The statistical analysis of compositional data: Coordinate representation

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

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

Summary

1. clr representation of compositions
2. Orthonormal basis. Balances
3. Enhancing interpretation using balance-coordinates

Definition of clr coefficients

Composition $\mathbf{x} \in \mathcal{S}^D$

Centered log-ratio of $\mathbf{x}$, $\mathrm{clr}(\mathbf{x})$, is the unique $\mathbb{R}^D$-vector $\boldsymbol{\xi} = [\xi_1, \xi_2, \dots, \xi_D]$, satisfying

$$\mathbf{x} = \mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}\left(\exp(\boldsymbol{\xi})\right) \;,\; \sum_{i=1}^{D} \xi_i = 0 \,.$$

The i-th clr coefficient is

$$\xi_i = \frac{\ln x_i}{g(\mathbf{x})} \quad , \quad g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$$

Properties of clr coefficients

If $\sum_1^D \xi_i = 0$, $\boldsymbol{\xi} \in \mathbb{R}_0$

- clr inverse
  $\mathrm{clr} : \mathcal{S}^D \to \mathbb{R}_0^D \subset \mathbb{R}^D$ is one-to-one and
  $$\mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}[\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_D)] = \mathbf{x}.$$

- clr transforms $\oplus$, $\odot$ into $+$, $\cdot$:
  $\mathrm{clr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathrm{clr}(\mathbf{x}_1) + \beta \cdot \mathrm{clr}(\mathbf{x}_2)$
- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) \rangle$
- $\|\mathbf{x}_1\|_a = \|\mathrm{clr}(\mathbf{x}_1)\|$, $d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2))$

Orthonormal basis

Definition

Compositions $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$ are an orthonormal basis if

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \langle \mathrm{clr}(\mathbf{e}_i), \mathrm{clr}(\mathbf{e}_j) \rangle = \delta_{ij}$$

clr matrix of the basis $(D-1, D)$

$$\mathbf{\Psi} = \begin{pmatrix} \mathrm{clr}(\mathbf{e}_1) \\ \mathrm{clr}(\mathbf{e}_2) \\ \dots \\ \mathrm{clr}(\mathbf{e}_{D-1}) \end{pmatrix} \quad , \quad \mathbf{\Psi}\mathbf{\Psi}' = I_{D-1} \quad , \quad \mathbf{\Psi}'\mathbf{\Psi} = I_D - (1/D)\mathbf{1}'_D \mathbf{1}_D$$

Coordinates

Given an orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$,
Expression in coordinates

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i \ , \ x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$$

Isometric log-ratio: assigns coordinates to a composition
$\mathrm{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is one-to-one.

$$\mathbf{x} \overset{\mathrm{ilr}}{\longrightarrow} \mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_{D-1}^*]$$

Properties of ilr-coordinates

Given an orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_{D-1}$ in $\mathcal{S}^D$

ilr and ilr^{-1}

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x}) \cdot \boldsymbol{\Psi}' \quad , \quad \mathbf{x} = \mathcal{C}\left(\exp(\mathbf{x}^* \boldsymbol{\Psi})\right)$$

Isometry: $\mathrm{ilr} : \mathcal{S}^D \longrightarrow \mathbb{R}^{D-1}$

$$\mathrm{ilr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathrm{ilr}(\mathbf{x}_1) + \beta \cdot \mathrm{ilr}(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^*$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2) \rangle = \langle \mathbf{x}_1^*, \mathbf{x}_2^* \rangle$$

$$\|\mathbf{x}\|_a = \|\mathrm{ilr}(\mathbf{x})\| \quad , \quad d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2))$$

Building an orthonormal basis of balances

the intuitive approach

example: for $\mathbf{x} \in \mathcal{S}^5$ define a sequential binary partition and obtain the coordinates in the corresponding orthonormal basis

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

Balances and balancing elements

Coordinates in an orthonormal basis obtained from a sequential binary partition:

$$y_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \ln \frac{(\prod_{j \in R_i} x_j)^{1/r_i}}{(\prod_{\ell \in S_i} x_\ell)^{1/s_i}}$$

where $i =$ order of partition, R_i and S_i index sets, r_i the number of indices in R_i, s_i the number in S_i
The corresponding balancing element is

$$\mathbf{e}_i = \mathcal{C}(\exp[\psi_{i1}, \psi_{i2}, \ldots, \psi_{iD}])$$

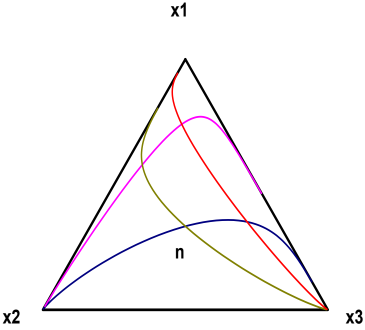
$$\psi_{i+} = +\sqrt{\frac{s_i}{r_i(r_i + s_i)}} \quad , \quad \psi_{i-} = -\sqrt{\frac{r_i}{s_i(r_i + s_i)}} \quad , \quad \psi_{i0} = 0$$

parallel lines

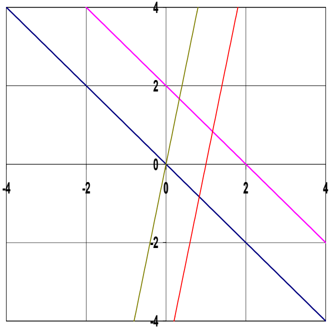
Processes of exponential growth or decay are straight-lines:

$$x_i(t) = x_i(0) \cdot \exp(\lambda_i t) \ , \ i = 1, 2, \ldots, D$$

$$\mathbf{x}(t) = \mathbf{x}(0) \oplus (t \odot \exp(\boldsymbol{\lambda}))$$

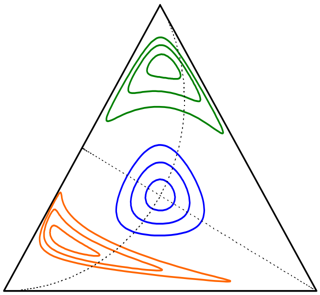


in $\mathcal{S}^3$

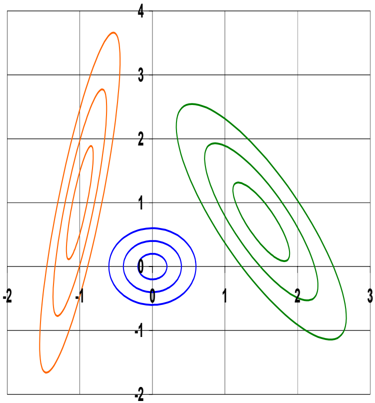


coordinate representation

circles and ellipses

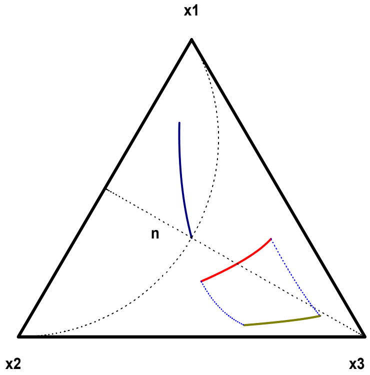


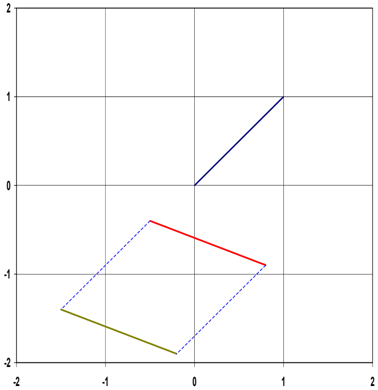
in $\mathcal{S}^3$



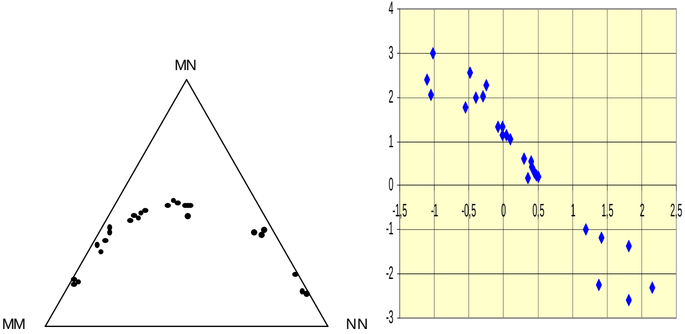
coordinate representation

circles and other geometric figures





Example: genetic hypothesis (Hardy-Weinberg)



data: genotypes in the MN system of blood groups;

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

Example of orthogonal coordinates (using SBP)

Votes in a district.
Left wing parties L_i and right wing parties R_i

level	L_1	L_2	R_1	R_2	L_3	L_4	r	s
1	$+1$	$+1$	-1	-1	$+1$	$+1$	4	2
2	$+1$	-1	0	0	-1	-1	1	3
3	0	$+1$	0	0	-1	-1	1	2
4	0	0	0	0	$+1$	-1	1	1
5	0	0	-1	$+1$	0	0	1	1
1	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{3}}$	$-\frac{1}{\sqrt{3}}$	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$		
2	$+\frac{\sqrt{3}}{2}$	$-\frac{1}{\sqrt{12}}$	0	0	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$		Ψ
3	0	$+\frac{\sqrt{2}}{\sqrt{3}}$	0	0	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$		
4	0	0	0	0	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$		
5	0	0	$+\frac{1}{\sqrt{2}}$	0	0	$-\frac{1}{\sqrt{2}}$		

Balances and projections

What information conveys a balance (of two groups of parts)?
Information which remains after:

- Removing information not within the subcomposition made of the two groups
- Filter out information within each group of parts

The remaining information is the balance
This is equivalent to set all balances to zero, except that one corresponding to the separation of the two groups

Elections example

- Only left-right: only balance 1
  The projection is:

$$\langle \mathbf{x}, \mathbf{e}_1 \rangle_a = \mathcal{C}[g(L), g(L), g(R), g(R), g(L), g(L)]$$

- Information only within the L group:
  balances 2,3,4
  (balance 1, between $L-R$ groups; balance 5, within R)
  The projection is

$$\bigoplus_{i=2,3,4} \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \mathcal{C}[L_1, L_2, g(L), g(L), L_3, L_4]$$

Elections example

- Only left-right: only balance 1
  The projection is:

$$\langle \mathbf{x}, \mathbf{e}_1 \rangle_a = \mathcal{C}[g(L), g(L), g(R), g(R), g(L), g(L)]$$

- Information only within the L group:
  balances 2,3,4
  (balance 1, between $L - R$ groups; balance 5, within R)
  The projection is

$$\bigoplus_{i=2,3,4} \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \mathcal{C}[L_1, L_2, g(L), g(L), L_3, L_4]$$

Elections example (continued)

- Remove info within R: balance 5 null
  The projection is:

$$\bigoplus_{i=1}^{4} \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \mathcal{C}[L_1, L_2, g(R), g(R), L_3, L_4]$$

  Information from $L - R$ balance is still in the projection.

- Assume L_2, L_3, L_4 are nationalist (I); and L_1 is not nationalist (I);
  Examine balance $LI - LN$: only balance 2
  The projection is:

$$\langle \mathbf{x}, \mathbf{e}_2 \rangle_a = \mathcal{C}[g(LI), g(LN), g(L), g(L), g(LN), g(LN)]$$

Elections example (continued)

- Remove info within R: balance 5 null
  The projection is:

$$\bigoplus_{i=1}^{4} \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \mathcal{C}[L_1, L_2, g(R), g(R), L_3, L_4]$$

Information from $L - R$ balance is still in the projection.

- Assume L_2, L_3, L_4 are nationalist (I); and L_1 is not nationalist (I);
  Examine balance $LI - LN$: only balance 2
  The projection is:

$$\langle \mathbf{x}, \mathbf{e}_2 \rangle_a = \mathcal{C}[g(LI), g(LN), g(L), g(L), g(LN), g(LN)]$$