IAMG Distinguished Lecturer – 2007

Prof. Dr. Vera Pawlowsky-Glahn

Department of Computer Science and Applied Mathematics
University of Girona, Spain

The Aitchison geometry of the simplex and the statistical analysis of compositional data

Prof. Dr. Vera Pawlowsky-Glahn

Department of Computer Science and Applied Mathematics
University of Girona, Spain

Facultad de Ciencias – Departamento de Matemáticas
Universidad de los Andes
Bogotá, Colombia, March 15, 2007

what are compositional data?

- **definition: parts of some whole which only carry relative information**
- **usual units of measurement:** parts per unit, percentages, ppm, ppb, concentrations, ... **(constant sum constraint)**
- **examples:** geochemical analysis; (sand, silt, clay) composition; proportions of minerals in a rock; ...

historical remarks: end of the XIX*th* century

Karl Pearson, 1897: "On a form of spurious correlation which may arise when indices are used in the measurement of organs"

- he was the first to point out dangers that may befall the analyst who attempts to interpret correlations between ratios whose numerators and denominators contain common parts
- the ***closure problem*** was stated within the **framework of classical statistics**, and thus within the **framework of Euclidean geometry in real space**

the problem: negative bias & spurious correlation

example: scientists A and B record the composition of aliquots of soil samples; A records (animal, vegetable, mineral, water) compositions, B records (animal, vegetable, mineral) after drying the sample; both are absolutely accurate (adapted from Aitchison, 2005)

sample A	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

sample B	x'_1	x'_2	x'_3
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

corr A	x_1	x_2	x_3	x_4
x_1	1.00	**0.50**	0.00	**-0.98**
x_2		1.00	**-0.87**	**-0.65**
x_3			1.00	0.19
x_4				1.00

corr B	x'_1	x'_2	x'_3
x'_1	1.00	**-0.57**	**-0.05**
x'_2		1.00	**-0.79**
x'_3			1.00

historical remarks: from 1897 to 1980 (and beyond)

- the fact that correlations between closed data are induced by numerical constraints caused **Felix Chayes** to attempt to separate the ***spurious*** part from the ***real*** **correlation**

  ("On correlation between variables of constant sum", 1960)

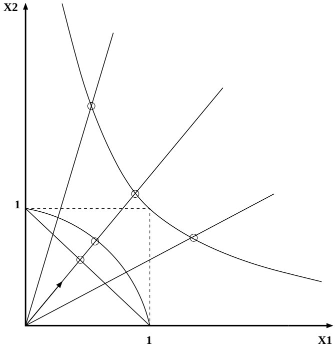
- many studied the **effects of closure** on methods related to correlation and covariance analysis (principal component analysis, partial and canonical correlation analysis) or distances (cluster analysis)
- an **exhaustive search** was initiated within the **framework of classical (applied) statistics**

historical remarks: end of the XX*th* century

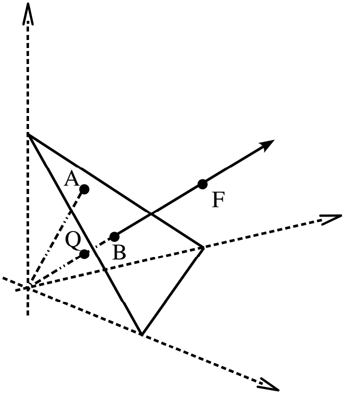
John Aitchison, 1982, 1986: "The statistical analysis of compositional data"

- **key idea:** compositional data represent parts of some whole; they only carry ***relative information***
- by analogy with the log-normal approach, Aitchison projected the sample space of compositional data, the D-part simplex $\mathcal{S}^D$, to real space $\mathbb{R}^{D-1}$ or $\mathbb{R}^D$, using log-ratio transformations
- **the log-ratio approach was born ...**

compositional data are equivalence classes



compositional data in $\mathbb{R}^2$



compositional data in $\mathbb{R}^3$

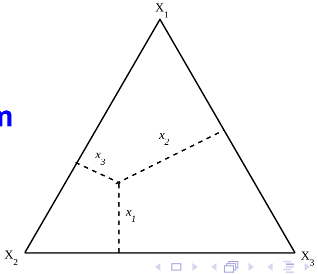
usual representation: the simplex

compositional data and their sample space

- **parts of some whole which only carry relative information**
- sample space: the **simplex** (for κ a constant)

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D \,\middle|\, x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- standard representation for $D = 3$: **ternary diagram**



Euclidean space structure of $\mathcal{S}^D$

for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$, and $\mathcal{C}$ the closure operation,

- **perturbation**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$
- **powering**: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$
- **Aitchison inner product, norm and distance**:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \quad \|\mathbf{x}\|_a^2 = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2$$

$$d_a^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2$$

- **dimension**: $(D-1)$

advantages of Euclidean spaces

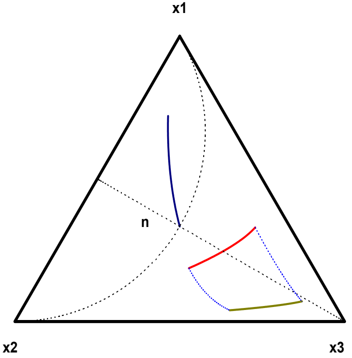
- **orthonormal basis** can be constructed: $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$
- **coordinates obey the rules** of real Euclidean space:

  $\mathbf{x} \in \mathcal{S}^D \Rightarrow \mathbf{y} = [y_1, \ldots, y_{D-1}] \in \mathbb{R}^{D-1}$, with $y_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$
- **standard methods** can be directly applied to coordinates
- **expressing results as compositions is easy**:

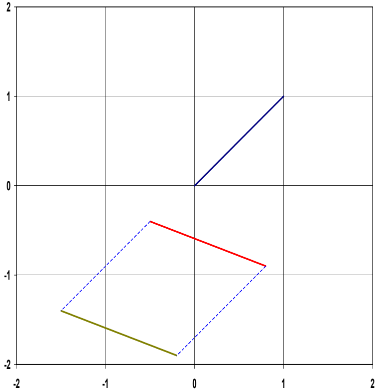
  if $h : \mathcal{S}^D \mapsto \mathbb{R}^{D-1}$ assigns to each $\mathbf{x} \in \mathcal{S}^D$ its coordinates, i.e. $h(\mathbf{x}) = \mathbf{y}$, then

$$h^{-1}(\mathbf{y}) = \mathbf{x} = \bigoplus_{i=1}^{D-1} y_i \odot \mathbf{e}_i$$

understanding the Aitchison geometry: perturbation (shifting)



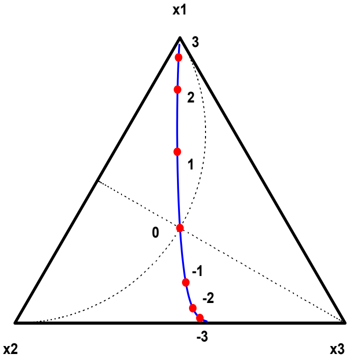
in $\mathcal{S}^3$



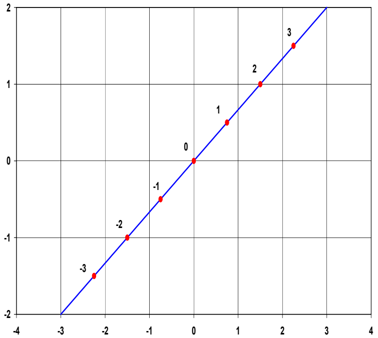
coordinate representation

$$\left(\frac{1}{\sqrt{6}}\ln\frac{x_1\cdot x_2}{x_3\cdot x_3},\ \frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}\right)$$

understanding the Aitchison geometry: powering (exponential decay or growth)



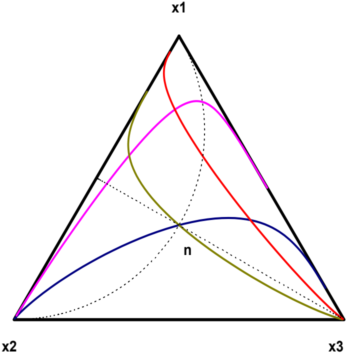
in $\mathcal{S}^3$



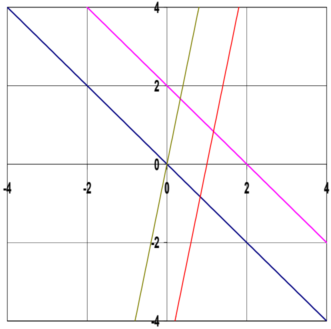
coordinate representation

$$\left(\frac{1}{\sqrt{6}}\ln\frac{x_1\cdot x_2}{x_3\cdot x_3}, \frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}\right)$$

understanding the Aitchison geometry: parallel lines (different origins)



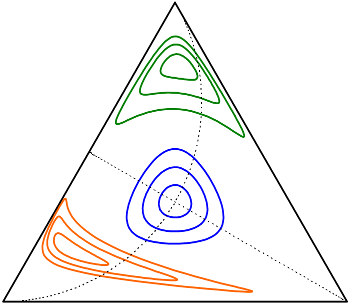
in $\mathcal{S}^3$



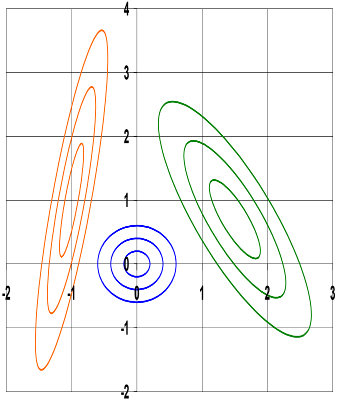
coordinate representation

$$\left(\frac{1}{\sqrt{6}}\ln\frac{x_1\cdot x_2}{x_3\cdot x_3},\ \frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}\right)$$

understanding the Aitchison geometry: circles and ellipses (normal densities)



in $\mathcal{S}^3$



coordinate representation

$$\left(\frac{1}{\sqrt{6}}\ln\frac{x_1\cdot x_2}{x_3\cdot x_3}, \frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}\right)$$

orthonormal basis: example of construction

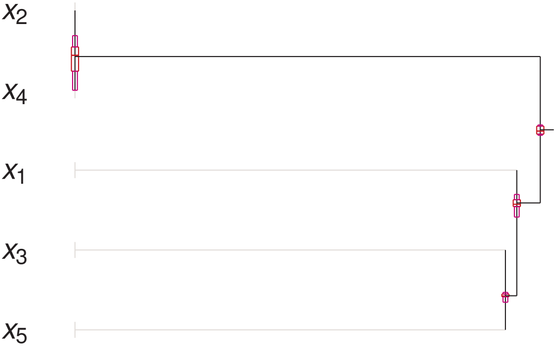
define a sequential binary partition and compute the coefficients for each sample; e.g. for $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5] \in \mathcal{S}^5$

order	x_1	x_2	x_3	x_4	x_5	coefficient
1	-1	$+1$	-1	$+1$	-1	$y_1 = \sqrt{\frac{2\cdot 3}{2+3}} \ln \frac{(x_2 \cdot x_4)^{1/2}}{(x_1 \cdot x_3 \cdot x_5)^{1/3}}$
2	0	$+1$	0	-1	0	$y_2 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_2}{x_4}$
3	$+1$	0	-1	0	-1	$y_3 = \sqrt{\frac{1\cdot 2}{1+2}} \ln \frac{x_1}{(x_3 \cdot x_5)^{1/2}}$
4	0	0	$+1$	0	-1	$y_4 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_3}{x_5}$

these type of coordinates are called **balances**

orthonormal basis: visualisation

- compute summary statistics of each balance
- plot the summary statistics in a dendrogram-type graph



boxplots: ($p_{0.05}$, Q_1, Q_2, Q_3, $p_{0.95}$) of empirical distributions;

boxplot scale: $(-2, 2)$; horizontal bars $\approx$ variance

the treatment of zeros

case 1: the part with zeros is not important for the study
$\Rightarrow$ the part should be omitted

case 2: the part is important, the zeros are essential
$\Rightarrow$ divide the sample into two or more populations, according to the presence/absence of zeros

case 3: the part is important, the zeros are rounded zeros
$\Rightarrow$ use imputation techniques

the principle of working on coordinates

- select a convenient **orthonormal basis**
- **perform any statistical analysis on the coordinates**
- **interpret test results** directly
- **interpret coordinates** if results are meaningful in coordinates, e.g. geochemical processes
- obtain results in $\mathcal{S}^D$ using the inverse if you prefer to **interpret compositions**

the principle of working on coordinates in $\mathcal{S}^D$ is equivalent to use the Aitchison geometry and the Aitchison measure

(Aitchison measure = Lebesgue measure on coordinates)

example

- granitoid rocks of a **progressive chemical weathering profile** developed on the Toorongo granodiorite in South Australia (Nesbitt and Markovics, 1977)
  - 15 samples and 12 major elements
  - sample space: $\mathcal{S}^{12}$
- **data used to model compositional change** by von Eynatten, Barceló-Vidal, and Pawlowsky-Glahn (2003, Math. Geol. 35(3))

sequential binary partition

order	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	H_2O
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	+1
2	−1	−1	−1	+1	+1	−1	−1	−1	−1	−1	−1	0
3	0	0	0	+1	−1	0	0	0	0	0	0	0
4	−1	−1	−1	0	0	−1	−1	+1	+1	−1	−1	0
5	0	0	0	0	0	0	0	+1	−1	0	0	0
6	−1	−1	+1	0	0	−1	−1	0	0	−1	−1	0
7	−1	−1	0	0	0	−1	−1	0	0	+1	−1	0
8	+1	−1	0	0	0	−1	−1	0	0	0	−1	0
9	0	+1	0	0	0	−1	−1	0	0	0	−1	0
10	0	0	0	0	0	−1	−1	0	0	0	+1	0
11	0	0	0	0	0	−1	+1	0	0	0	0	0

order 1: balance H_2O *vs.* others

order 2: balance $\{FeO, Fe_2O_3\}$ *vs.* others except H_2O

order 3: balance FeO *vs.* Fe_2O_3

order 4: balance $\{CaO, Na_2O\}$ *vs.* $\{SiO_2, TiO_2, Al_2O_3, MnO, MgO, K_2O, P_2O_5\}$

order 5: balance CaO *vs.* Na_2O

order 6: balance $\{Al_2O_3\}$ *vs.* $\{SiO_2, TiO_2, MnO, MgO, K_2O, P_2O_5\}$

summary statistics of coordinates (balances)

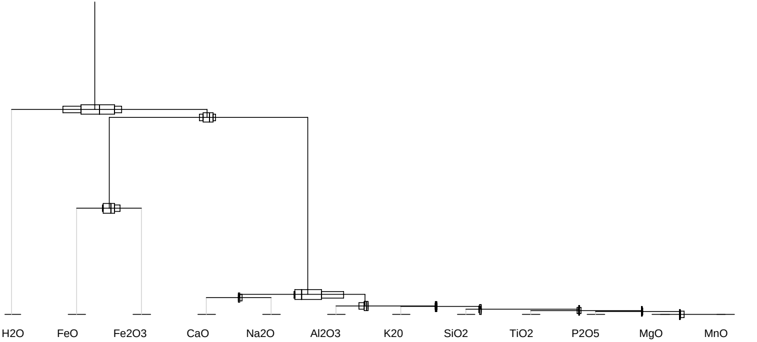
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
mean	0.92	0.08	-0.01	-0.55	0.02	2.52	0.70	4.51	0.78	-0.55	2.39
std	1.23	0.33	1.13	1.57	0.21	0.40	0.09	0.20	0.12	0.12	0.20
var	**1.52**	**0.11**	**1.27**	**2.47**	**0.04**	**0.16**	**0.01**	**0.04**	**0.01**	**0.02**	**0.04**
min	-0.71	-0.39	-1.97	-3.93	-0.60	2.19	0.57	4.36	0.57	-0.69	1.71
max	2.98	0.59	1.34	0.80	0.22	3.60	0.84	5.15	1.04	-0.14	2.66

total variance = 5.69

correlation matrix

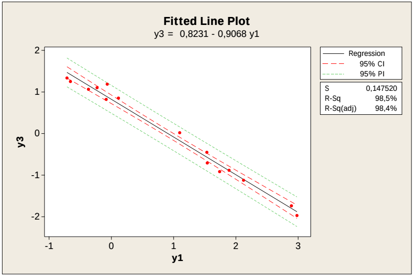
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
y_1	1.00	0.69	**-0.99**	**-0.93**	-0.59	**0.90**	0.62	0.81	0.55	0.08	-0.06
y_2		1,00	-0.69	-0.46	-0.02	0.32	0.65	0.30	0.62	-0.33	0.26
y_3			1.00	**0.94**	0.57	**-0.90**	-0.67	**-0.83**	-0.58	-0.11	0.03
y_4				1.00	0.57	**-0.95**	-0.63	**-0.90**	-0.49	-0.26	0.12
y_5					1.00	-0.77	-0.15	-0.73	-0.29	-0.53	0.02
y_6						1.00	0.49	**0.94**	0.47	0.39	-0.10
y_7							1.00	0.62	**0.84**	0.26	0.51
y_8								1.00	0.66	0.60	0.19
y_9									1.00	0.42	0.77
y_{10}										1.00	0.49
y_{11}											1.00

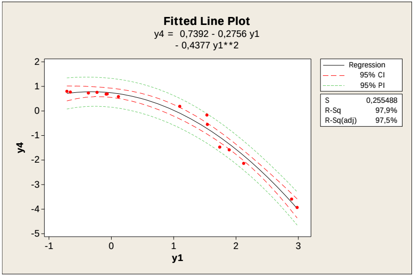
balances-dendrogram



boxplot scale: $(-6, 6)$

regression equations of balances



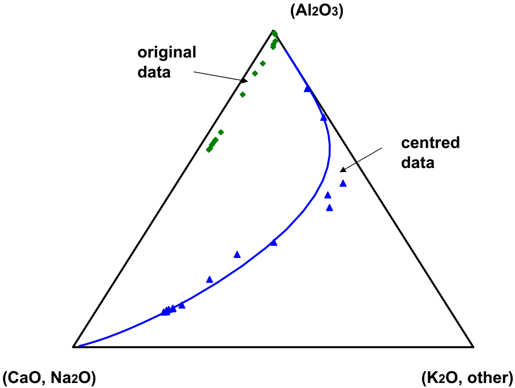


$$Y_1 = \sqrt{\frac{1\cdot 11}{1+11}} \ln \frac{H_2O}{(FeO\cdot Fe_2O_3\cdot CaO\cdot Na_2O\cdot Al_2O_3\cdot K_2O\cdot SiO_2\cdot TiO_2\cdot P_2O_5\cdot MgO\cdot MgO)^{1/11}}$$

$$Y_3 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{(FeO)}{(Fe_2O_3)}$$

$$Y_4 = \sqrt{\frac{2\cdot 7}{2+7}} \ln \frac{(CaO\cdot Na_2O)^{1/2}}{(Al_2O_3\cdot K_2O\cdot SiO_2\cdot TiO_2\cdot P_2O_5\cdot MgO\cdot MgO)^{1/7}}$$

principal component: chemical weathering profile



groups represented: (Al_2O_3), (CaO, Na_2O) and $(SiO_2, TiO_2, MnO, MgO, P_2O_5, K_2O)$ = inverse of balances b_4 and b_6

conclusions

- compositional data have a **constraint sample space**
- the natural measure of difference is a **relative measure**
- the **Aitchison geometry** offers the possibility of **working in coordinates**, which is a simple way to take these facts into account
- **main problems: appropriate representation and interpretation**
  - the **balance-dendrogram** facilitates finding an appropriate basis for interpretation
  - **classical statistical analysis** can be applied to coordinates