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:

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

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

Compositional operations

Take \( x = [x_1, \ldots; x_D] \), \( y = [y_1, \ldots; y_D] \), \( z = [z_1, \ldots; z_D] \) compositions of \( D \) parts, and \( \lambda \) a real value. The compositional operations are

• closure:

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

• perturbation (replacing sum and subtraction):

\[ z = x \oplus y = C[x_1 \cdot y_1; \ldots; x_D \cdot y_D] \]
\[ z = x \odot y = C[x_1/y_1; \ldots; x_D/y_D] \]

• power transformation (replacing scaling):

\[ z = \lambda \odot x = C[x_1^\lambda; \ldots; x_D^\lambda] \]

• Aitchison scalar product (repl. dot product):

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

• Aitchison distance (repl. Euclidean distance):

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

Log-ratio transformations

• additive log-ratio transform (and inverse)

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

\[ \text{alr}^{-1}(y) = C[\exp(\langle [y]; 0 \rangle)] \]

• centered log-ratio transform \( (g(x) = \sqrt{x_1 \cdots x_D}) \)

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

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

• isometric log-ratio transform

\[ \text{ilr}(x) = \text{clr}(x) \cdot V = \ln(x) \cdot V \]

for a given matrix \( V \) of \( D \) rows and \( (D - 1) \) columns such that \( V \cdot V^t = I_{D-1} \) (identity matrix of D-1 elements) and \( V \cdot V^t = I_D + a1 \), where \( a \) may be any value, and \( I \) is a matrix full of ones.

The inverse is

\[ \text{ilr}^{-1}(x) = C[\exp(x \cdot V^t)] \]

• examples for \( D = 3 \):

\[ \text{alr}(x) = [y_1; y_2] = \begin{pmatrix} \ln \frac{x_1}{x_3}; \ln \frac{x_2}{x_3} \end{pmatrix} \]

\[ x = \begin{pmatrix} \exp(y_1); \exp(y_2); 1 \end{pmatrix} \cdot \text{exp}(y_1) + \text{exp}(y_2) + 1 \]

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

\[ x_i = \frac{\text{exp}(z_i)}{\exp(z_1) + \exp(z_2) + \exp(z_3)} \]

\[ \text{ilr}(x) = \begin{pmatrix} \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}; \frac{1}{\sqrt{6}} \ln \frac{x_1^2}{x_2 x_3} \end{pmatrix} \]

\[ V = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \]
Statistics

Descriptive statistics

Take \( \mathbf{X} \) as a compositional data set, with \( N \) rows (individuals) and \( D \) columns (compositional variables). Notation *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: \( \mathbf{X}' = \mathbf{X} \odot \text{Mean}_A[\mathbf{X}] \)

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

\[
\tau_{ij} = \text{Var} \left[ \ln \frac{x_i}{x_j} \right]
\]
• if \( \tau_{ij} \to 0 \), then \( \ln(x_i/x_j) \approx \text{constant} \), then \( x_i \) and \( x_j \) proportional
• larger \( \tau_{ij} \), less proportional \( x_i \) and \( x_j \)

*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 \( \Sigma \) of eigendecomposition \( \Sigma = \mathbf{V} \cdot \Lambda \cdot \mathbf{V}' \):

\[
\mathbf{x} \sim N_3^D(\mathbf{m}, \Sigma) \iff -2 \ln f(x|m, \Sigma) = (D - 1) \ln(2\pi) + \sum_{i=1}^{D-1} \ln \lambda_i + \text{ilr}_v(\mathbf{x} \odot \mathbf{m}) \cdot \Lambda^{-1} \cdot \text{ilr}_v(\mathbf{x} \odot \mathbf{m}),
\]

where \( \text{ilr}_v(\cdot) \) is the ilr with matrix \( \mathbf{V} \) giving the eigenvectors in columns, and \( \lambda_i \) are the diagonal elements of \( \Lambda \), the non-zero eigenvalues of \( \Sigma \).

Given \( \mathbf{m} \) and \( \Sigma \) 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 \( r \):

\[
- (1 - \alpha)-\text{probability regions for observations,} \quad r = \sqrt{\chi^2_{\alpha/2}(2)}
\]

\[
- (1 - \alpha)-\text{confidence regions on the mean,} \quad r = \sqrt{F_{\alpha/2}(2, N - 2)} \cdot \sqrt{2/(N - 2)}.
\]

• Test statistic on equivalence of population of two groups, with \( \mathbf{m}_1 \) and \( \Sigma_1 \) center and dispersion in group i:

\[
Q(X) = N \ln |\Sigma_0| - N_1 \ln |\Sigma_1| - N_2 \ln |\Sigma_2| \sim \chi^2(\nu)
\]

1. \( = \) center, \( = \) dispersion: \( \nu = D(D - 1)/2 \), and \( \Sigma_0 \) 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 \( \Sigma_0 \) expression;

\[
\ln |\Sigma| = \text{log-determinant, computed as the sum of logs of the non-zero eigenvalues of } \Sigma.
\]

Most basic references


http://hdl.handle.net/10256/297

Ongoing research several CoDaWork proceedings, available online at:

http://dugi-doc.udg.edu/handle/10256/150