Why are microbiome data compositional?

Vera Pawlowsky-Glahn

Emeritus Prof., Dep. Computer Science, Applied Mathematics & Statistics, University of Girona, Spain
Past-President of the Association for Compositional Data

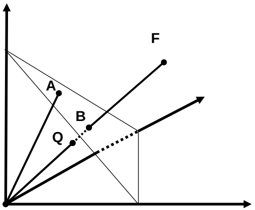
joint work with **Juan José Egozcue**
Emeritus Prof., Dep. Civil & Environmental Engineering, Technical University of Catalonia, Barcelona, Spain
President of the Association for Compositional Data

The Barcelona Debates on the Human Microbiome
Barcelona, Spain, 20-21 June 2019

What are compositional data (CoDa)?

- **historically:** sum constraint data, like proportions or percentages
- **after 1980:** strictly positive data that carry relative information
- **after 2001:** **parts of some whole that carry relative information, equivalence classes** of strictly positive, proportional vectors

representative: $$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D \;\middle|\; x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}$$



- $\mathcal{S}^D \subset \mathbb{R}^D_+ \subset \mathbb{R}^D$; $\kappa =$ constant, frequently 1 or 100
- CoDa need not to be closed
- scale invariant properties hold for any subcomposition*
- analyses can be based on any representative

* **subcomposition:** equivalence class of a subset of parts

Microbiome data: usually tables of counts or proportions

part of a table of oral microbiome data*

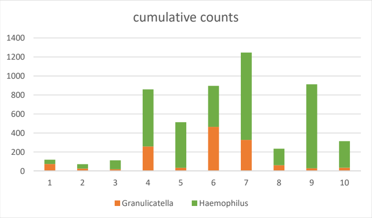
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	...
Fusobacterium	13	7	25	10	10	10	70	1575	221	73	...
Gardnerella	0	0	0	0	0	0	0	0	0	0	...
Gemella	12	6	0	70	10	54	95	79	39	12	...
Geobacillus	0	0	0	0	0	0	0	0	0	0	...
Gillisia	0	0	0	0	0	0	0	0	0	0	...
Granulicatella	74	26	19	258	34	465	328	61	29	35	...
Haemophilus	45	46	94	601	480	431	918	174	883	279	...
Haloanella	0	0	0	0	0	0	0	0	0	0	...
Helicobacter	0	0	0	0	0	0	0	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

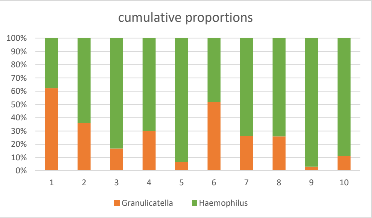
*Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature, 486.

table of (relative) abundances of features (OTUs, bacteria, phyla, genera, ...)

- how many times a sequence aligns to a reference annotation, classification of genomic sequences
- large proportion of zeros, positive numbers representing portions of a whole

Information in barplots of Granulicatella and Haemophilus





Do both representations carry the same information?

- **NOT** in absolute scale, **YES** in relative scale
- counts can not be estimated from proportions
- but proportions can be estimated from counts

Important characteristics of microbiome data

microbiome data are compositional!!!

- **the total number of sequenced reads** depends on the capacity of the instrument and **is not informative**
- absolute and relative abundances carry the same relative information
- information in microbiome data is relative
- data are strictly positive or zero, never negative
- zeros may be due to undersampling, high heterogeneity, or real absence

note

- absolute abundances are not recoverable from sequence data alone
- each count is not compositional itself, but the share out of counts is

Why is the compositional nature of data a problem?

typical problems

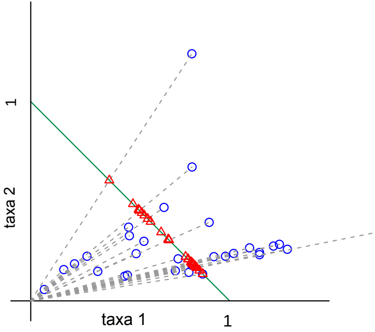
- discrimination and clustering are affected by sequencing depth
- correlation between two taxa depends on the subcomposition considered: it is spurious (Pearson, 1897); some are necessarily negative (negative bias)
- many methods are subcompositionally incoherent

actual practice does not avoid the problems

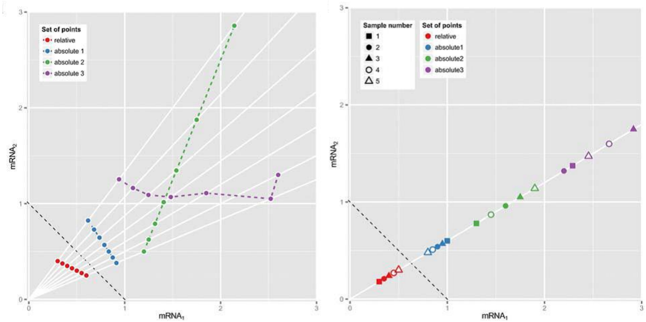
- rarefaction and count normalization do not change the compositional nature of data, but might introduce noise
- some dissimilarities (UniFrac; Bray-Curtis; Jensen-Shannon divergence) used for clustering and discrimination are not subcompositionally coherent

Problems with compositional data

changes in proportions do not reflect changes in absolute abundance



Egozcue and Pawlowsky-Glahn (2018)



Lovell et al. (2015)

Which is the origin of these problems?

experiments produce results (data); **data** can be categorical, numerical, functional, sets, ...; results are observed and recorded in a **sample space**;

examples: real space, positive orthant of real space, simplex, hypersphere, ...

desirable (ideal) properties of the sample space

- includes **only possible results** and has a **structure**
- a **scale** is defined (how are differences measured?)
- **operations** are defined (sum, product, shift, ...)
- a **metric** is available (angle, orthogonality, distance, ...)

an inappropriate sample space can produce spurious results!!!

Problems with compositional data

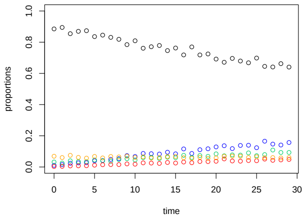
most methods assume the sample space to be $\mathcal{S}^D \subset \mathbb{R}^D$ with the usual Euclidean geometry; this can lead to nonsensical results

examples with closed (constant sum) CoDa:

1. standard Euclidean distances are not dominant
2. correlations are spurious
3. the standard covariance matrix is singular
4. covariance matrices are spurious $\Rightarrow$ all methods based on covariance or correlation are flawed
5. Bray-Curtis dissimilarity and Unifrac (weighted and unweighted) distances are not subcompositionally coherent

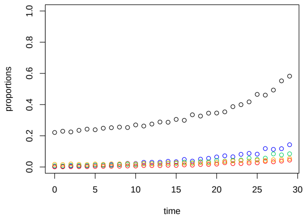
spurious correlation (simulated data)

proportions in $\mathcal{S}^5$



	x1	x2	x3	x4	x5
x1	1.00	**-0.99**	-0.97	-0.98	0.15
x2	-0.99	1.00	0.95	0.98	-0.22
x3	-0.97	0.95	1.00	0.92	-0.21
x4	-0.98	0.98	0.92	1.00	-0.18
x5	0.15	-0.22	**-0.21**	-0.18	1.00

proportions in $\mathcal{S}^6$



	x1	x2	x3	x4	x5
x1	1.00	**0.98**	0.97	0.98	0.98
x2	0.98	1.00	0.98	0.99	0.97
x3	0.97	0.98	1.00	0.97	0.96
x4	0.98	0.99	0.97	1.00	0.97
x5	0.98	0.97	**0.96**	0.97	1.00

Principles underlying CoDa analysis

1. scale invariance

- scaling factors do not alter the analysis
- avoids the need for rarefaction
- ratios of components are relevant!

2. subcompositional coherence (compatibility)

- subcompositional scale invariance
- subcompositional dominance ($d_a(x_1, x_2) \geq d_a(s_1, s_2)$), distances will never decrease if additional taxa are observed)
- ratios of common parts are preserved

Aitchison geometry

$\mathcal{S}^D(\oplus, \odot, \langle,\rangle_a)$ is a $(D-1)$-dimensional Euclidean space

For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathbb{R}$, $\mathcal{C}$ 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**: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$
- **norm**, **distance**: $\|\mathbf{x}\|_a^2 = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2$, $\quad d_a^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2$

Aitchison (1982, 1986), operations and distance;
Pawlowsky-Glahn and Egozcue (2001), Aitchison geometry

Advantages of the Aitchison geometry

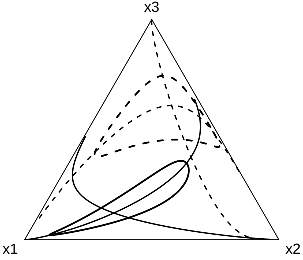
- olr-**coordinates** (orthonormal, isometric log-ratio coordinates, previously known as ilr) are available, e.g. balances
- operations and metrics in $\mathcal{S}^D$ are equivalent to ordinary operations and metrics in coordinates **(principle of working in coordinates)**
- **Aitchison measure** in $\mathcal{S}^D$ = Lebesgue measure in olr-coordinates in $\mathbb{R}^{D-1}$
- standard statistical tools can be used on olr-coordinates

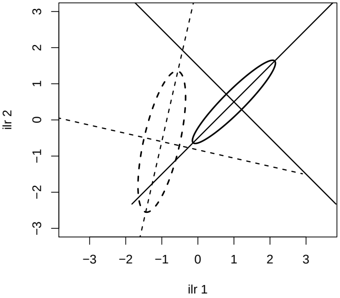
Special features of the Aitchison geometry

- correlation between parts is not valid
  $\Rightarrow$ **alternatives are based on proportionality**
- questions need reformulation
  $\Rightarrow$ **always two or more parts are involved**
- questions and statements on **single parts are nonsensical**

The Aitchison geometry: ellipses and lines

what you see in proportions **... and in olr-coordinates**



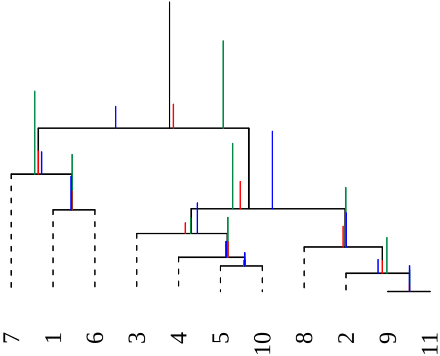


$$\mathrm{olr}_1(\mathbf{x}) = \sqrt{\frac{2}{3}} \log \frac{x_1}{(x_2 x_3)^{\frac{1}{2}}}$$

$$\mathrm{olr}_2(\mathbf{x}) = \sqrt{\frac{1}{2}} \log \frac{x_2}{x_3}$$

CoDa-dendrogram: partition, means, variances, olr-coordinates

1-*Actinomyces*
2-*Fusobacterium*
3-*Gemella*
4-*Granulicatella*
5-*Haemophilus*
6-*Leptotrichia*
7-*Neisseria*
8-*Porphyromonas*
9-*Prevotella*
10-*Streptococcus*
11-*Veillonella*



keratinized gingiva
buccal mucosa
supragingival plaque

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

visual ANOVA for each balance

application of balances in microbiome studies: SELBAL
(selection of a balance to predict a condition or disease)

Concluding remarks

microbiome data are compositional!!!

- interest is (or should be) in the relative information carried by proportions
- the simplex corresponds to the set of possible observations
- an interpretable measure of difference and scale of variables is available
- a suitable, well known algebraic-geometric structure allows building coherent models
- for CoDa, it is better to think in terms of ratios

some references (I)

- **Aitchison J (1982)**: The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, B*, **44**(2), 139–177.
- **Aitchison J (1983)**: Principal component analysis of compositional data. *Biometrika*, **70**(1), 57–65.
- **Aitchison J (1986)**: *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman, London (UK).
- **Aitchison J; Shen SM (1980)**: Logistic-normal distributions. Some properties and uses. *Biometrika*, **67**(2), 261–272.
- **Barceló-Vidal C; Martín-Fernández JA (2016)**: The Mathematics of Compositional Analysis. *Austrian Journal of Statistics 45: 57–71*.
- **Egozcue JJ; Pawlowsky-Glahn V (2005)**: Groups of parts and their balances in compositional data analysis. *Math. Geol.*, **37**(7), 795–828.
- **Egozcue JJ; Pawlowsky-Glahn V (2006)**: *Simplicial geometry for compositional data*. In: Buccianti et al (Eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Soc., London (UK), SP 264.
- **Egozcue JJ; Pawlowsky-Glahn V (2018)**: *Modelling Compositional Data. The Sample Space Approach*. In: Daya Sagar B et al (Eds) *Handbook of Mathematical Geosciences*. Springer, Cham.
- **Egozcue JJ; Pawlowsky-Glahn V (2019)**: Compositional data: the sample space and its structure. *TEST* (in press).

some references (II)

- **Egozcue JJ et al (2018)**: Linear Association in Compositional Data Analysis. *Austrian Journal of Statistics*, **47**(1).
- **Egozcue JJ et al (2003)**: Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**(3).
- **Gloor GB et al (2017)**: Microbiome datasets are compositional: and this is not optional. *Frontiers Microbiology*, Mini Review article.
- **Lovell D et al (2015)**: Proportionality: A Valid Alternative to Correlation for Relative Data, *PLoS Computational Biology*, **11**(3).
- **Martín-Fernández JA et al (2011)**: Dealing with zeros. In: Pawlowsky-Glahn and Buccianti (Eds) *Compositional Data Analysis: Theory and Applications.* Wiley (UK).
- **Pawlowsky-Glahn V; Egozcue JJ (2001)**: Geometric approach to statistical analysis on the simplex. *SERRA*, **15**(5).
- **Pawlowsky-Glahn V et al (2015)**: *Modeling and Analysis of Compositional Data*, Wiley, Chichester (UK).
- **Rivera-Pinto J et al (2018)**: Balances: a new perspective for microbiome analysis. *mSystems* 3:e00053-18.
- **Tsilimigras MC; Fodor AA (2016)**: Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*, **26**(5).