Module 18 Multivariate Analysis for Genetic data
Session 05: Log-ratio principal component analysis

Jan Graffelman[1,2]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

26th Summer Institute in Statistical Genetics (SISG 2021)

Contents

What is compositional data?

- Compositional data consists of variables that are parts of some whole.
- Typical examples are proportions, percentages, concentrations.
- The data vectors are constrained and reside in a simplex.
- Compositions provide information about the relative values of the parts.
- Aitchison (1986) proposed the log-ratio approach to deal with compositional data.

Compositional data

Compositional data arise in many contexts:

- Mineral composition of rocks (geology)
- Time or budget expenditure (economy)
- Bacterial composition of the gut (microbiology)
- Allele frequencies and genotype frequencies (genetics)
- ...

Compositional data and spurious correlations

full composition

	A	B	C	D	E
s1	0.07	0.20	0.13	0.33	0.27
s2	0.00	0.20	0.13	0.07	0.60
s3	0.13	0.53	0.07	0.20	0.07

subcomposition

	A	B	C	D
s1	0.09	0.27	0.18	0.45
s2	0.00	0.50	0.33	0.17
s3	0.14	0.57	0.07	0.21

counts

	A	B	C	D	E
s1	10	30	20	50	40
s2	0	30	20	10	90
s3	20	80	10	30	10

R

	A	B	C	D	E
A	1.00	0.87	-0.87	0.50	-0.99
B	0.87	1.00	-1.00	0.00	-0.79
C	-0.87	-1.00	1.00	-0.00	0.79
D	0.50	0.00	-0.00	1.00	-0.62
E	-0.99	-0.79	0.79	-0.62	1.00

R

	A	B	C	D
A	1.00	0.07	-1.00	0.31
B	0.07	1.00	-0.14	-0.93
C	-1.00	-0.14	1.00	-0.24
D	0.31	-0.93	-0.24	1.00

- Correlations can be spurious due to the existence of a linear constraint
- Ordinary PCA of the data will display a spurious correlation structure

Principles of Compositional Data Analysis (CoDA)

Principles:

- Scale invariance
- Permutation invariance
- Subcompositional coherence

Typical CoDA approach:

- In order to satisfy these principles, we use a log-ratio transformation of the data.
- Analyse the data by applying the classical statistical methods to the log-ratio transformed data.

Some notation

A composition of D parts

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)$$

The sample space is the simplex

$$S^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D) | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}$$

The closure operation $\mathcal{C}$ to the constant $\kappa > 0$ (usually 1)

$$\mathcal{C}(x) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right)$$

Log-ratio transformations

- Additive log-ratio transformation (alr; ratios of two parts)

$$alr(\mathbf{x}) = \left[\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \cdots, \ln\left(\frac{x_{D-1}}{x_D}\right) \right],$$

- Centred log-ratio transformation (clr; ratios of one part against all)

$$clr(\mathbf{x}) = \left[\ln\left(\frac{x_1}{g_m(\mathbf{x})}\right), \ln\left(\frac{x_2}{g_m(\mathbf{x})}\right), \cdots, \ln\left(\frac{x_D}{g_m(\mathbf{x})}\right) \right],$$

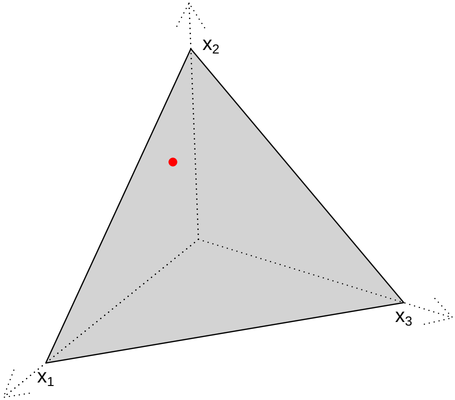
- Isometric log-ratio transformation (ilr; ratios of geometric means of subcompositions)

$$ilr(\mathbf{x}) = \left[\ln\left(\frac{g_m(\mathbf{x}_a)}{g_m(\mathbf{x}_b)}\right), \cdots, \ln\left(\frac{g_m(\mathbf{x}_c)}{g_m(\mathbf{x}_d)}\right) \right],$$

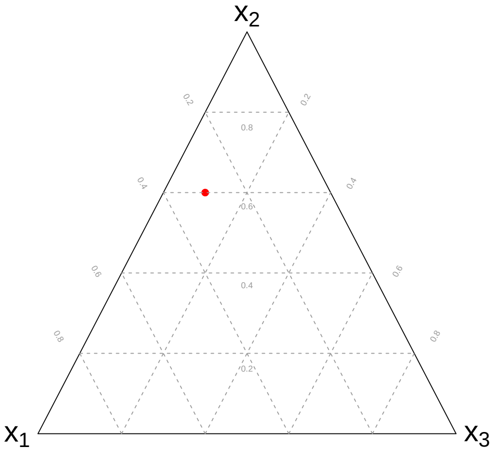
- $g_m(\mathbf{x})$ is the geometric mean of the components of the composition $\mathbf{x}$

$$g_m(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D} \quad \ln\left(g_m(\mathbf{x})\right) = \frac{1}{D}\sum_{i=1}^{D} \ln\left(x_i\right) \quad g_m(\mathbf{x}) = e^{\overline{y}} \quad y_i = \ln\left(x_i\right)$$

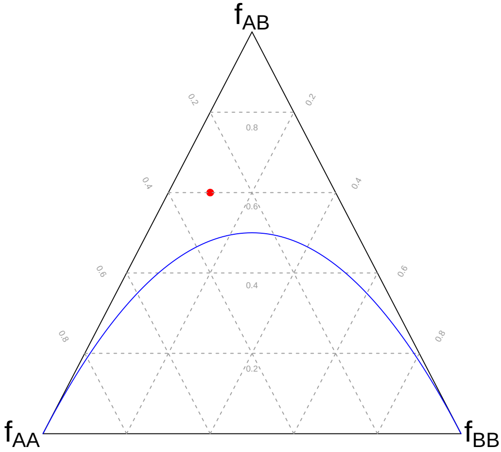
Visualizing 3 part compositions: ternary diagram



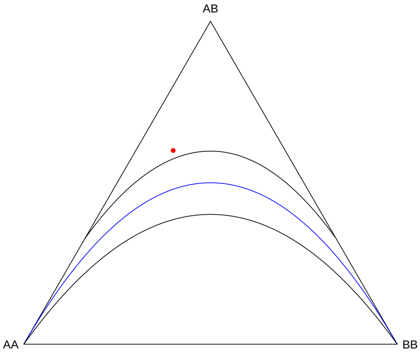
How to read a ternary diagram?



The ternary diagram in genetics: De Finetti diagram

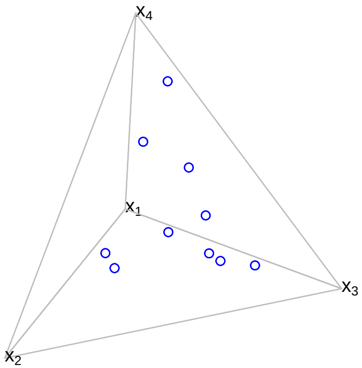


Is there Hardy-Weinberg equilibrium?



Graffelman, J. & Morales-Camarena, J. (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity* 65(2): 77-84

Plotting four-part compositions



Plotting compositional data

- Three part compositions can be visualised in a ternary diagram
- Four part compositions can be visualised in a tetrahedron
- Larger compositions can be visualised, in an approximate manner, in a compositional biplot
- Even three- and four part compositions are often better shown in a compositional biplot, as this represents them in an unconstrained space.
- Compositional biplots are obtained by log-ratio principal component analysis (LR-PCA).

Log-ratio PCA based on the centered log-ratio transformation

$$clr(x_i) = \ln\left(\frac{x_i}{g_m(\mathbf{x})}\right) = \ln\left(\frac{x_i}{(\prod x_i)^{1/D}}\right) = \ln(x_i) - \frac{1}{D}\sum_{i=1}^{D} \ln(x_i)$$

In matrix form:

$$\mathbf{X}_l = \ln(\mathbf{X})$$

$$\mathbf{X}_{\mathrm{clr}} = \mathbf{X}_l \mathbf{H}_r$$

$$\mathbf{X}_{c\mathrm{clr}} = \mathbf{H}_c \mathbf{X}_{\mathrm{clr}} = \mathbf{H}_c \mathbf{X}_l \mathbf{H}_r$$

where

$$\mathbf{H}_r = \mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}' \qquad \mathbf{H}_c = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}'.$$

The log-transformed data is double-centered

Do a standard PCA of the $\mathbf{X} = \mathbf{X}_{c\mathrm{clr}}$

Singular value decomposition (SVD)

Log-ratio PCA can be performed by the SVD:

$$\mathbf{X}_{cclr} = \mathbf{UDV}' \text{ with } \mathbf{U}'\mathbf{U} = \mathbf{I} \text{ and } \mathbf{V}'\mathbf{V} = \mathbf{I}.$$

Possible biplot coordinates (row markers **F** and column markers **G**)

- $\mathbf{F} = \mathbf{UD}$ and $\mathbf{G} = \mathbf{V}$ (the form biplot)
- $\mathbf{F} = \mathbf{U}$ and $\mathbf{G} = \mathbf{VD}$ (the covariance biplot)
- $\mathbf{F} = \mathbf{UD}^{1/2}$ and $\mathbf{G} = \mathbf{VD}^{1/2}$ (the symmetric biplot)

- The form biplot will approximate the Aitchison distances between the compositions.
- The Aitchison distance is the Euclidean distance between the clr transformed compositions.

The zero problem

The zero issue:

- Compositional data analysis generally considers the simplex to be open
- Zeros are not admitted

Important questions:

- How many zeros do you have?
- What kind of zeros do you have?
  - Rounding zeros (below detection limit)
  - Count zeros (related to sampling effort)
  - Essential or structural zeros (impossible outcome)

Solutions:

- For rounding or count zeros, impute a reasonable non-zero amount, for structural zeros, stratify.

Compositional biplot interpretation

- The origin represents the geometric mean of the compositions.
- Biplot vectors represent clr transformed parts.
- Links between vectors represent pairwise log-ratios:

$$clr(x_1) - clr(x_2) = \ln\left(\frac{x_1}{g_m(\mathbf{x})}\right) - \ln\left(\frac{x_2}{g_m(\mathbf{x})}\right) = \ln\left(\frac{x_1}{x_2}\right)$$

- The link length represents the standard deviation of the corresponding log-ratio.

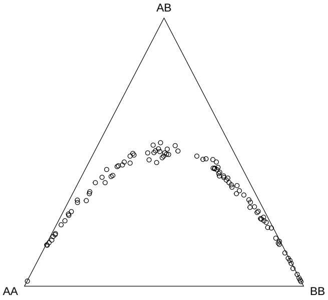
$$\begin{aligned}\| \mathbf{f}_i - \mathbf{f}_j \|^2 &= \mathbf{f}_i{}'\mathbf{f}_i + \mathbf{f}_j{}'\mathbf{f}_j - 2\mathbf{f}_i{}'\mathbf{f}_j \\ &= \mathrm{Var}\left(\mathrm{clr}(x_i)\right) + \mathrm{Var}\left(\mathrm{clr}(x_j)\right) - 2Cov\left(\mathrm{clr}(x_i), \mathrm{clr}(x_j)\right) \\ &= \mathrm{Var}\left(\ln\left(\frac{x_i}{\mathrm{g_m}(\mathbf{x})}\right) - \ln\left(\frac{x_j}{\mathrm{g_m}(\mathbf{x})}\right)\right) = \mathrm{Var}\left(\ln\left(\frac{x_i}{x_j}\right)\right).\end{aligned}$$

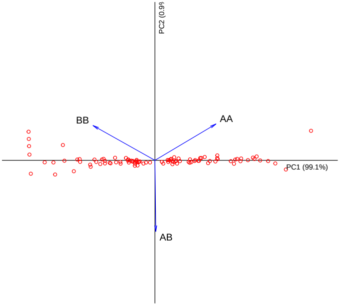
- Close to coincident biplot vectors suggest proportionality of parts.
- Cosines of angles between links represent correlations between log-ratios
- Collinear biplot vectors suggest a one-dimensional pattern for a subcomposition.

Experiment

- Take repeated samples from bi-allelic genetic variants that are in Hardy-Weinberg equilibrium
- Record the genotype composition
- Do a log-ratio PCA of the compositions obtained

Results






```
     PC1     PC2      PC3                 [,1]   [,2]  [,3]
la   4.837 0.04373 7.28e-32          AA  0.7019  0.417 0.577
laf  0.991 0.00896 1.49e-32          BB -0.7121  0.399 0.577
lac  0.991 1.00000 1.00e+00          AB  0.0102 -0.816 0.577
```

- PC1: log odds of the allele frequency
- PC2: deviation from Hardy-Weinberg equilibrium

$$\ln(f_{AA}) + \ln(f_{BB}) - 2\ln(f_{AB}) = \ln\left(\frac{f_{AA}f_{BB}}{f_{AB}^2}\right) = \ln\left(\frac{p^2q^2}{(2pq)^2}\right) = c$$

Worldwide Y-STR dataset

- Purps, J. et al. (2014) A global analysis of Y–chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics* 12: 12–23.
- Data consists of 23 Y-STRs typed for 19,630 males in 129 populations sampled world-wide,
- Samples stemming from Africa, Asia, Europe, Latin and North America.

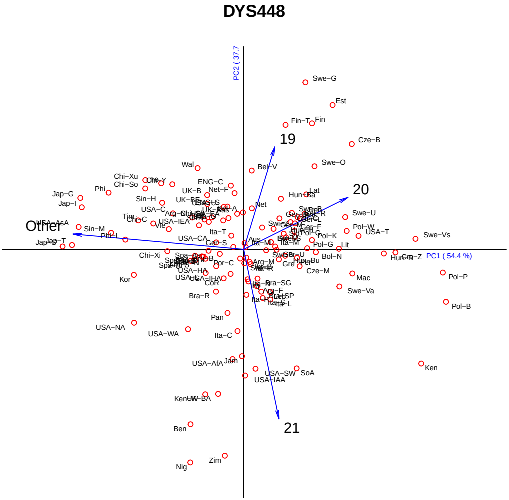
Example: Allele frequencies of Y-STR DYS448 over 129 populations worldwide

Sample	19	20	21	other
Arg-B	37	31	10	14
Arg-F	30	22	13	6
Arg-M	45	34	12	10
Arg-N	23	14	3	10
Arg-S	25	10	6	9
Aus	108	100	24	27
Bel-A	115	61	12	18
Bel-V	63	32	3	7
Ben	4	7	32	8
Bol-M	19	18	4	3
Bol-N	16	33	5	2
Bos	44	46	6	4
Bra-R	53	28	24	18
Bra-SG	35	12	10	4
Bra-SP	54	33	24	9
Chi-B	91	80	27	48
Chi-C	44	19	6	31
Chi-Sh	35	49	6	19
Chi-So	17	4	1	8
Chi-Xi	6	53	4	29
Chi-Xu	53	37	4	51
Chi-Y	31	39	3	28
CoR	75	42	27	22
Cro-C	54	57	8	6
Cro-Z	40	63	10	1
Cze-B	25	45	1	1
Cze-M	13	22	5	2
Den	76	92	10	7
ENG-C	46	24	3	8
ENG-S	66	30	6	12
⋮	⋮	⋮	⋮	⋮
Wal	85	17	3	13
Zim	3	6	42	4

Sample	19	20	21	other
Arg-B	0.40	0.34	0.11	0.15
Arg-F	0.42	0.31	0.18	0.08
Arg-M	0.45	0.34	0.12	0.10
Arg-N	0.46	0.28	0.06	0.20
Arg-S	0.50	0.20	0.12	0.18
Aus	0.42	0.39	0.09	0.10
Bel-A	0.56	0.30	0.06	0.09
Bel-V	0.60	0.30	0.03	0.07
Ben	0.08	0.14	0.63	0.16
Bol-M	0.43	0.41	0.09	0.07
Bol-N	0.29	0.59	0.09	0.04
Bos	0.44	0.46	0.06	0.04
Bra-R	0.43	0.23	0.20	0.15
Bra-SG	0.57	0.20	0.16	0.07
Bra-SP	0.45	0.28	0.20	0.07
Chi-B	0.37	0.33	0.11	0.20
Chi-C	0.44	0.19	0.06	0.31
Chi-Sh	0.32	0.45	0.06	0.17
Chi-So	0.57	0.13	0.03	0.27
Chi-Xi	0.07	0.58	0.04	0.32
Chi-Xu	0.37	0.26	0.03	0.35
Chi-Y	0.31	0.39	0.03	0.28
CoR	0.45	0.25	0.16	0.13
Cro-C	0.43	0.46	0.06	0.05
Cro-Z	0.35	0.55	0.09	0.01
Cze-B	0.35	0.62	0.01	0.01
Cze-M	0.31	0.52	0.12	0.05
Den	0.41	0.50	0.05	0.04
ENG-C	0.57	0.30	0.04	0.10
ENG-S	0.58	0.26	0.05	0.11
⋮	⋮	⋮	⋮	⋮
Wal	0.72	0.14	0.03	0.11
Zim	0.05	0.11	0.76	0.07

Purps, J. et al. (2014) A global analysis of Y–chromosomal haplotype diversity for 23 STR loci. Forensic Science International: Genetics 12: 12–23.

LR-PCA biplot of allele frequencies



References:

- Aitchison, J. (1986) *The statistical analysis of compositional data*. Chapman & Hall.
- Aitchison, J. (1983) Principal component analysis of compositional data. *Biometrika* 70(1) pp. 57-65.
- Pawlowsky-Glahn, V., Egozcue, J.J. & Tolosana-Delgado, R. (2015) *Modeling and Analysis of Compositional Data*, Chichester, United Kingdom, John Wiley & Sons.