

Compositional Data Analysis in a Nutshell

report errors to: Raimon Tolosana-Delgado, raimon.tolosana@geo.uni-goettingen.de

version: November 10, 2008

Geometry

Characteristics

- Compositional data are vectors of non-negative components showing the *relative* weight or importance of a set of *parts in a total*.
- The total sum of a compositional vector is *considered* irrelevant, or an artifact of the sampling procedure.
- No individual component can be interpreted isolated from the other. A composition carries no absolute information on increment/decrement of mass.
- The sample space (or set of possible values) is called the *simplex*: this is the set of vectors of positive (or zero) components and constant sum:

$$S^D = \{\mathbf{x} = [x_1; \dots; x_D] | x_i \geq 0 \text{ and } \sum_{j=1}^D x_j = \kappa\}$$

with $\kappa = 1, 100, 10^6, 10^9$ (proportions, %, ppm, ppb), etc.

Compositional operations

Take $\mathbf{x} = [x_1, \dots, x_D]$, $\mathbf{y} = [y_1, \dots, y_D]$, $\mathbf{z} = [z_1, \dots, z_D]$ compositions of D parts, and λ a real value. The compositional operations are

- closure:

$$\mathbf{x} = \mathcal{C}[\mathbf{x}'] = \frac{\kappa}{\sum_{i=1}^D x'_i} \mathbf{x}'$$

- perturbation (replacing sum and subtraction):

$$\begin{aligned} \mathbf{z} &= \mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 \cdot y_1; \dots; x_D \cdot y_D] \\ \mathbf{z} &= \mathbf{x} \ominus \mathbf{y} = \mathcal{C}[x_1/y_1; \dots; x_D/y_D] \end{aligned}$$

- power transformation (replacing scaling):

$$\mathbf{z} = \lambda \odot \mathbf{x} = \mathcal{C}[x_1^\lambda; \dots; x_D^\lambda]$$

- Aitchison scalar product (repl. dot product):

$$\langle \mathbf{x} | \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

- Aitchison distance (repl. Euclidean distance):

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

Log-ratio transformations

- additive log-ratio transform (and inverse)

$$\begin{aligned} \text{alr}(\mathbf{x}) &= \mathbf{y} = \left[\ln \frac{x_1}{x_D}; \dots; \ln \frac{x_{D-1}}{x_D} \right] = \\ &= \ln(\mathbf{x}) \cdot \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{pmatrix} \end{aligned}$$

$$\text{alr}^{-1}(\mathbf{y}) = \mathcal{C}[\exp(\mathbf{y}; 0)]$$

- centered log-ratio transform ($g(\mathbf{x}) = \sqrt[D]{x_1 \cdots x_D}$)

$$\begin{aligned} \text{clr}(\mathbf{x}) &= \mathbf{z} = \left[\ln \frac{x_1}{g(\mathbf{x})}; \dots; \ln \frac{x_D}{g(\mathbf{x})} \right] \\ &= \frac{\ln(\mathbf{x})}{D} \cdot \begin{pmatrix} D-1 & -1 & \dots & -1 \\ -1 & D-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & D-1 \end{pmatrix} \end{aligned}$$

$$\text{clr}^{-1}(\mathbf{z}) = \mathcal{C}[\exp(\mathbf{z})]$$

- isometric log-ratio transform

$$\text{ilr}_V(\mathbf{x}) = \text{clr}(\mathbf{x}) \cdot \mathbf{V} = \ln(\mathbf{x}) \cdot \mathbf{V},$$

for a given matrix \mathbf{V} of D rows and $(D-1)$ columns such that $\mathbf{V} \cdot \mathbf{V}^t = \mathbf{I}_{D-1}$ (identity matrix of $D-1$ elements) and $\mathbf{V} \cdot \mathbf{V}^t = \mathbf{I}_D + a\mathbf{1}$, where a may be any value, and $\mathbf{1}$ is a matrix full of ones. The inverse is

$$\text{ilr}_V^{-1}(\mathbf{x}) = \mathcal{C}[\exp(\mathbf{x} \cdot \mathbf{V}^t)].$$

- examples for $D = 3$:

$$\text{alr}(\mathbf{x}) = [y_1; y_2] = \left[\ln \frac{x_1}{x_3}; \ln \frac{x_2}{x_3} \right]$$

$$\mathbf{x} = \frac{[\exp(y_1); \exp(y_2); 1]}{\exp(y_1) + \exp(y_2) + 1}$$

$$\text{clr}_i(\mathbf{x}) = z_i = \ln \frac{x_i}{\sqrt[3]{x_1 x_2 x_3}}$$

$$x_i = \frac{\exp(z_i)}{\exp(z_1) + \exp(z_2) + \exp(z_3)}$$

$$\text{ilr}_V(\mathbf{x}) = \left[\frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}; \frac{1}{\sqrt{6}} \ln \frac{x_1^2}{x_2 x_3} \right]$$

$$\mathbf{V} = \begin{pmatrix} 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \end{pmatrix}$$

Statistics

Descriptive statistics

Take \mathbf{X} as a compositional data set, with N rows (individuals) and D columns (compositional variables). Notation $\ast\text{lr}$ means one of the log-ratio transforms.

center (repl. average)

$$\text{Mean}_A[\mathbf{X}] = \text{clr}^{-1}(\text{Mean}[\ln \mathbf{X}]) = \ast\text{lr}^{-1}(\text{Mean}[\ast\text{lr}(\mathbf{X})])$$

- **centering**: $\mathbf{X}' = \mathbf{X} \ominus \text{Mean}_A[\mathbf{X}]$

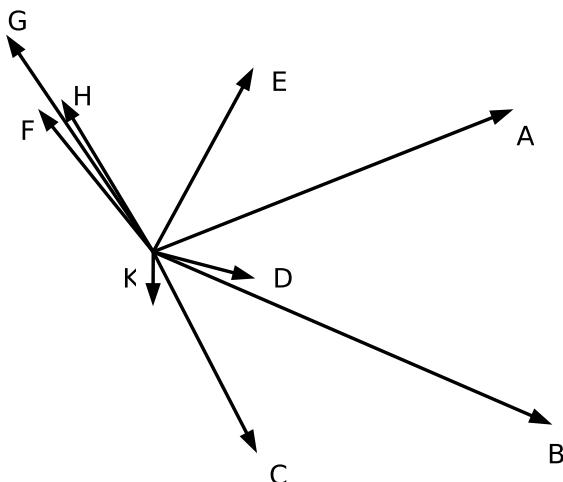
variation matrix (repl. correlation) $\mathbf{T} = [\tau_{ij}]$ with

$$\tau_{ij} = \text{Var} \left[\ln \frac{x_i}{x_j} \right]$$

- if $\tau_{ij} \rightarrow 0$, then $\ln(x_i/x_j) \approx \text{constant}$, then x_i and x_j proportional
- larger τ_{ij} , less proportional x_i and x_j

$\ast\text{lr}$ -variance matrix (repl. covariance) $\text{Var}[\ast\text{lr}(\mathbf{X})]$
(no back-transformation, difficult to interpret)

Compositional biplot



Best 2D simultaneous representation of data variability and relationships between variables; linked to principal components of the covariance matrix of a centered clr -transformed data set:

- **warning**: do not interpret rays; focus on links
- short link: small t_{ij} , x_i and x_j proportional (FH)
- 3 separate, very long rays: subcomposition defining a high-variance ternary diagram (ABG)
- collinear links: subcomposition showing a one-dimensional pattern (AFH, AEG or CDE)
- orthogonal links: the two subcompositions are uncorrelated (AFH vs. CDE)

Normal inference on the simplex

Normal on the simplex: normal distribution of a $\ast\text{lr}$ -transformed composition, with parameters: a central composition \mathbf{x} and a dispersion (positive-semidefinite symmetric) matrix Σ of eigen-decomposition $\Sigma = \mathbf{V} \cdot \Lambda \cdot \mathbf{V}^t$:

$$\mathbf{x} \sim \mathcal{N}_S^D(\mathbf{m}, \Sigma) \Leftrightarrow -2 \ln f(\mathbf{x}|\mathbf{m}, \Sigma) = (D-1) \ln(2\pi) + \sum_{i=1}^{D-1} \ln \lambda_i + \text{ilr}_V(\mathbf{x} \ominus \mathbf{m}) \cdot \Lambda^{-1} \cdot \text{ilr}_V^t(\mathbf{x} \ominus \mathbf{m}),$$

where $\text{ilr}_V(\cdot)$ is the ilr with matrix \mathbf{V} giving the eigenvectors in columns, and λ_i are the diagonal elements of Λ , the non-zero eigenvalues of Σ .

Given \mathbf{m} and Σ mean composition and dispersion matrix (theoretical or estimated)

- Regions on a ternary diagram ($D = 3$): ellipses, centered on \mathbf{m} , with principal axes along the eigenvectors of the columns of \mathbf{V} , semiaxes $\sqrt{\lambda_i}$ and radius r :

$$\begin{aligned} & - (1 - \alpha)\text{-probability regions for observations, } r = \sqrt{\chi_\alpha^2(2)} \\ & - (1 - \alpha)\text{-confidence regions on the mean, } r = \sqrt{\mathcal{F}_\alpha(2, N-2) \cdot 2/(N-2)}. \end{aligned}$$

- Test statistic on equivalence of population of two groups, with \mathbf{m}_i and Σ_i center and dispersion in group i :

$$Q(\mathbf{X}) = N \ln |\Sigma_0| - N_1 \ln |\Sigma_1| - N_2 \ln |\Sigma_2| \sim \chi^2(\nu)$$

1. = center, = dispersion: $\nu = D(D-1)/2$, and Σ_0 the joint covariance matrix (computed as if no groups existed)
2. \neq center, = dispersion: $\nu = (D-1)(D-2)/2$ and $\Sigma_0 = \frac{N_1}{N} \Sigma_1 + \frac{N_2}{N} \Sigma_2$ the pooled covariance matrix
3. = center, \neq dispersion: $\nu = (D-1)$; see lecture notes or book for Σ_0 expression;

$\ln |\Sigma|$ = log-determinant, computed as the sum of logs of the non-zero eigenvalues of Σ .

Most basic references

Grounding book: Aitchison, J. (1986) *The statistical analysis of compositional data*. Reprinted in 2003 by The Blackburn Press.

General paper: Pawlowsky-Glahn, V. (2003) Statistical modelling in coordinates. In: Proceedings of the 1st CoDaWork.

Lecture notes: Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2007) *Lecture notes on compositional data analysis*
<http://hdl.handle.net/10256/297>

Ongoing research several CoDaWork proceedings, available online at:
<http://dugi-doc.udg.edu/handle/10256/150>