Statistical analysis of compositional data

G. Mateu-Figueras

Dep. d'Informàtica, Matemàtica Aplicada i Estadística
Universitat de Girona

February 26, 2014

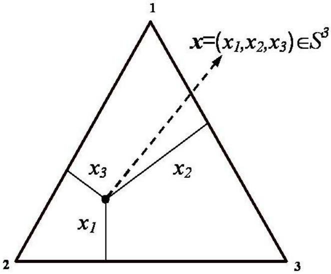
Outline

compositional data

- **compositional data** are parts of some whole which only carry **relative information**
- 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**

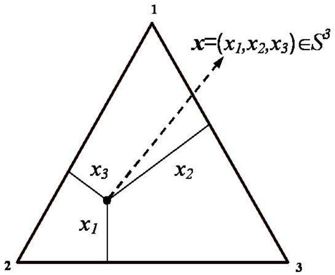


compositional data

- **compositional data** are parts of some whole which only carry **relative information**
- 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**

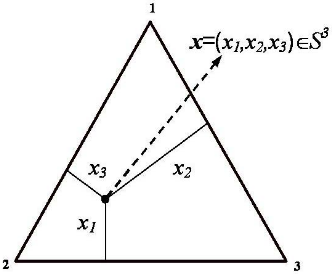


compositional data

- **compositional data** are parts of some whole which only carry **relative information**
- 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**



some compositional problems

- **MN blood system:** frequencies of MM, NN and MN blood types and the ethnic population. Despite the hight variability, is there any stability in the data? do they follow any genetic law?
- **elections to the Parlament de Catalunya:** the total votes achieved by each party in each counties. To characterize the regions.
- **skye lavas:** relative proportions of A ($Na_2O + K_2O$), F (Fe_2O_3) and M (MgO) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.

some compositional problems

- **MN blood system:** frequencies of MM, NN and MN blood types and the ethnic population. Despite the hight variability, is there any stability in the data? do they follow any genetic law?
- **elections to the Parlament de Catalunya:** the total votes achieved by each party in each counties. To characterize the regions.
- **skye lavas:** relative proportions of A ($Na_2O + K_2O$), F (Fe_2O_3) and M (MgO) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.

some compositional problems

- **MN blood system:** frequencies of MM, NN and MN blood types and the ethnic population. Despite the hight variability, is there any stability in the data? do they follow any genetic law?
- **elections to the Parlament de Catalunya:** the total votes achieved by each party in each counties. To characterize the regions.
- **skye lavas:** relative proportions of A ($Na_2O + K_2O$), F (Fe_2O_3) and M (MgO) of 23 basalt specimens from the Isle of Skye. To describe the variability of the geochemical composition.

spurious correlations (Pearson, 1897)

$$\mathbf{x} = (x_1, \ldots, x_D) \quad \sum_{i=1}^{D} x_i = \kappa \qquad cov(x_i, x_1) + \cdots + cov(x_i, x_D) = 0$$

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

cov	x_1	x_2	x_3	x_4
x_1	0.007	0.003	0.000	-0.010
x_2	0.003	0.002	-0.002	-0.003
x_3	0.000	-0.002	0.009	-0.007
x_4	-0.010	-0.003	-0.007	0.020

corr	x_1	x_2	x_3	x_4
x_1	1.000	0.866	0.000	-0.866
x_2	0.866	1.000	-0.500	-0.500
x_3	0.000	-0.500	1.000	-0.500
x_4	-0.866	-0.500	-0.500	1.000

spurious correlations (Pearson, 1897)

$$\mathbf{x} = (x_1, \ldots, x_D) \quad \sum_{i=1}^{D} x_i = \kappa \qquad cov(x_i, x_1) + \cdots + cov(x_i, x_D) = 0$$

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

cov	x_1	x_2	x_3	x_4
x_1	0.007	0.003	0.000	-0.010
x_2	0.003	0.002	-0.002	-0.003
x_3	0.000	-0.002	0.009	-0.007
x_4	-0.010	-0.003	-0.007	0.020

corr	x_1	x_2	x_3	x_4
x_1	1.000	0.866	0.000	-0.866
x_2	0.866	1.000	-0.500	-0.500
x_3	0.000	-0.500	1.000	-0.500
x_4	-0.866	-0.500	-0.500	1.000

subcompositional incoherence (Aitchison, 1997)

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

subcompositional incoherence (Aitchison, 1997)

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

subcompositional incoherence (Aitchison, 1997)

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

principles

- **scale invariance:** the analysis should not depend on the closure constant κ

$$f(\alpha \mathbf{x}) = f(\mathbf{x}) \quad , \quad \alpha > 0$$

- **subcompositional coherence:** studies performed on subcompositions should not stand in contradiction with those performed on the full composition

principles

- **scale invariance:** the analysis should not depend on the closure constant κ

$$f(\alpha \mathbf{x}) = f(\mathbf{x}) \quad , \quad \alpha > 0$$

- **subcompositional coherence:** studies performed on subcompositions should not stand in contradiction with those performed on the full composition

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

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

- **perturbation**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)$
- **powering**: $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)$
- **inner product**:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

- associated **norm** and **distance**:

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

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

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

- **perturbation**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)$
- **powering**: $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)$
- **inner product**:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

- associated **norm** and **distance**:

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

orthonormal coordinates

- **orthonormal basis** on $\mathcal{S}^D$: $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ (not unique)
- **coordinates** in this basis for $\mathbf{x} \in \mathcal{S}^D$ or **ilr** coordinates
  $\mathbf{x}^* = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)$
- example:
  $\mathbf{e}_1 = \mathcal{C}(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}})), \quad \mathbf{e}_2 = \mathcal{C}(\exp(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0))$

$$\mathbf{x}^* = \left(\sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}, \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \right)$$

Egozcue et al. (2003)

- compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **the principle of working on coordinates**

orthonormal coordinates

- **orthonormal basis** on $\mathcal{S}^D$: $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ (not unique)
- **coordinates** in this basis for $\mathbf{x} \in \mathcal{S}^D$ or **ilr** coordinates
  $\mathbf{x}^* = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)$
- example:
  $\mathbf{e}_1 = \mathcal{C}(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}})), \quad \mathbf{e}_2 = \mathcal{C}(\exp(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0))$

$$\mathbf{x}^* = \left(\sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}, \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \right)$$

Egozcue et al. (2003)

- compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **the principle of working on coordinates**

orthonormal coordinates

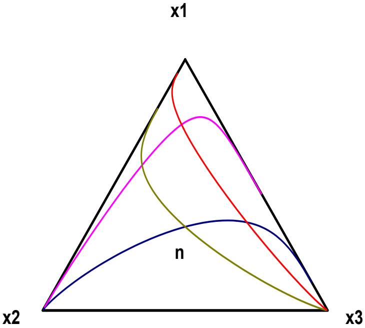
- **orthonormal basis** on $\mathcal{S}^D$: $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ (not unique)
- **coordinates** in this basis for $\mathbf{x} \in \mathcal{S}^D$ or **ilr** coordinates
  $\mathbf{x}^* = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \dots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)$
- example:
  $\mathbf{e}_1 = \mathcal{C}(\exp(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}})), \quad \mathbf{e}_2 = \mathcal{C}(\exp(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0))$

$$\mathbf{x}^* = \left(\sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}, \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \right)$$

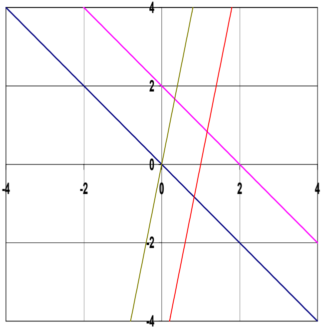
Egozcue et al. (2003)

- compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates
- **the principle of working on coordinates**

parallel lines

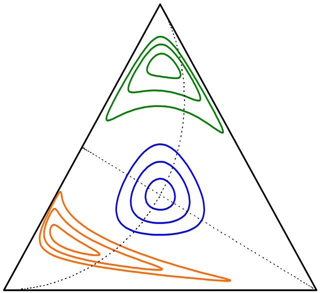


in $\mathcal{S}^3$

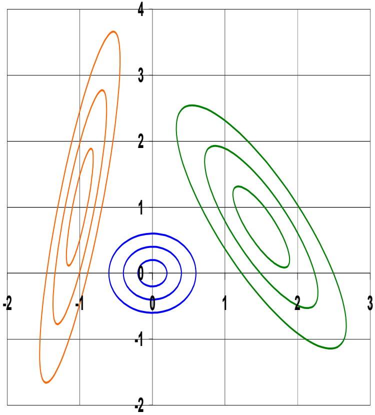


coordinate representation

circles and ellipses

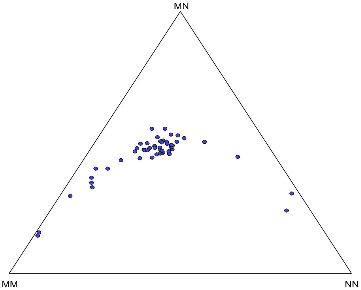


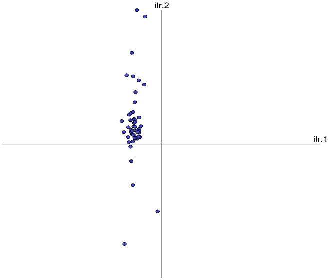
in $\mathcal{S}^3$



coordinate representation

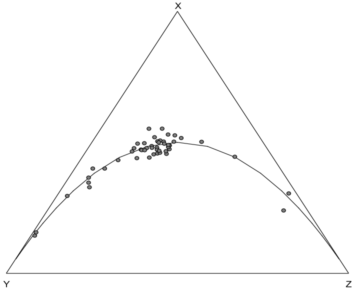
the MN blood system

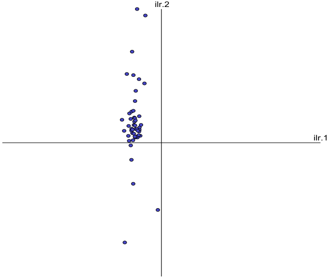




$$\sqrt{\frac{2}{3}}\ \ln \frac{(MM \cdot NN)^{1/2}}{MN} = -0.57$$

the MN blood system





Hardy-Weinberg law: $MN^2 = 4MM \cdot NN$

$$\sqrt{\frac{2}{3}} \ln \frac{(MM \cdot NN)^{1/2}}{MN} = -0.57$$

building an orthonormal basis

using sequential binary partitions (SBP)

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

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

coordinates ⇒ balances

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

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

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

Egozcue, Pawlowsky-Glahn (2005)

Log-ratio approach (Aitchison, 1980-86)

log-ratio transformations introduced by J. Aitchison:

- **alr**: $\mathcal{S}^D \to \mathbb{R}^{D-1}$, $\quad \mathrm{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right)$

  drawback: not an isometry

- **clr**: $\mathcal{S}^D \to \mathbb{R}^D$, $\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)$,

  $$g(\mathbf{x}) = \prod_{i=1}^{D} x_i^{1/D}$$

  drawback: a constrained transformed vector

Log-ratio approach (Aitchison, 1980-86)

log-ratio transformations introduced by J. Aitchison:

- **alr**: $\mathcal{S}^D \to \mathbb{R}^{D-1}$, $\quad \mathrm{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right)$

  drawback: not an isometry
- **clr**: $\mathcal{S}^D \to \mathbb{R}^D$, $\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)$,

  $g(\mathbf{x}) = \prod_{i=1}^{D} x_i^{1/D}$

  drawback: a constrained transformed vector

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

for a review, see Martín-Fernández et al. (2011)

center and variability

let $\mathbf{X} = \{\mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in \mathcal{S}^D : i = 1, \ldots, n\}$

- **center** (closed geometric mean) of $\mathbf{X}$:

$$\mathbf{g} = \mathcal{C}(g_1, g_2, \ldots, g_D), \text{ with } g_j = \left(\prod_{i=1}^{n} x_{ij} \right)^{1/n}$$

- **total variance** of $\mathbf{X}$: $\quad \mathrm{TotVar}[\mathbf{X}] = \frac{1}{n} \sum_{i=1}^{n} d_a^2(\mathbf{x}_i, \mathbf{g})$
- **variation array** of $\mathbf{X}$:

$$\begin{pmatrix} - & \mathrm{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_1}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_2}\right] & - & \ddots & \vdots \\ \vdots & \ddots & - & \mathrm{var}\left[\ln \frac{x_{D-1}}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_D}\right] & \cdots & \mathrm{E}\left[\ln \frac{x_{D-1}}{x_D}\right] & - \end{pmatrix}$$

center and variability

let $\mathbf{X} = \{\mathbf{x}_i = (x_{i1}, \dots, x_{iD}) \in \mathcal{S}^D : i = 1, \dots, n\}$

- **center** (closed geometric mean) of $\mathbf{X}$:

$$\mathbf{g} = \mathcal{C}(g_1, g_2, \dots, g_D), \text{ with } g_j = \left(\prod_{i=1}^{n} x_{ij} \right)^{1/n}$$

- **total variance** of $\mathbf{X}$: $\quad \mathrm{TotVar}[\mathbf{X}] = \frac{1}{n} \sum_{i=1}^{n} d_a^2(\mathbf{x}_i, \mathbf{g})$

- **variation array** of $\mathbf{X}$:

$$\begin{pmatrix} - & \mathrm{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_1}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_2}\right] & - & \ddots & \vdots \\ \vdots & \ddots & - & \mathrm{var}\left[\ln \frac{x_{D-1}}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_D}\right] & \cdots & \mathrm{E}\left[\ln \frac{x_{D-1}}{x_D}\right] & - \end{pmatrix}$$

center and variability

let $\mathbf{X} = \{\mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in \mathcal{S}^D : i = 1, \ldots, n\}$

- **center** (closed geometric mean) of $\mathbf{X}$:

$$\mathbf{g} = \mathcal{C}(g_1, g_2, \ldots, g_D), \text{ with } g_j = \left(\prod_{i=1}^{n} x_{ij} \right)^{1/n}$$

- **total variance** of $\mathbf{X}$: $\mathrm{TotVar}[\mathbf{X}] = \frac{1}{n} \sum_{i=1}^{n} d_a^2(\mathbf{x}_i, \mathbf{g})$
- **variation array** of $\mathbf{X}$:

$$\begin{pmatrix} - & \mathrm{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_1}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_2}\right] & - & \ddots & \vdots \\ \vdots & \ddots & - & \mathrm{var}\left[\ln \frac{x_{D-1}}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_D}\right] & \cdots & \mathrm{E}\left[\ln \frac{x_{D-1}}{x_D}\right] & - \end{pmatrix}$$

example: ParlCat2010 data set

votes achieved by PP, CiU, SI, C's, ERC, PSC, ICV

$$\mathbf{g} = (0.097, 0.505, 0.044, 0.017, 0.102, 0.179, 0.056)$$

Variation array:

Variance ln(Xi/Xj)

Xi\Xj	P1S_PP	P2C_CiU	P3C_SI	P4S_Cs	P5C_ERC	P6S_PSC	P7S_ICV	clr variances
P1S_PP		0.2839	0.6362	0.1580	0.4618	0.1161	0.1852	0.1244
P2C_CiU	1.6503		0.1860	0.5452	0.0732	0.1639	0.1597	0.0631
P3C_SI	-0.7934	-2.4436		0.8915	0.1386	0.4575	0.3146	0.2363
P4S_Cs	-1.7543	-3.4045	-0.9609		0.8344	0.3118	0.2582	0.2898
P5C_ERC	0.0491	-1.6012	0.8424	1.8033		0.2732	0.2434	0.1506
P6S_PSC	0.6154	-1.0349	1.4087	2.3696	0.5663		0.1015	0.0648
P7S_ICV	-0.5464	-2.1967	0.2470	1.2079	-0.5955	-1.1618		0.0417

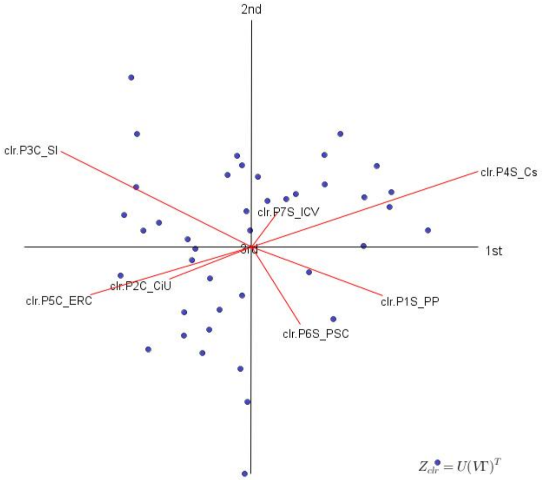
Mean ln(Xi/Xj) — 0.9705 Total Variance

clr biplot

- graphical display of a multivariate data set (individuals and variables)
- **clr**-biplot
- particular **rules of interpretation**
  - $\|ray\| \approx$ variance clr component
  - $\|link\| \approx$ variance logratio
  - perpendicular links $\Rightarrow$ possible incorrelated logratios
  - parallel links $\Rightarrow$ possible hight correlated logratios
  - coincident vertices $\Rightarrow$ two redundant parts
  - collinear vertices $\Rightarrow$ possible one-dimensional variability

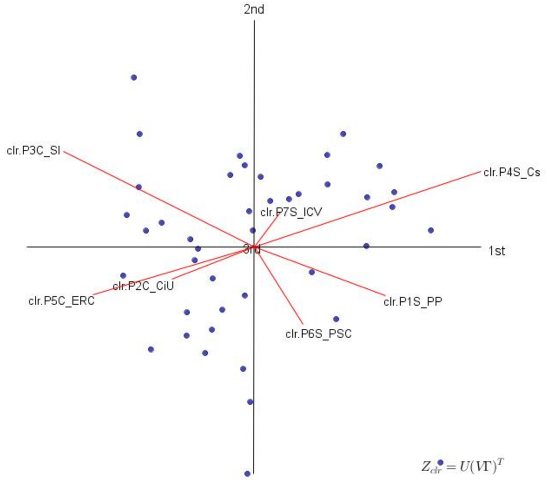
Aitchison and Greenacre (2002)

example: ParlCat2010 data set (explains 86% variance)



$$var\left(\ln\left(\frac{ICV}{g}\right)\right) = 0.0417 \qquad var\left(\ln\left(\frac{C's}{g}\right)\right) = 0.2898$$

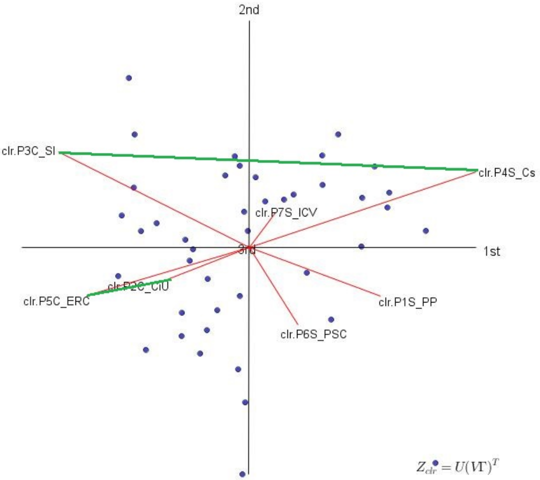
example: ParlCat2010 data set (explains 86% variance)



$$var\left(\ln\left(\frac{ICV}{\mathbf{g}}\right)\right) = 0.0417 \qquad var\left(\ln\left(\frac{C's}{\mathbf{g}}\right)\right) = 0.2898$$

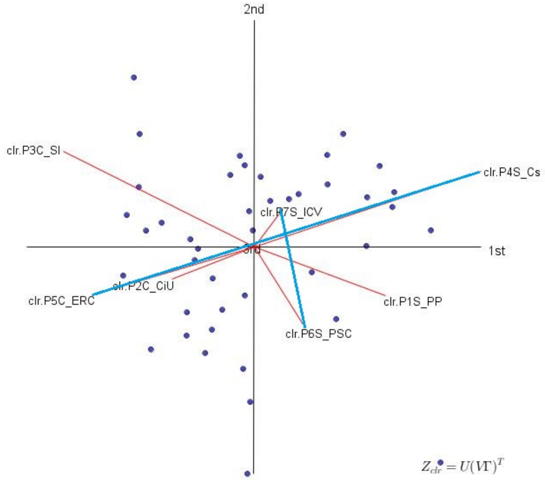
example: ParlCat2010 data set

(explains 86% variance)



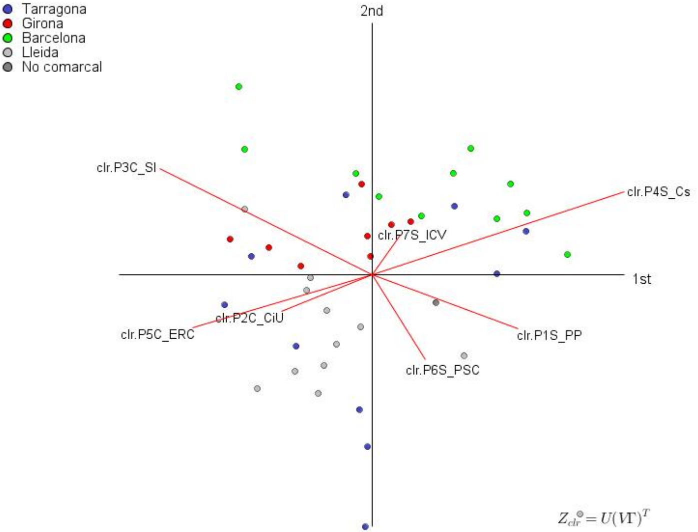
$$var\left(\ln\left(\frac{SI}{C's}\right)\right) = 0.8915 \qquad var\left(\ln\left(\frac{CiU}{ERC}\right)\right) = 0.0732$$

example: ParlCat2010 data set (explains 86% variance)



$$corr\left(\ln\left(\frac{C's}{ERC}\right), \ln\left(\frac{PSC}{ICV}\right)\right) = -0.041$$

example: ParlCat2010 data set (explains 86% variance)



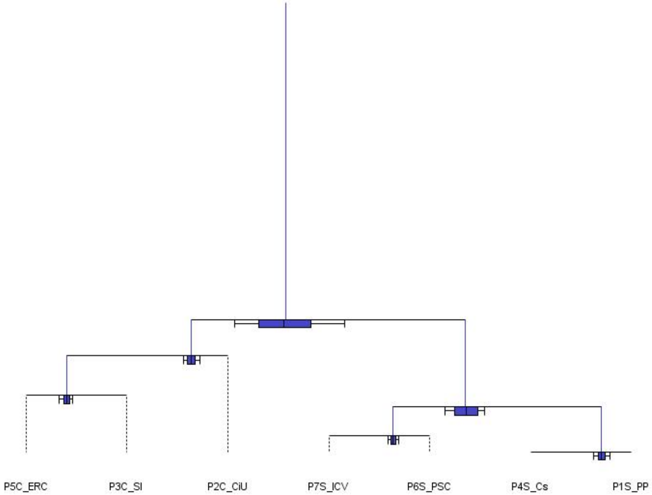
coda-dendrogram

to visualize

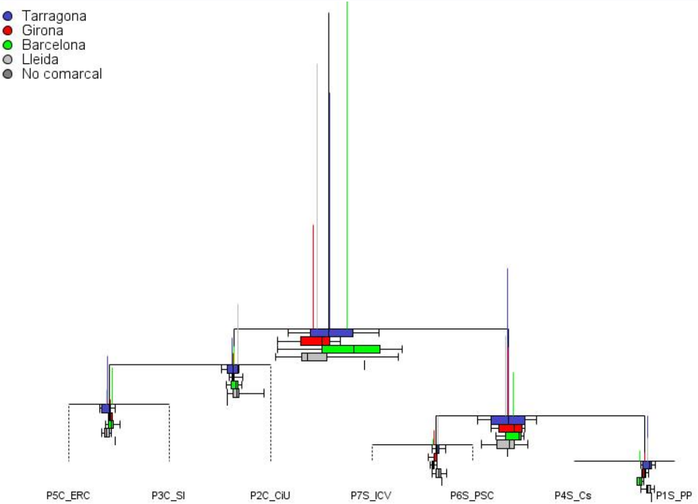
- sequential binary **partition**
- **center** of each balance
- proportion of the sample total **variance** corresponding to each balance.
- **summary statistics** of each balance (box-plot of percentiles 5, 25, 50, 75, 95)
- adequate to represent different **groups**

Pawlowsky-Glahn and Egozcue (2011)

example: ParlCat2010 data set



example: ParlCat2010 data set



logistic normal (Aitchison 1980-86)

$$\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$$

- **transform** $\mathbf{x}$ to $\mathbb{R}^{D-1}$ using a log-ratio transformation
- define the density of the **transformed vector** and go back to $\mathcal{S}^D$ using the **change of variable** theorem
- the result is a density function for $\mathbf{x}$ with respect to λ on $\mathcal{S}^D$

$$\Downarrow$$

(Aitchison, 1997)

$E[\mathbf{x}]$ is not a meaningful measure of central location

$\mathrm{cen}[\mathbf{x}]$ is the alternative which minimizes $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$

logistic normal (Aitchison 1980-86)

$$\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$$

- **transform** $\mathbf{x}$ to $\mathbb{R}^{D-1}$ using a log-ratio transformation
- define the density of the **transformed vector** and go back to $\mathcal{S}^D$ using the **change of variable** theorem
- the result is a density function for $\mathbf{x}$ with respect to λ on $\mathcal{S}^D$

$$\Downarrow$$

(Aitchison, 1997)

$E[\mathbf{x}]$ is not a meaningful measure of central location
$\mathrm{cen}[\mathbf{x}]$ is the alternative which minimizes $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$

densities and measures

- on $\mathcal{S}^D$: density functions expressed with respect to the **Aitchison measure** λ_a
- density functions of the vector of **coordinates** with respect to λ.

$$d\lambda/d\lambda_a = \sqrt{D}\, x_1 x_2 \cdots x_D, \qquad \lambda_a(A) = \lambda(A^*)$$

densities and measures

- on $\mathcal{S}^D$: density functions expressed with respect to the **Aitchison measure** λ_a
- density functions of the vector of **coordinates** with respect to λ.

$$d\lambda/d\lambda_a = \sqrt{D}\, x_1 x_2 \cdots x_D, \qquad \lambda_a(A) = \lambda(A^*)$$

normal on $\mathcal{S}^D$

$$\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$$

a random composition $\mathbf{x}$ is **normally distributed on $\mathcal{S}^D$** with parameters μ and Σ if its density function is

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}^* - \mu^*\right)' \Sigma^{-1} \left(\mathbf{x}^* - \mu^*\right)\right]$$

usual **normal density** applied to coordinates $\mathbf{x}^*$ and $f_{\mathbf{x}} = \frac{dP}{d\lambda_a}$

$$\mu = \mathrm{E}_a[\mathbf{x}] = \mathrm{cen}[\mathbf{x}]$$

Mateu-Figueras et al (2013)

normal on $\mathcal{S}^D$

$$\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$$

a random composition $\mathbf{x}$ is **normally distributed on $\mathcal{S}^D$** with parameters μ and Σ if its density function is

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x}^* - \boldsymbol{\mu}^*)' \, \Sigma^{-1} \, (\mathbf{x}^* - \boldsymbol{\mu}^*) \right]$$

usual **normal density** applied to coordinates $\mathbf{x}^*$ and $f_{\mathbf{x}} = \frac{dP}{d\lambda_a}$

$$\mu = \mathrm{E}_a[\mathbf{x}] = \mathrm{cen}[\mathbf{x}]$$

Mateu-Figueras et al (2013)

normal on $\mathcal{S}^D$

$$\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$$

a random composition $\mathbf{x}$ is **normally distributed on $\mathcal{S}^D$** with parameters μ and Σ if its density function is

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}^* - \boldsymbol{\mu}^*\right)' \Sigma^{-1} \left(\mathbf{x}^* - \boldsymbol{\mu}^*\right)\right]$$

usual **normal density** applied to coordinates $\mathbf{x}^*$ and $f_{\mathbf{x}} = \frac{dP}{d\lambda_a}$

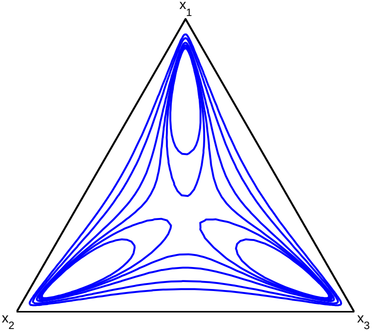
$$\boldsymbol{\mu} = \mathrm{E}_a[\mathbf{x}] = \mathrm{cen}[\mathbf{x}]$$

Mateu-Figueras et al (2013)

comparison

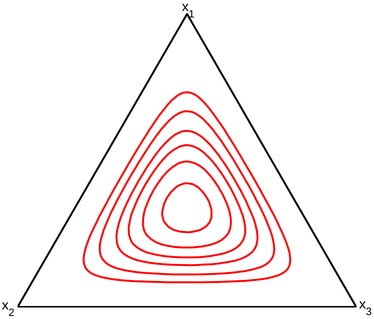
$$\mu^* = (0,0), \Sigma = Id$$

$\mathcal{S}^D \subset \mathbb{R}^D$



logistic normal
Lebesgue measure λ

$\mathcal{S}^D$ as Euclidian space

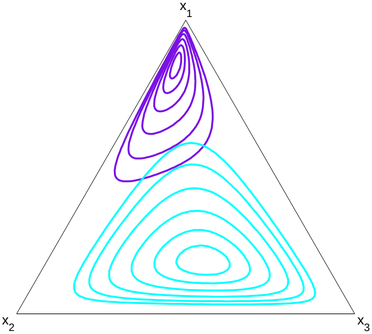


normal on $\mathcal{S}^D$
Aitchison measure λ_a

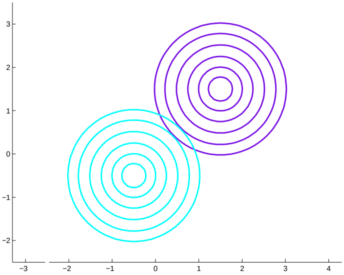
invariance under perturbation

p=(0.93, 0.05,0.02)

$$\mathbf{x}^* = \left(\frac{1}{\sqrt{2}} \ln\left(\frac{x_1}{x_2}\right), \frac{1}{\sqrt{6}} \ln\left(\frac{x_1 x_2}{x_3 x_3}\right) \right)$$



normal on $\mathcal{S}^3$



coordinate representation

$\mu^* = (-0.5, -0.5)$, $\mu^* = (1.5, 1.5)$, $\Sigma = Id$

tests of normality on $\mathcal{S}^D$

H_0: the sample of **coordinates** comes from a multivariate normal distribution

- based on empirical distribution function **(EDF) tests**
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D-1)$ marginal, univariate distributions
  - all $(D-1)(D-2)/2$ bivariate angle distributions
  - the $(D-1)$-dimensional radius distribution
- problem: **dependence** of the orthonormal basis

tests of normality on $\mathcal{S}^D$

H_0: the sample of **coordinates** comes from a multivariate normal distribution

- based on empirical distribution function **(EDF) tests**
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D-1)$ marginal, univariate distributions
  - all $(D-1)(D-2)/2$ bivariate angle distributions
  - the $(D-1)$-dimensional radius distribution
- problem: **dependence** of the orthonormal basis

tests of normality on $\mathcal{S}^D$

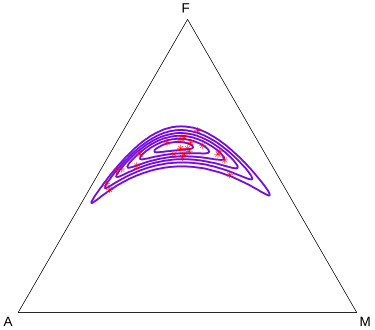
H_0: the sample of **coordinates** comes from a multivariate normal distribution

- based on empirical distribution function **(EDF) tests**
- Anderson-Darling, Cramer-von Mises and Watson statistics
- three possible cases
  - all $(D-1)$ marginal, univariate distributions
  - all $(D-1)(D-2)/2$ bivariate angle distributions
  - the $(D-1)$-dimensional radius distribution
- problem: **dependence** of the orthonormal basis

example: aphyric Skye lavas

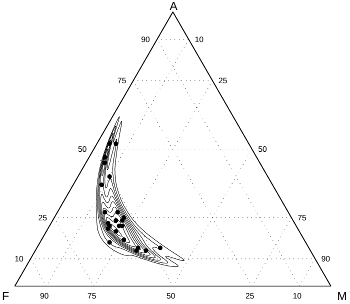
X=(A,F,M) composition of 23 basalt specimens from the Isle of Skye (Aitchison,1986)

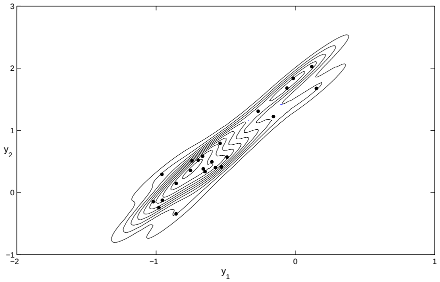
$$\widehat{\mu}^* = (0.555, 0.639) \qquad \widehat{\Sigma} = \begin{pmatrix} 0.126 & -0.229 \\ -0.229 & 0.456 \end{pmatrix}$$



kernel density estimation

- the normal on $\mathcal{S}^D$ for the **kernel** in the density estimator
- **invariance** with respect to the orthonormal basis





Chacón et al (2010)

other distributions on $\mathcal{S}^D$

- the skew-normal distribution on $\mathcal{S}^D$
- the Dirichlet distribution
- the shifted-scaled Dirichlet distribution
- ...

conclusions

- treat compositional data (CoDa) in the **simplex**, with its specific geometry
- do **not** apply ordinary multivariate statistics **directly** to CoDa
- the simplex has an **Euclidean** structure: **orthonormal coordinates** are available
- multivariate statistical models and methods **work properly** on coordinates of CoDa
- problem (or advantage): **interpretation** of coordinates

references

Aitchison, J. (1986): The statistical analysis of compositional data. Monographs on statistics and applied Probability: Chapman and Hall, London.
Aitchison, J., Greenacre, M. (2002): Biplots for compositional data .Journal of the Royal Statistical Society, Series C (Applied Statistics) 51 (4), 375–392. 2002
Billheimer, D.; Guttorp, P.; Fagan, W. (2001): Statistical interpretation of species composition.*J. Am. Statistical Ass.*, 96(456), 1205–1214.
Chacón, J.E.; Mateu-Figueras, G.; Martín-Fernández, J.A. (2010): Gaussian kernels for density estimation with compositional data.*Computers and Geosciences.*, 37, 702–711.
Egozcue, J.J.; Pawlowsky-Glahn, V. (2005): Groups of parts and their balances in compositional data analysis. *Math. Geol.*, 37(7), 795–828.
Egozcue, J.J.; Pawlowsky-Glahn, V., Mateu-Figueras, G.; Barceló-Vidal, C. (2003): Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.
Martín-Fernández, J.A.; Palarea-Albaladejo, J.; Olea, R.A. (2011): Dealing with zeros. In Pawlowsky-Glahn, V. and Buccianti A. (Eds.) *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester UK.
Mateu-Figueras, G.; Pawlowsky-Glahn, V.; Egozcue, J.J. (2013): The normal distribution in some constrained sample spaces. *SORT*, 37(1),29-56.
Pawlowsky-Glahn, V.; Egozcue, J.J. (2001): Geometric approach to statistical analysis on the simplex. *SERRA*, 15(5), 384–398.
Pawlowsky-Glahn, V.; Egozcue, J.J. (2011): Exploring Compositional Data with the Coda-Dendrogram, *Austrian Journal of Statistics*, 40, 1-2.