The statistical analysis of compositional data: The Aitchison geometry

Prof. Dr. Vera Pawlowsky-Glahn
Prof. Dr. Juan José Egozcue
Ass. Prof. Dr. René Meziat

Instituto Colombiano del Petróleo
Piedecuesta, Santander, Colombia
March 20–23, 2007

IAMG Distinguished Lecturer – 2007

Prof. Dr. Vera Pawlowsky-Glahn

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

recall

- **compositional data** are parts of some whole which only carry **relative information**
- **usual units of measurement:** parts per unit, percentages, ppm, ppb, concentrations, ...
- **historically:** data subject to a **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**

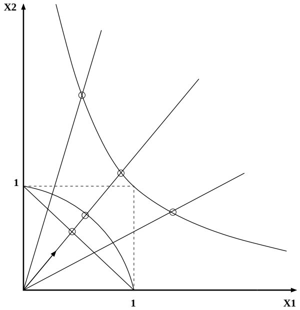
historical remarks: end of the XX*th* century

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

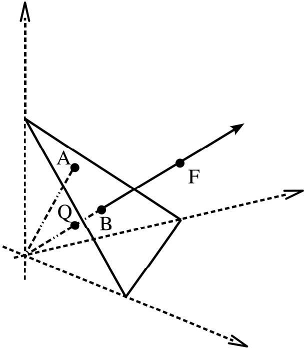
- **key idea:** compositional data represent parts of some whole; they only carry ***relative information***
- by analogy with the log-normal approach, Aitchison projected the sample space of compositional data, the D-part simplex $\mathcal{S}^D$, to real space $\mathbb{R}^{D-1}$ or $\mathbb{R}^D$, using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

compositional data: definition

definition: parts of some whole which carry only **relative information** $\Longleftrightarrow$ compositional data are **equivalence classes**



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



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

usual representation: subject to a **constant sum constraint**

compositional data: usual representation

definition: $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ is a D-part **composition**

$$\iff \begin{cases} x_i > 0, & \text{for all } i = 1, ..., D \\ \sum_{i=1}^{D} x_i = \kappa & \text{(constant)} \end{cases}$$

$\kappa = 1 \iff$ measurements in parts per unit

$\kappa = 100 \iff$ measurements in percent

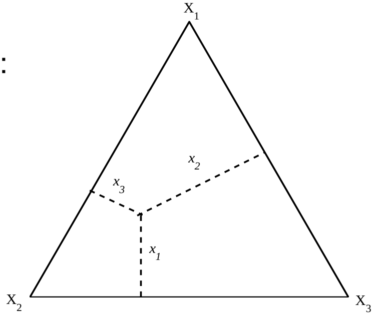
other frequent units: ppm, ppb, ...

a **subcomposition** $\mathbf{x}_s$ with s parts is obtained as the closure of a subvector $\left[x_{i_1}, x_{i_2}, \ldots, x_{i_s}\right]$ of $\mathbf{x}$

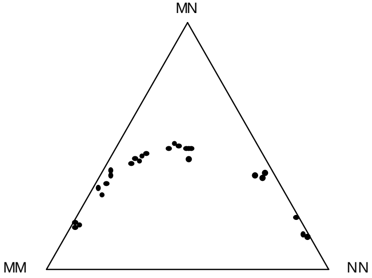
the simplex as sample space

$$\mathcal{S}^D = \{\mathbf{x} = [x_1, x_2, \ldots, x_D] | x_i > 0; \sum_{i=1}^{D} x_i = \kappa\}$$

standard representation for $D = 3$:
the ternary diagram



example 1: genetic hypothesis



data: genotyps in the MN system of blood groups; **code:** Ab = Aborigines; Ch = Chinese; In= Indian; AmIn = American Indian; Es = Eskimo;

question: despite the high variability which can be observed, is there any inherent stability in the data? do they follow any genetic law?

requirements for a proper analysis

- **scale invariance:** the analysis should not depend on the closure constant κ
- **permutation invariance:** the order of the parts should be irrelevant
- **subcompositional coherence:** studies performed on subcompositions should not stand in contradiction with those performed on the full composition

why a new geometry on the simplex?

in real space we **add** vectors, we **multiply** them by a constant, we look for **orthogonality** between vectors, we look for **distances** between points, ...

possible because $\Re^D$ is a linear vector space

BUT Euclidean geometry is not a proper geometry for compositional data because

- **results might not be in the simplex** when we **add** compositional vectors, **multiply** them by a constant, or compute **confidence regions**
- **Euclidean differences are not always reasonable:** from 0.05% to 0.10% the amount is doubled; from 50.05% to 50.10% the increase is negligible

basic operations

closure of $\mathbf{z} = [z_1, z_2, \ldots, z_D] \in \Re_+^D$

$$\mathcal{C}[\mathbf{z}] = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^D z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^D z_i}, \cdots, \frac{\kappa \cdot z_D}{\sum_{i=1}^D z_i} \right]$$

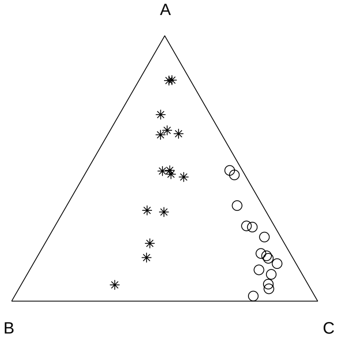
perturbation of $\mathbf{x} \in \mathcal{S}^D$ by $\mathbf{y} \in \mathcal{S}^D$

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D]$$

powering of $\mathbf{x} \in \mathcal{S}^D$ by $\alpha \in \Re$

$$\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]$$

interpretation of perturbation and powering



left: perturbation of initial compositions ($\circ$) by $\mathbf{p} = [0.1, 0.1, 0.8]$ resulting in compositions ($\star$)

right: powering of compositions ($\star$) by $\alpha = 0.2$ resulting in compositions ($\circ$)

comments

- **closure = projection** of a point in $\Re_+^D$ on $\mathcal{S}^D$
- points on a ray are projected onto the same point
  - a ray in $\Re_+^D$ is an equivalence class
  - the point on $\mathcal{S}^D$ is a representant of the class
  - a generalization to other representants is possible
- for $\mathbf{z} \in \Re_+^D$ and $\mathbf{x} \in \mathcal{S}^D$, $\quad \mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot \mathcal{C}[\mathbf{z}])$

vector space structure of $(\mathcal{S}^D, \oplus, \odot)$

- **commutative group structure** of $(\mathcal{S}^D, \oplus)$
  1. commutativity: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$
  2. associativity: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$
  3. neutral element: $\mathbf{e} = \mathcal{C}\,[1, 1, \ldots, 1]$ = barycentre of $\mathcal{S}^D$
  4. inverse of $\mathbf{x}$: $\mathbf{x}^{-1} = \mathcal{C}\left[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}\right]$

     $\Rightarrow \quad \mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{e}$ and $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$

- **properties of powering**
  1. associativity: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$;
  2. distributivity 1: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$
  3. distributivity 2: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$
  4. neutral element: $1 \odot \mathbf{x} = \mathbf{x}$

inner product space structure of $(\mathcal{S}^D, \oplus, \odot)$

inner product : $$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$$

norm : $$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}, \quad \mathbf{x} \in \mathcal{S}^D$$

distance : $$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}, \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$$

Aitchison geometry on the simplex

properties of the Aitchison geometry

distance and perturbation: $d_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}) = d_a(\mathbf{x}, \mathbf{y})$

distance and powering: $d_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_a(\mathbf{x}, \mathbf{y})$

compositional lines: $\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$
($\mathbf{x}_0$ = starting point, $\mathbf{x}$ = leading vector)

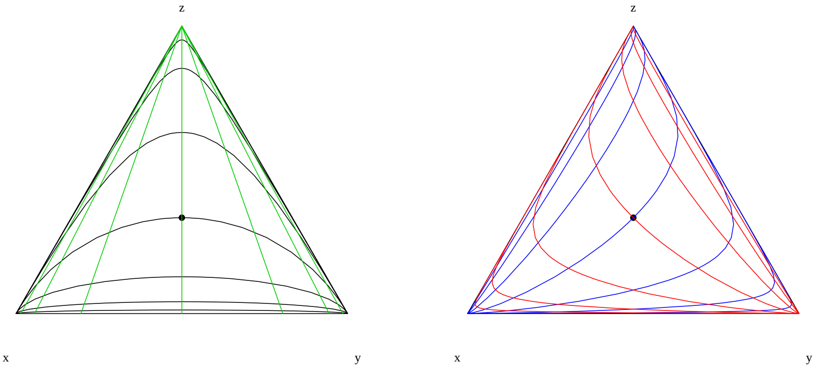
orthogonal lines: $\mathbf{y_1} = \mathbf{x_0} \oplus (\alpha_1 \odot \mathbf{x_1})$, $\mathbf{y_2} = \mathbf{x_0} \oplus (\alpha_2 \odot \mathbf{x_2})$,

$$\mathbf{y_1} \perp \mathbf{y_2} \iff \langle \mathbf{x_1}, \mathbf{x_2} \rangle_a = 0$$

(the inner product of the leading vectors is zero)

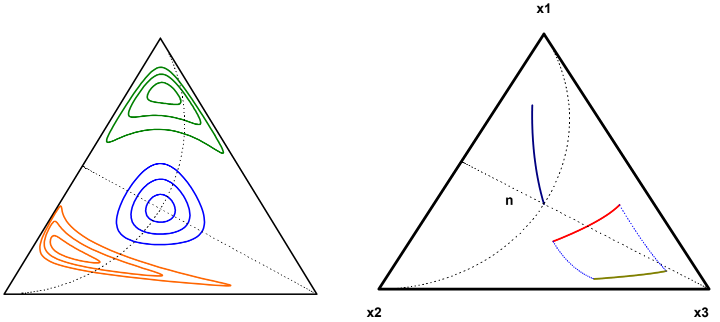
parallel lines: $\mathbf{y_1} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x}) \quad \| \quad \mathbf{y_2} = \mathbf{p} \oplus \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$

orthogonal compositional lines



orthogonal grids in $\mathcal{S}^3$, equally spaced, 1 unit in Aitchison distance; the right grid is rotated 45^o with respect to the left grid

circles and other geometric figures



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

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

conclusions

- the Aitchison geometry of the simplex offers a new tool to analyse CoDa
- the geometry is apparently complex, but it is completely equivalent to standard Euclidean geometry in real space
- the **key** is to use a **proper representation in coordinates**