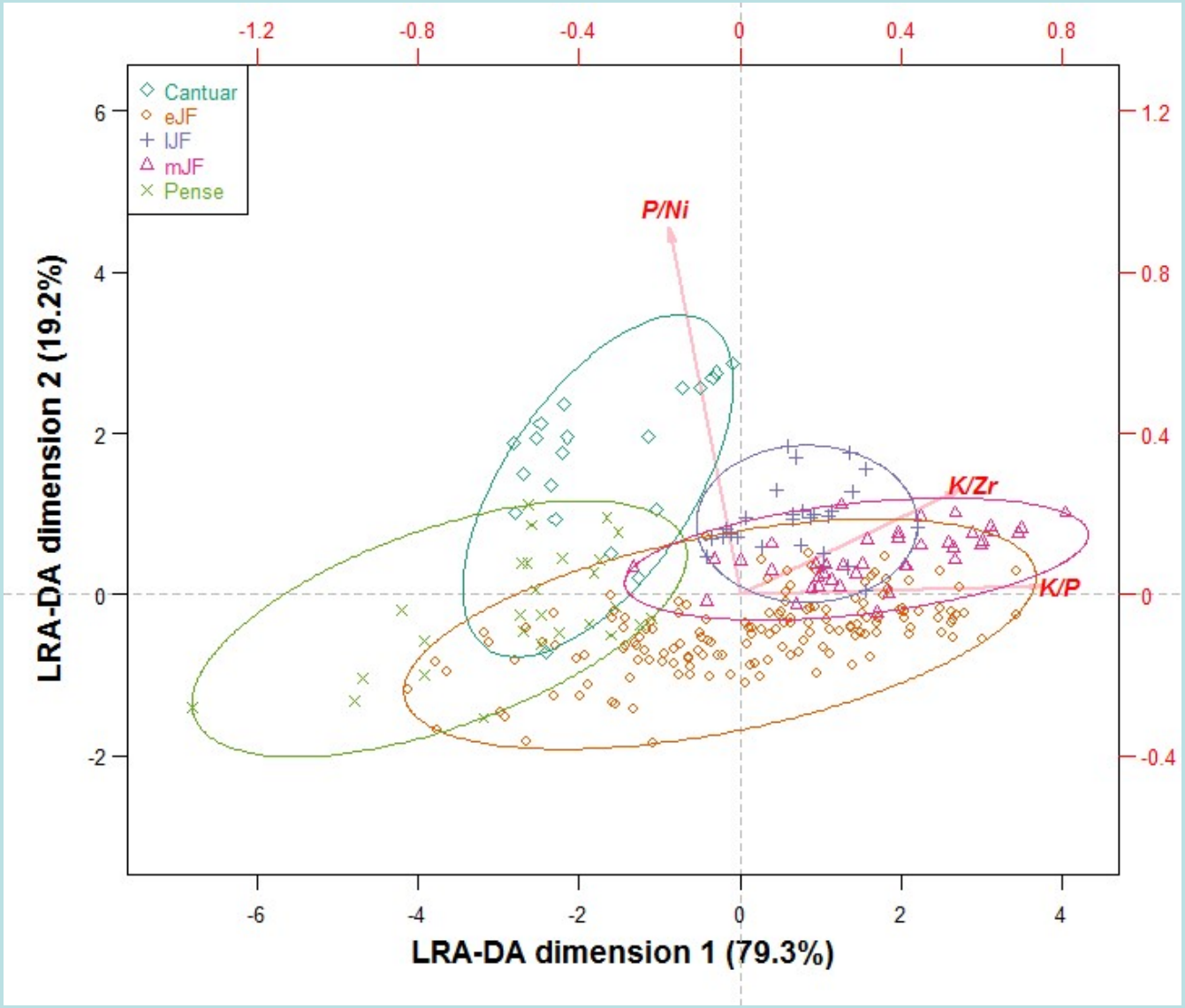
Best 3 logratios chosen by EG (99.2% of the between-group variance)



Between
54.5%
Within
45.5%

79.3% + 19.2% = 98.5% of that 99.2% is shown in this result

Best 3 logratios chosen by EG (90.5% of the variance)



Between
54.5%
Within
45.5%

Another approach: use additive logratios (ALRs)

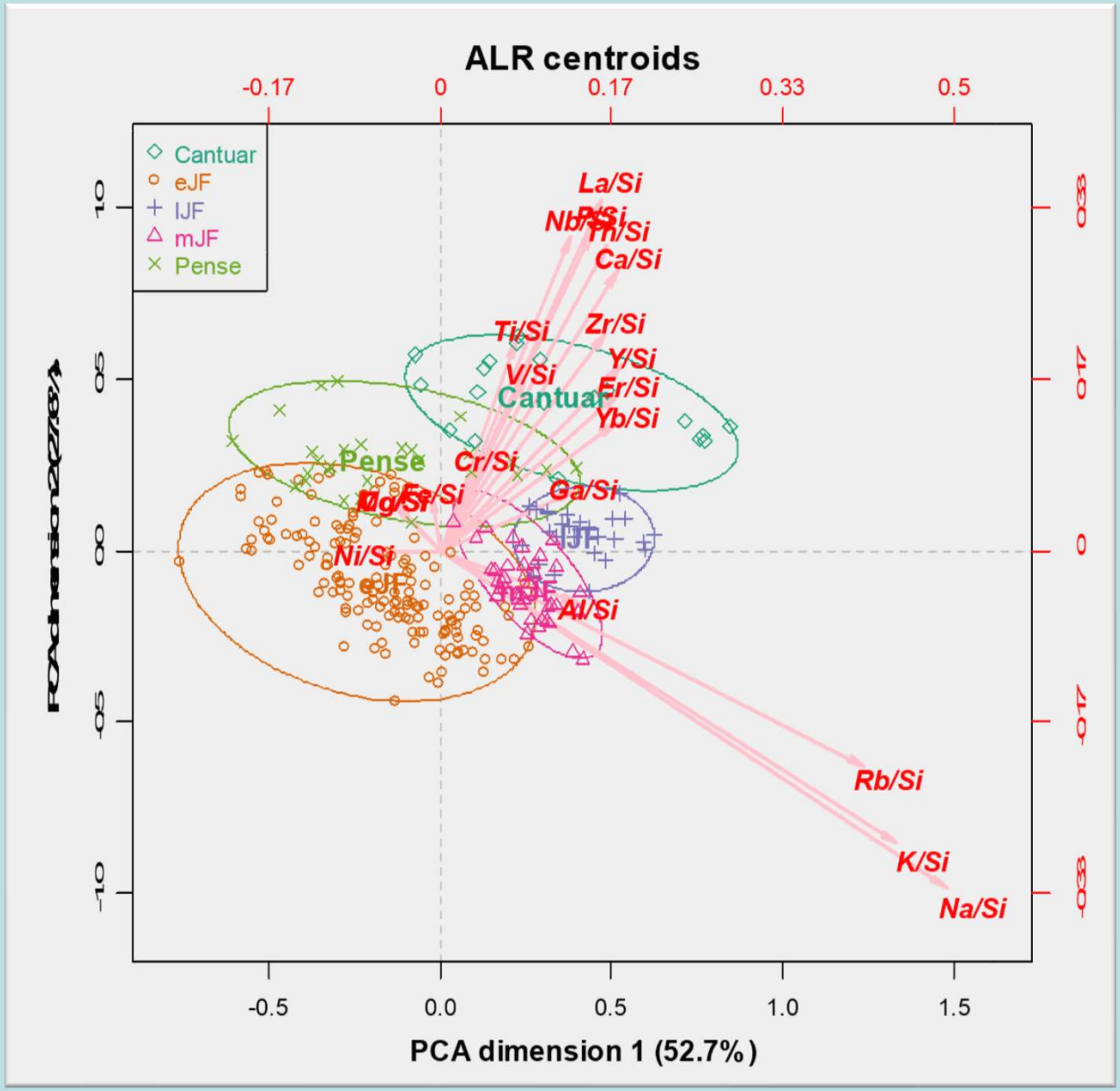
$$\mathrm{ALR}(\mathbf{x}, \mathrm{ref}) = \log(x_j / x_{\mathrm{ref}}) = \log(x_j) - \log(x_{\mathrm{ref}}) \quad j \neq \mathrm{ref}$$

ref	mean %	Procrustes	var of log
Zr	0.008	**0.977**	0.0956
Y	0.001	**0.977**	0.0967
Cr	0.131	0.973	0.0167
V	0.009	0.972	0.0426
Fe	10.045	0.968	**0.0049**
Si	42.334	0.966	**0.0030**
Nb	0.011	0.962	0.1194
Er	0.000	0.961	0.0952
Ti	1.566	0.959	0.0620
Mg	36.955	0.955	0.0068
Ga	0.001	0.955	0.0611
Co	0.011	0.945	0.0170
Th	0.002	0.944	0.1440
La	0.008	0.941	0.1602
Yb	0.000	0.938	0.1025
Al	2.676	0.937	0.0790
P	0.384	0.919	0.1491
Ni	0.166	0.918	0.0286
Rb	0.001	0.889	0.4940
K	0.225	0.877	0.6065
Na	0.286	0.835	0.8523
Ca	5.180	0.823	0.2279

highest Procrustes correlation

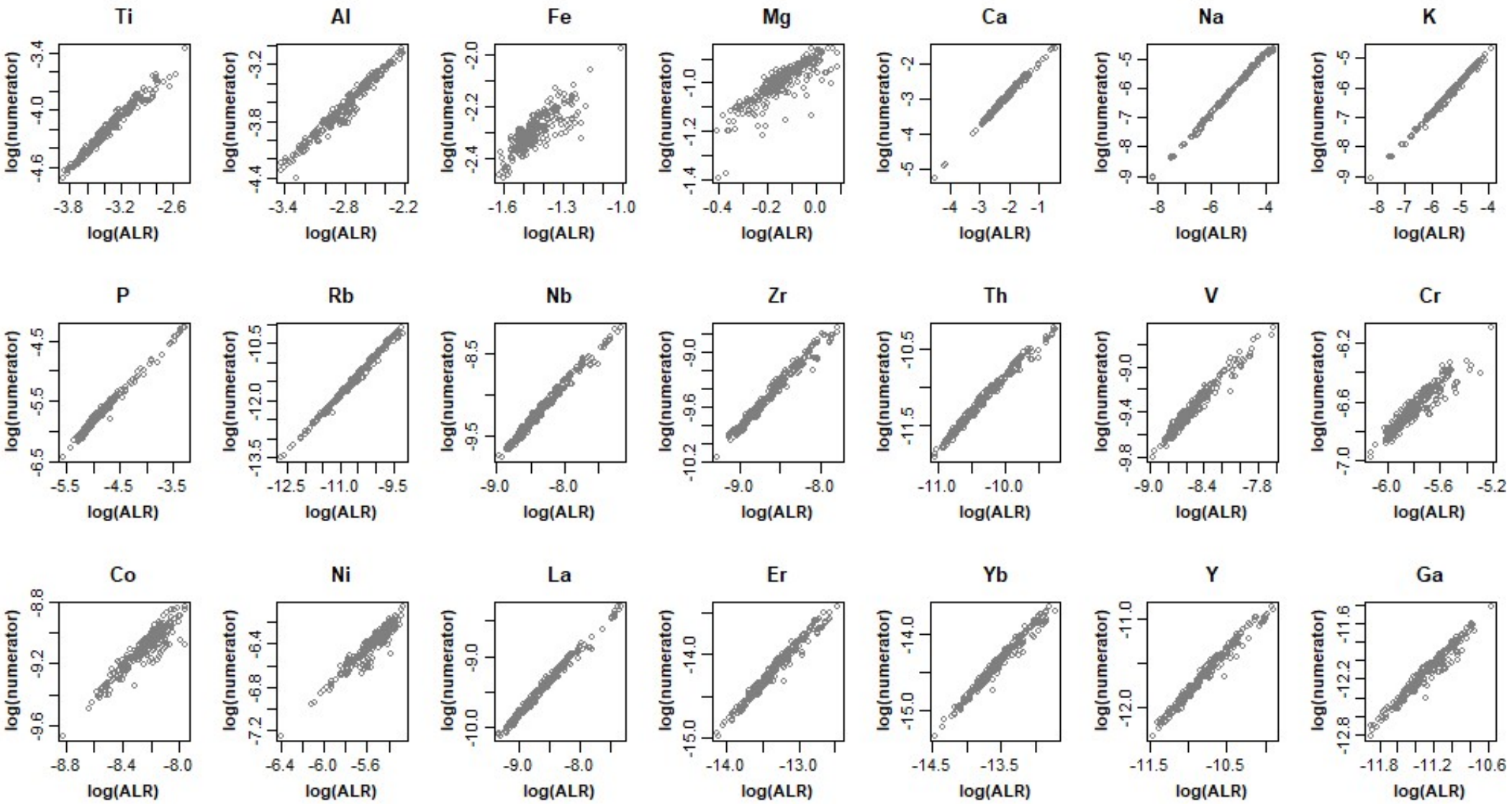
lowest $\log(x_{\mathrm{ref}})$

Using ALRs with Si as reference



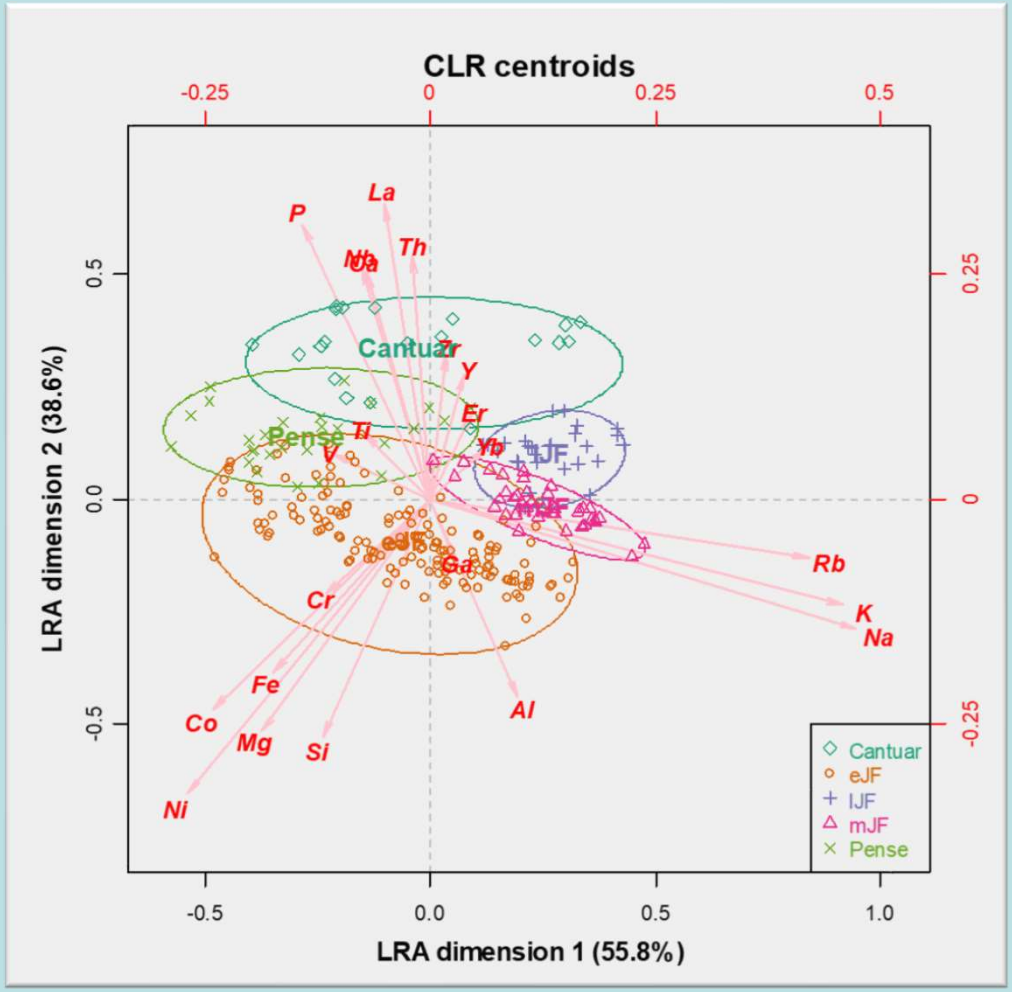
Between
49.3%
Within
50.7%

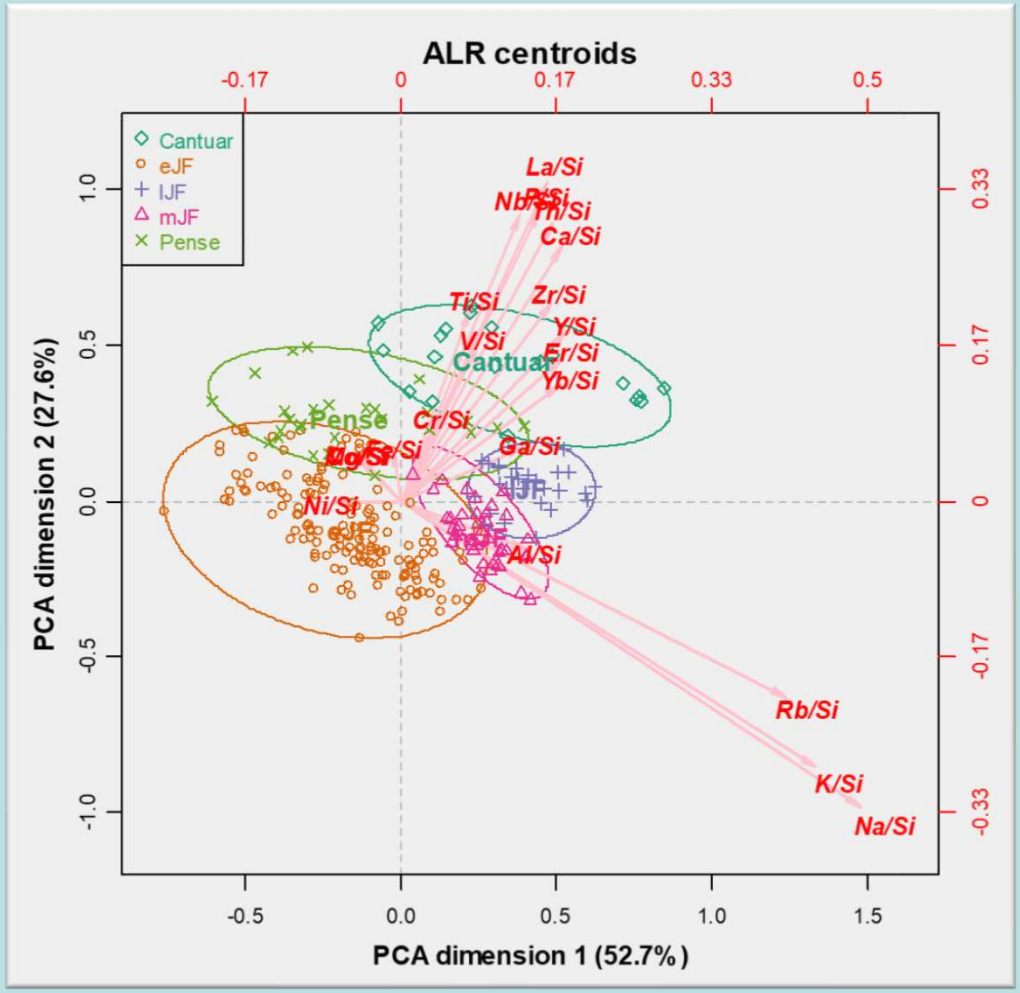
ALRs practically the same as numerator part



$$\mathrm{ALR}(\mathbf{x}, \mathrm{ref}) = \log(x_j / x_{\mathrm{ref}}) = \log(x_j) - \log(x_{\mathrm{ref}}) \quad j \neq \mathrm{ref}$$

CLRs (all logratios) compared with ALRs (reference Si)



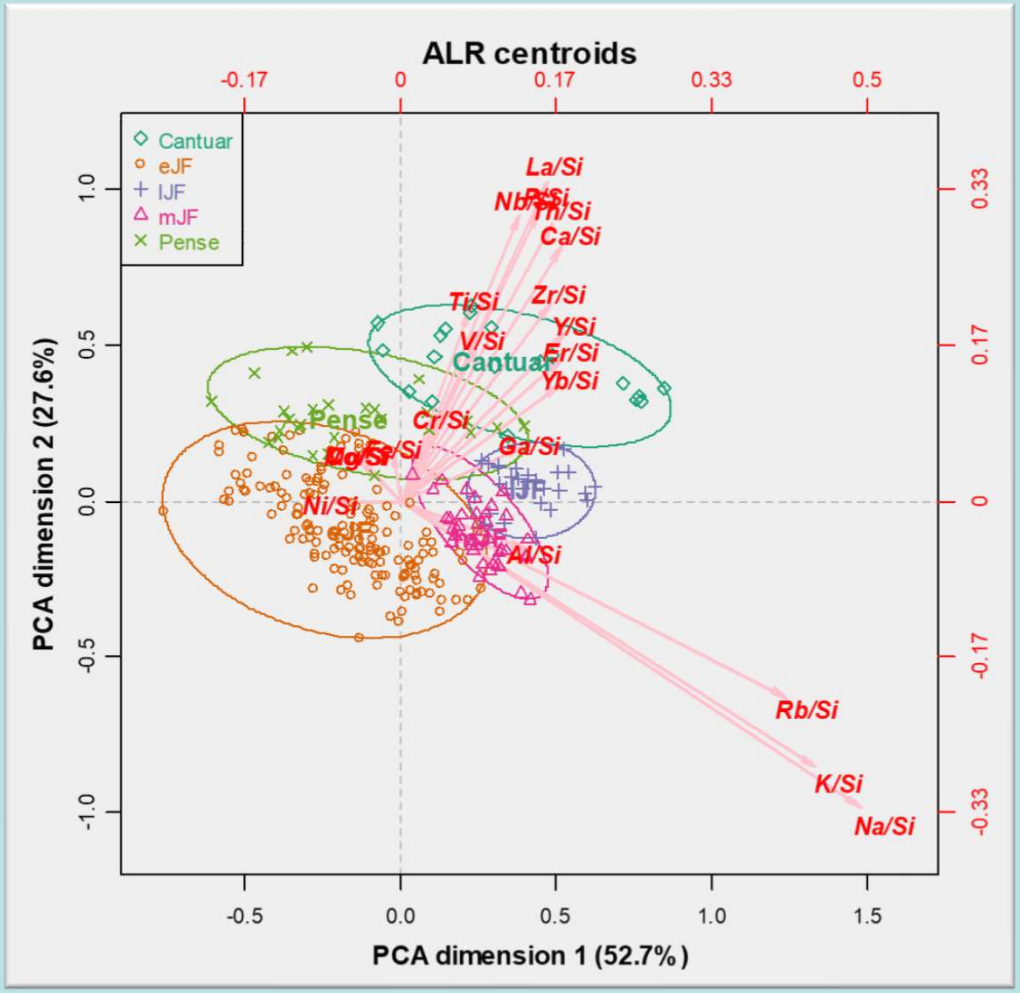


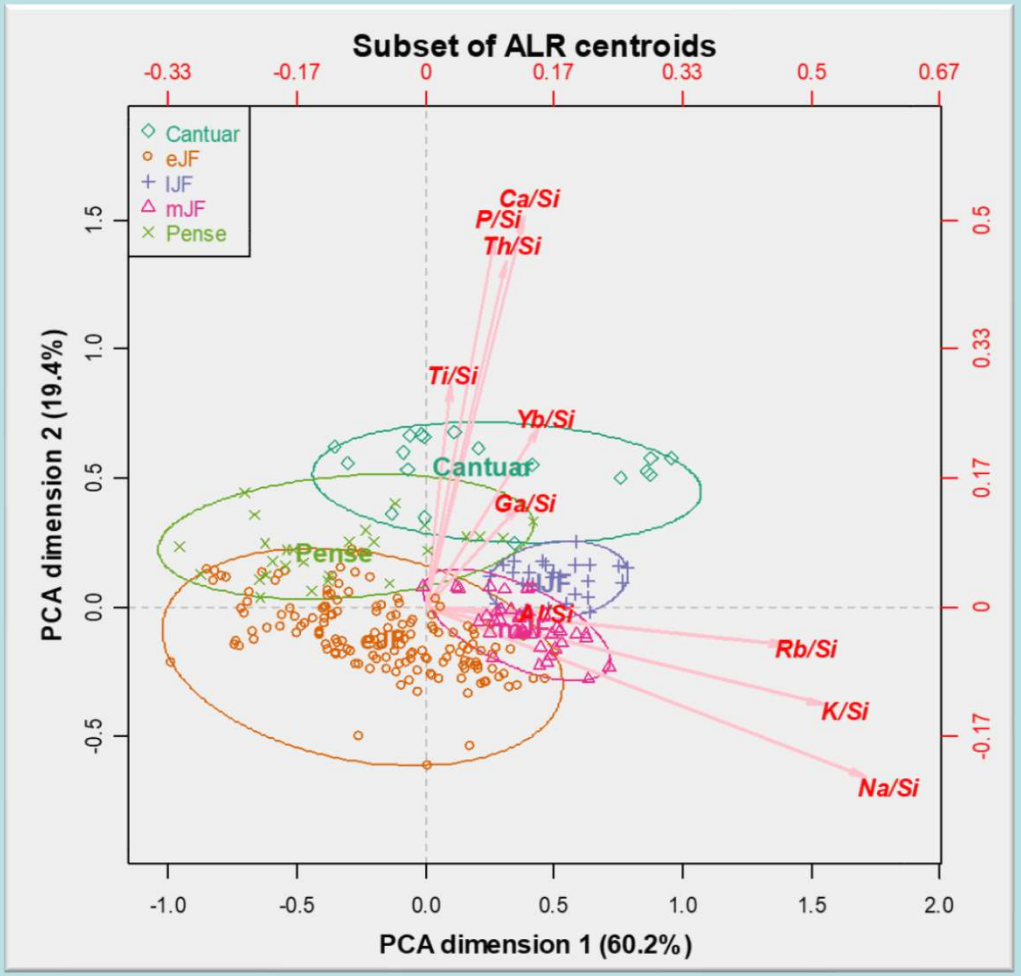
ratios = 22, but equivalent to all 231 pairwise logratios

ratios = 21, one less than the number of parts

It looks like there is an opportunity to reduce the number of ALRs, which I propose to do by backward elimination

ALRs (reference Si) compared to a subset of 10 of them



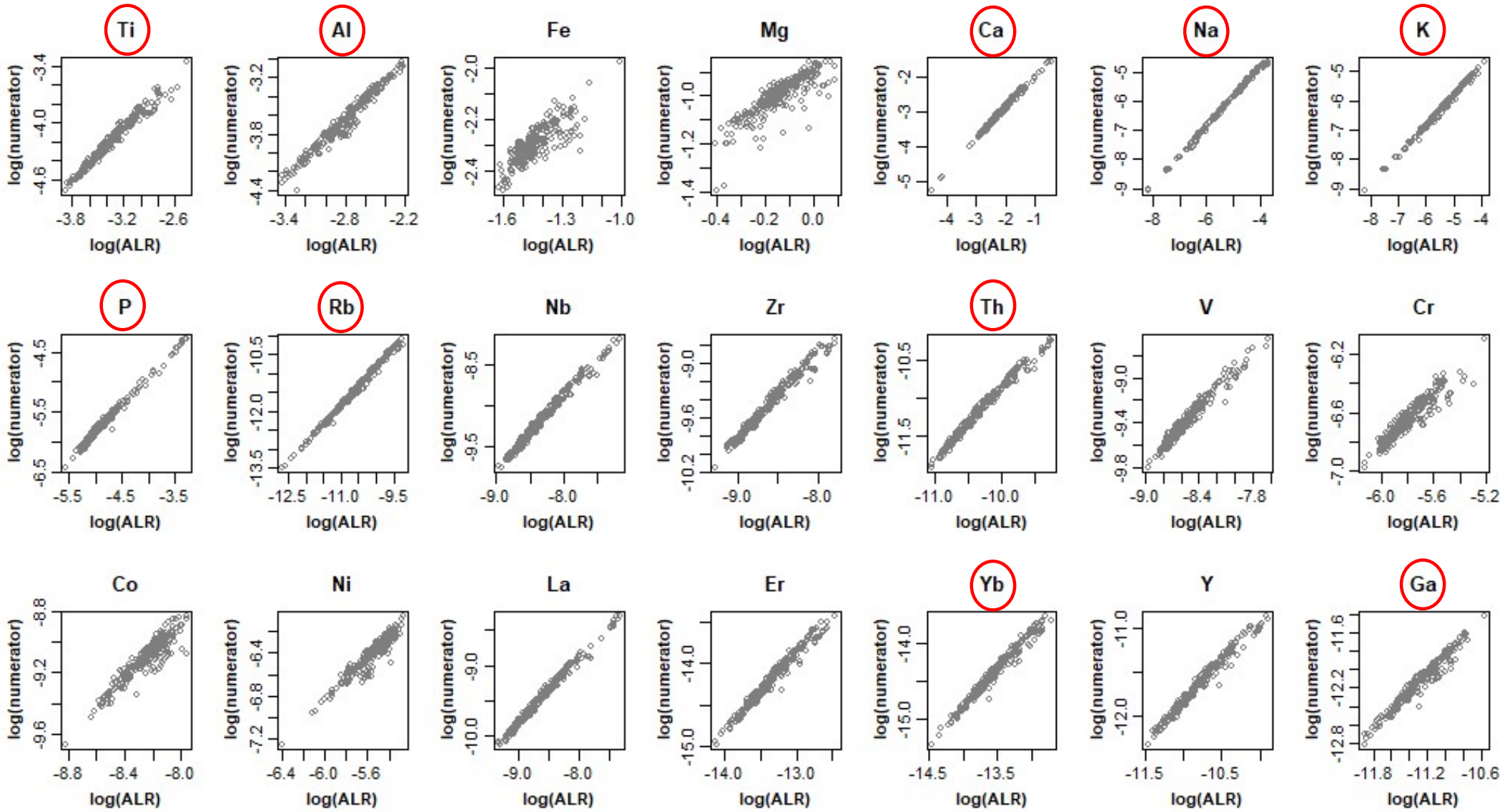


ratios = 21, one less than the number of parts

ratios = 10, i.e. a 11-part subcomposition (incl. Si)

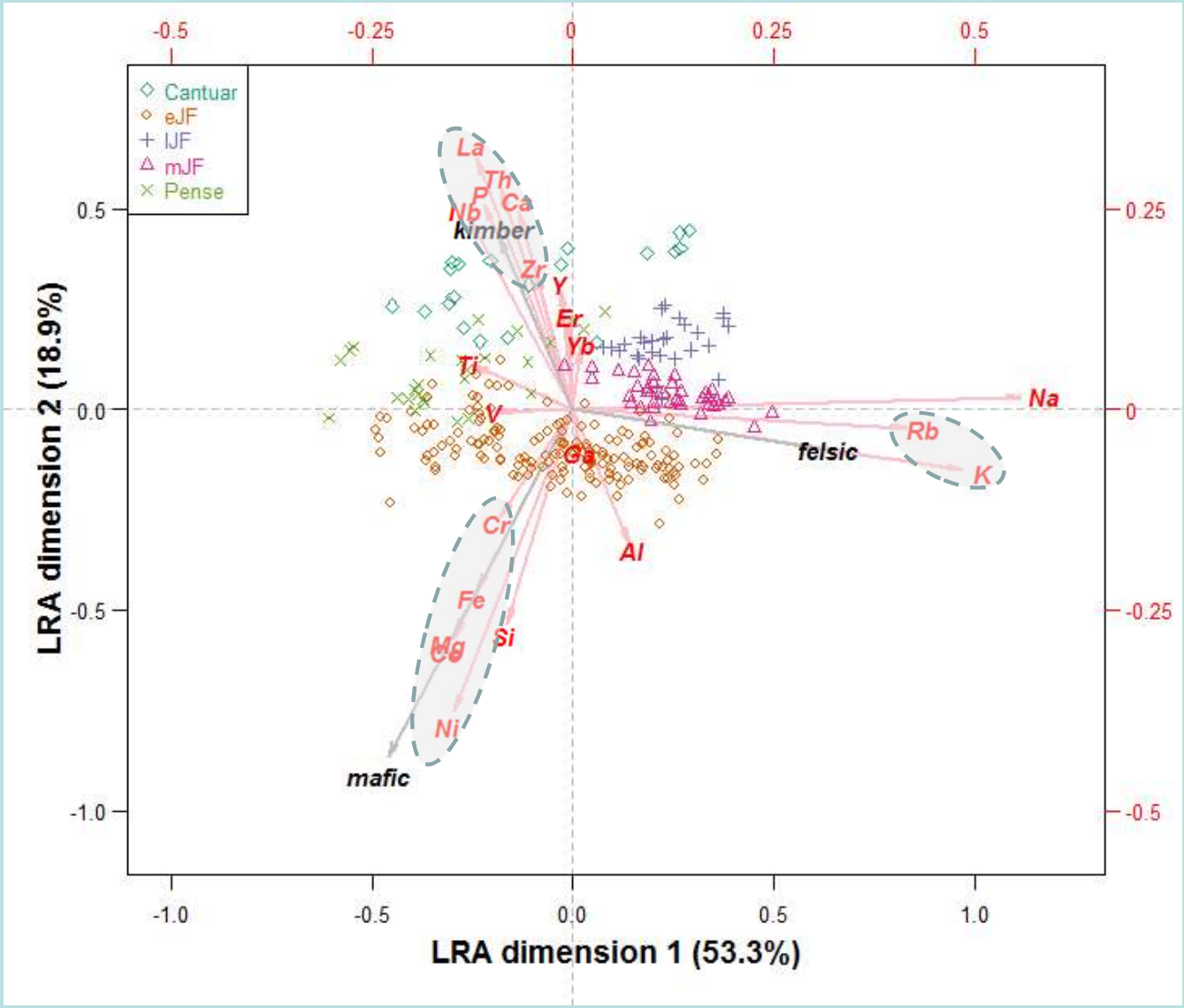
Ratios out	0	1	2	3	4	5	6	7	8	9	10	11
%expl:	1.000	0.998	0.996	0.994	0.989	0.986	0.984	0.983	0.981	0.978	0.975	0.968
Procr:	0.977	0.968	0.971	0.972	0.972	0.972	0.972	0.971	0.970	0.969	0.967	0.965

ALRs in subset practically the same as numerator part



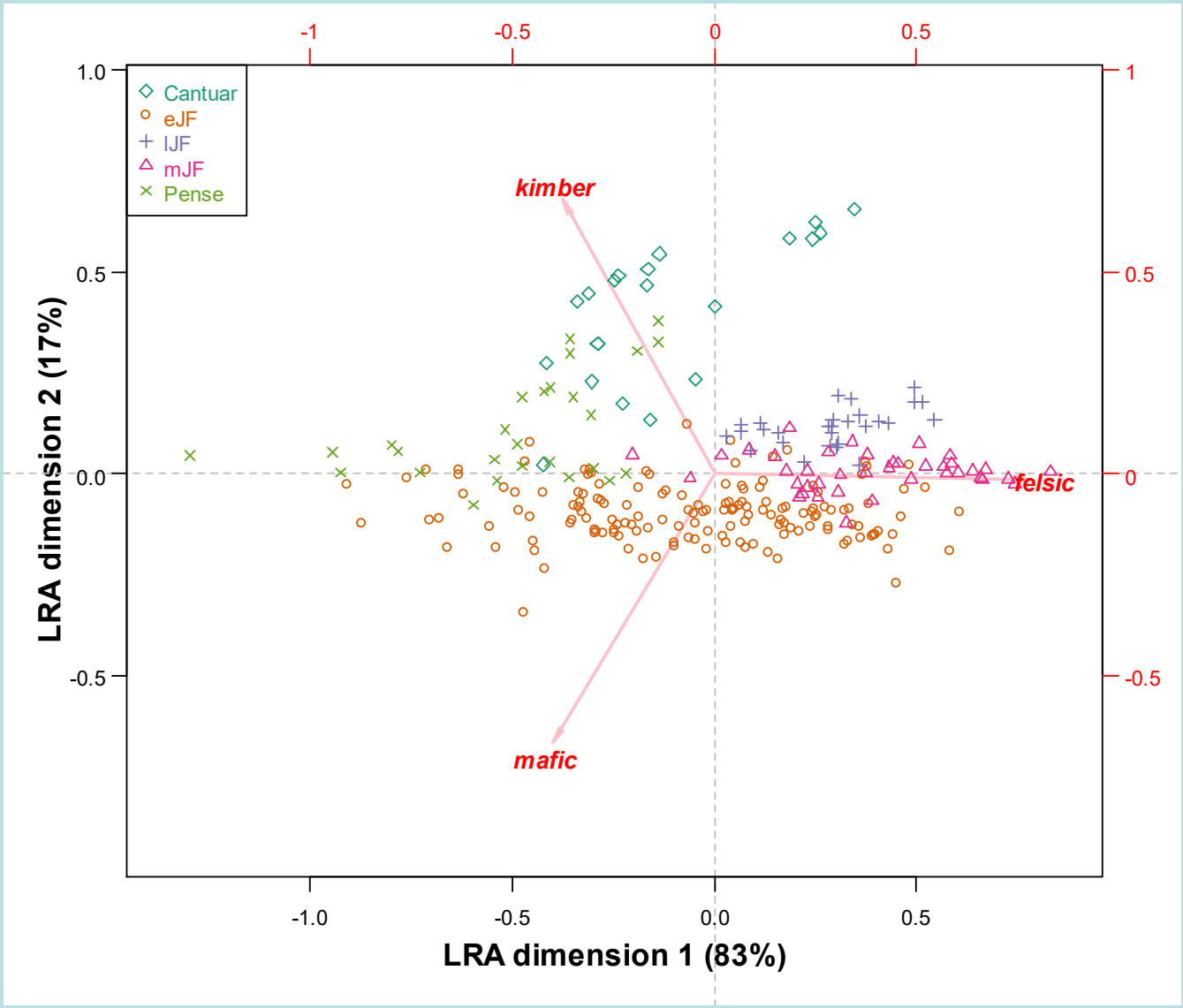
$$\mathrm{ALR}(\mathbf{x}, \mathrm{ref}) = \log(x_j / x_{\mathrm{ref}}) = \log(x_j) - \log(x_{\mathrm{ref}}) \quad j \neq \mathrm{ref}$$

Adding supplementary variables, e.g. amalgamations



mafic : Fe + Mg + Co + Cr + Ni
felsic : K + Rb
kimberlite Nb + La + Th + Zr + P

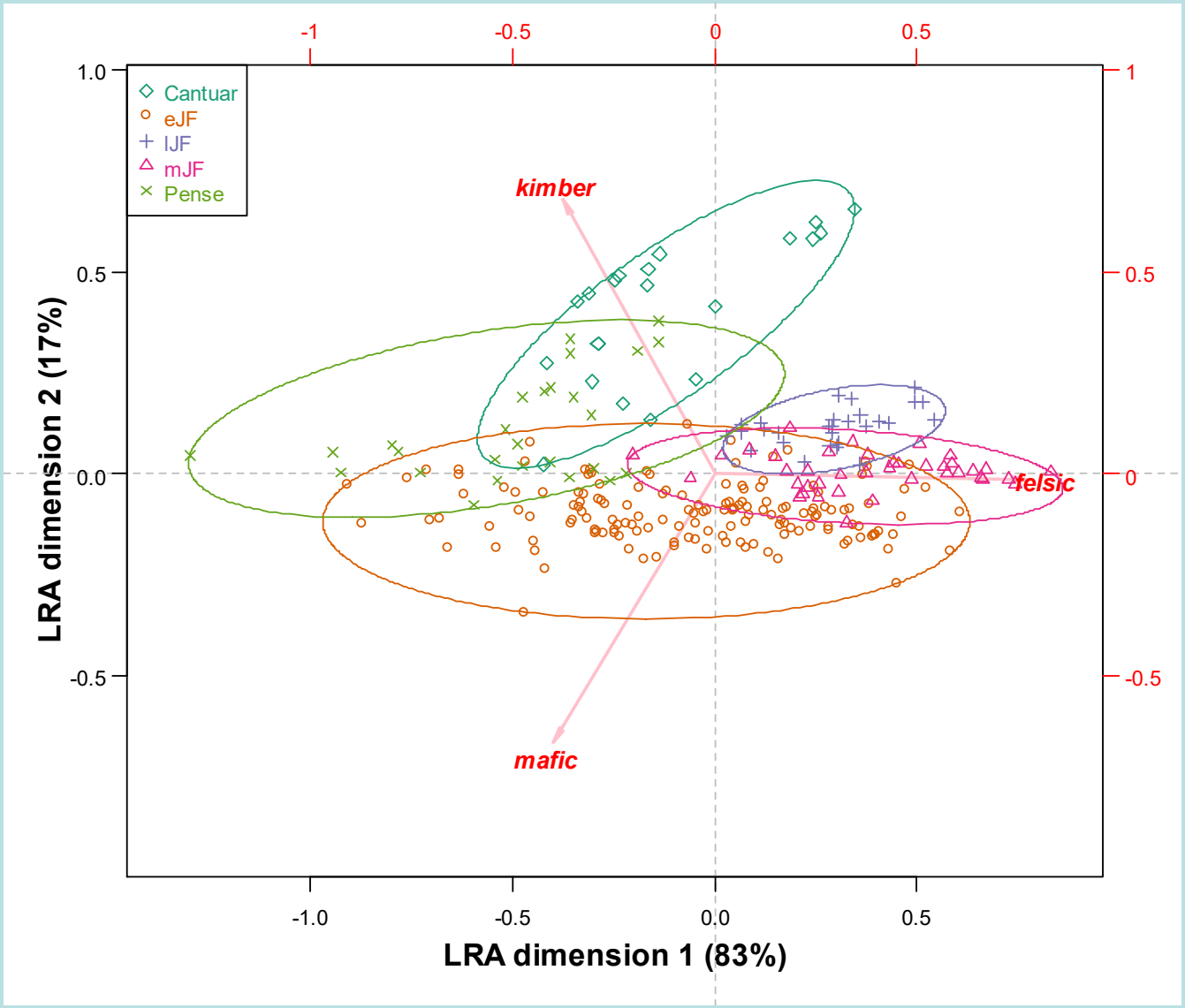
Amalgamations defining a subcomposition



Between
53.2%
Within
46.8%

mafic : Fe + Mg + Co + Cr + Ni
felsic : K + Rb
kimberlite Nb + La + Th + Zr + P

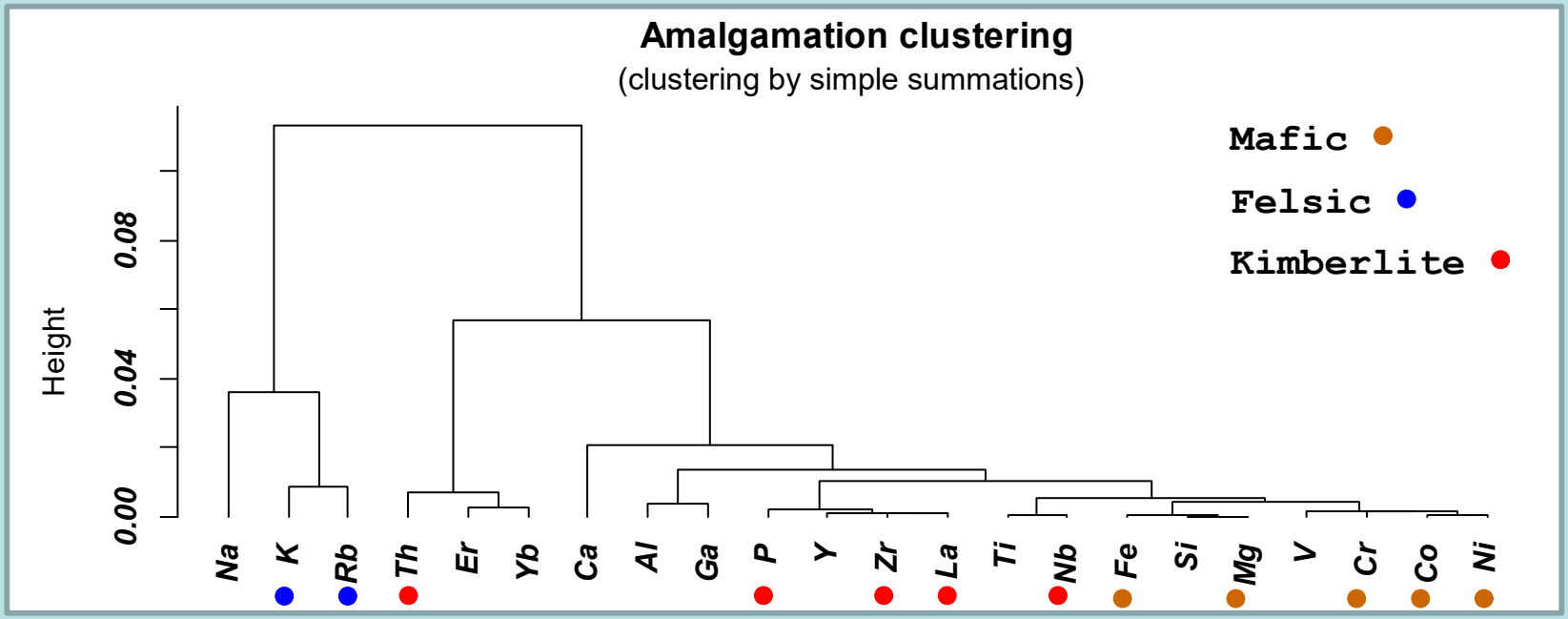
Amalgamations defining subcomposition

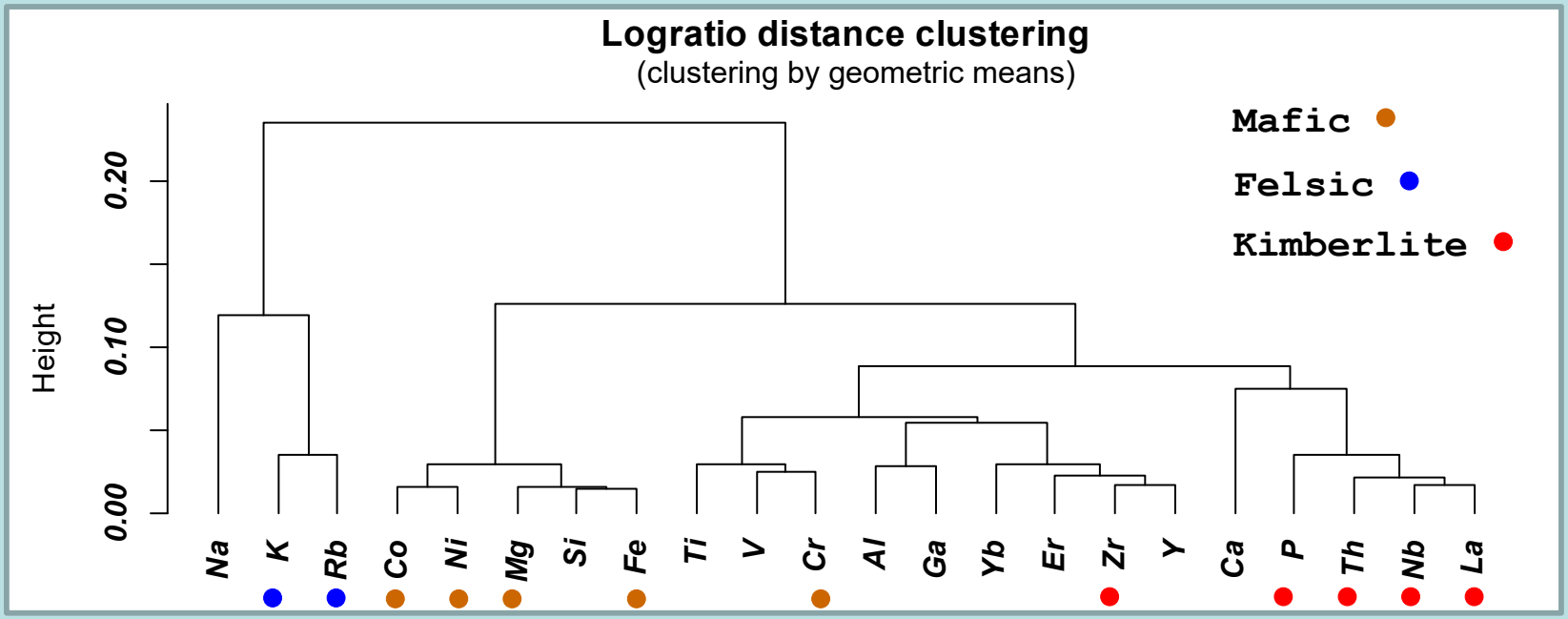


Between
53.2%
Within
46.8%

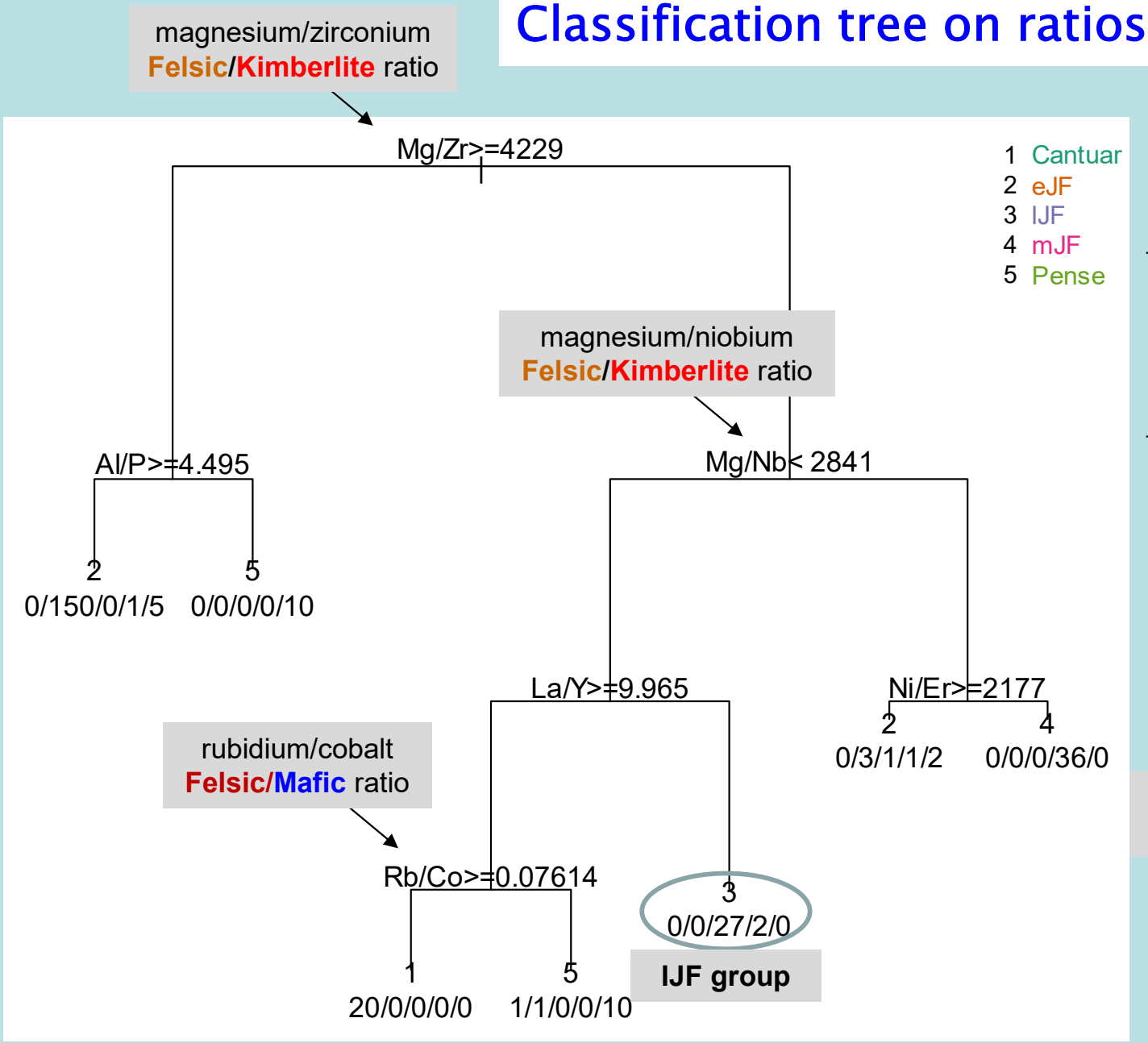
mafic : Fe + Mg + Co + Cr + Ni
felsic : K + Rb
kimberlite Nb + La + Th + Zr + P

Alternative cluster analyses





Classification tree on ratios



Classification matrix

Truth	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 (3rd group when

Mg/Zr > 4229 [39%]
& Mg/Nb < 2841 [24%]
& La/Y < 9.965 [68%]

lanthanum/yttrium
Kimberlite/...

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