Why are microbiome data compositional?

Vera Pawlowsky-Glahn[1] and **Juan José Egozcue**[2]

[1]Emeritus Prof., Dep. Computer Science, Applied Mathematics & Statistics, University of Girona, Spain
President of the Association for Compositional Data 2015-2017

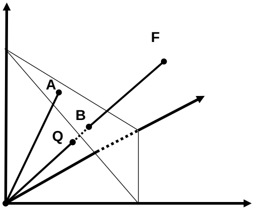
[2]Emeritus Prof., Dep. Civil & Environmental Engineering, Technical University of Catalonia, Barcelona, Spain
President of the Association for Compositional Data 2017-2021

NORBIS Summer School 2021, Norway
20 August 2021

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

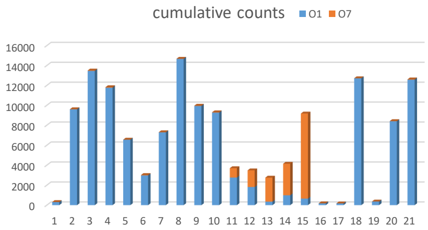
Microbiome data: usually tables of counts or proportions

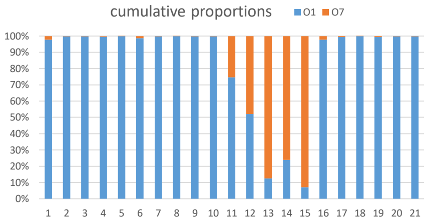
data for the hands-on session

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	sample	Watergr	01	07	02238	03	04	06	02	011	05	08	04580	09	014	020	028	024	016	015	026	025
2	S1	2	305	7	3	6715	4725	7127	3	8	14	34	2191	335	1068	9	0	42	395	0	0	2
3	S10	1	9594	14	2900	902	1350	83	2938	45	836	242	59	852	733	22	134	41	111	567	39	23
4	S11	1	13457	39	4630	511	1770	89	2682	38	637	554	99	708	493	87	106	168	315	1027	92	63
5	S12	1	11771	49	6743	759	1379	129	6129	30	1067	608	61	1571	1195	100	263	293	614	582	106	62
6	S13	1	6551	7	2830	658	803	53	4143	7	765	87	31	1192	685	6	181	17	71	368	20	16
7	S14	3	2977	40	5553	59	257	21	2062	90	139	678	16	1264	237	61	631	21	1146	224	244	201
8	S15	1	7287	12	1527	931	1270	1351	1739	10	663	172	51	272	464	16	83	53	60	440	21	19
9	S16	1	14662	11	3877	1015	2024	189	3184	29	978	321	87	815	854	47	159	94	138	940	24	24
10	S17	1	9936	23	1847	502	1742	154	1214	19	349	362	80	186	330	52	33	206	145	682	24	18
11	S18	1	9272	24	3566	572	1469	109	3619	6	722	165	54	1053	530	34	151	109	155	606	54	47
12	S19	3	2765	935	4161	0	457	15	129	344	34	610	0	1	389	208	1	532	494	272	235	241
13	S20	3	1815	1669	2426	0	314	11	74	761	1	1110	11	20	163	241	11	257	281	232	314	268
14	S21	3	346	2411	423	2	128	1	83	1072	0	837	4	0	33	322	1	115	168	55	166	215
15	S22	3	989	3157	1059	10	246	4	259	1479	30	1647	3	1	137	449	0	292	289	237	413	373
16	S23	3	652	8523	846	247	356	176	83	2861	14	1592	77	20	78	808	2	132	136	97	271	260
17	S24	2	180	4	4	3900	4273	3360	9	1	41	30	1506	103	591	4	0	34	43	1	1	3
18	S25	2	173	1	2	3316	2755	2371	6	5	51	35	1099	66	507	2	0	37	29	0	1	3
19	S3	1	12714	4	806	2512	3739	110	3473	15	5496	98	102	175	1038	7	96	26	20	1198	21	8
20	S4	2	362	2	7	7298	4618	3865	26	2	178	11	1707	95	1042	2	42	12	22	0	1	3
21	S6	1	8400	13	2594	1607	1378	160	4231	23	1355	289	80	1042	955	36	115	64	68	470	37	33
22	S7	1	12574	29	3951	1202	2128	225	3977	69	1009	422	109	1012	963	82	217	253	312	793	40	35

DataSet_WaterMasses_Os_cols

Information in barplots of O1 and O7





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

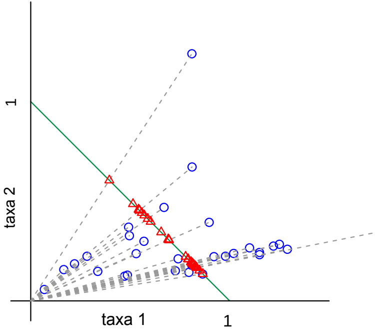
- discrimination and clustering are affected by sequencing depth
- correlation between two taxa depends on the subcomposition considered: it is spurious (Pearson, 1897); some are necessarily negative (negative bias)
- many methods are subcompositionally incoherent

actual practice does not avoid the problems

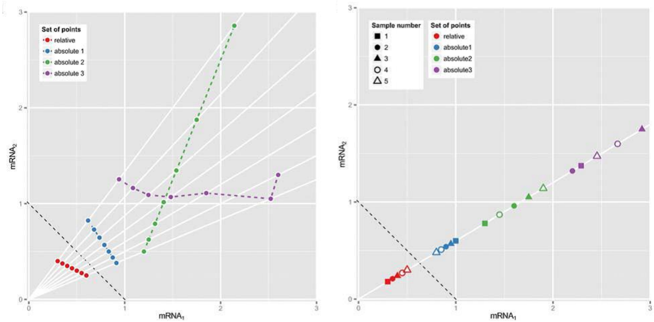
- rarefaction and count normalization do not change the compositional nature of data, but might introduce noise
- some dissimilarities (UniFrac; Bray-Curtis; Jensen-Shannon divergence) used for clustering and discrimination are not subcompositionally coherent

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

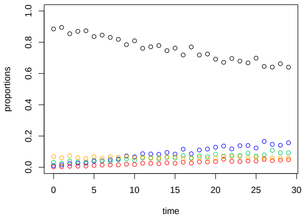
most methods assume the sample space to be $\mathcal{S}^D \subset \mathbb{R}^D$ with the usual Euclidean geometry; this can lead to nonsensical results

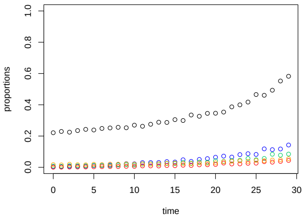
examples with closed (constant sum) CoDa:

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)

closed five simulated parts	adding a large sixth component
proportions in $\mathcal{S}^5$	**proportions in $\mathcal{S}^6$**





correlations between the five parts

	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

	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

spurious correlation (data for hands-on session)

Spurious correlations always appear, not only in simulated data

correlations between 5 OTUs for two closed subcompositions

20 OTU data after substitution of zeros and closure

	O20	O24	O25	O26	O28
O20	1.00	0.37	0.79	0.75	-0.27
O24	0.37	1.00	0.55	0.55	-0.18
O25	0.79	0.55	1.00	0.99	-0.05
O26	0.75	0.55	0.99	1.00	0.03
O28	-0.27	-0.18	-0.05	0.03	1.00

5 OTU closed subcomposition

	O20	O24	O25	O26	O28
O20	1.00	-0.22	0.64	0.51	-0.66
O24	-0.22	1.00	-0.43	-0.57	-0.48
O25	0.64	-0.43	1.00	0.89	-0.50
O26	0.51	-0.57	0.89	1.00	-0.32
O28	-0.66	-0.48	-0.50	-0.32	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**: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]$; $\mathbf{x} \ominus \mathbf{y} = \mathcal{C}[x_1/y_1, \ldots, x_D/y_D]$
- **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**

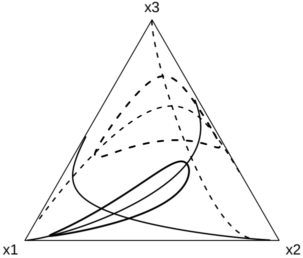
classes of zeros and how to deal with them

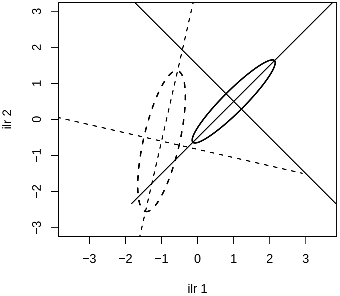
1. the part with zeros is **not important** for the study
   $\Rightarrow$ the part should be omitted; or treat it as essential zeros
2. the part is important, the **zeros are essential**
   $\Rightarrow$ divide the sample into two or more populations, according to the presence/absence of zeros
3. the part is important, the zeros are **rounded zeros**
   $\Rightarrow$ use imputation techniques
4. **zero counts:** the data are counts that can be zero, but the corresponding proportion is not zero
   $\Rightarrow$ Bayesian imputation techniques

Zeros are not parts of a composition, they are not relative to anything; they are either sampling defects or essential

The Aitchison geometry: ellipses and lines

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





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

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

variation array — looking for proportionality of parts

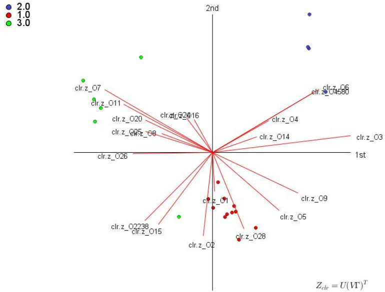
Variance ln(Xi/Xj)

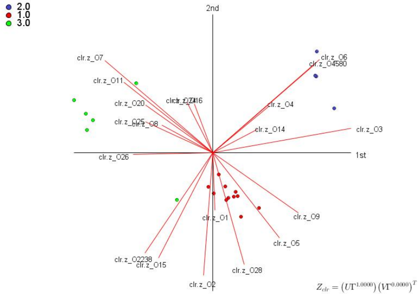
Xi\Xj	z_01	z_07	z_02238	z_03	z_04	z_06	z_02	z_011	z_05	z_08	z_04580	z_09	z_014	z_020	z_028	z_024	z_016	z_015	z_026	z_025	clr variances
z_01		9.6170	2.5312	9.6423	3.3302	8.9259	1.1922	7.1020	2.5228	2.8192	8.4124	4.1506	2.2042	4.3761	2.0947	2.7351	2.7264	1.8876	3.7571	3.9564	0.7838
z_07	-4.2224		7.8568	25.6900	11.8771	18.9942	12.8331	0.7889	19.1153	2.4660	17.9836	20.5152	10.9980	1.2795	16.4467	3.9044	4.4382	9.0671	2.2409	1.7388	6.4771
z_02238	-1.3935	2.8289		20.8374	10.6725	19.4128	1.8174	6.2789	8.2520	3.4108	18.5920	10.5617	8.3689	4.2811	4.8227	5.3931	5.5314	0.5208	2.1649	3.5421	3.8270
z_03	-2.1715	2.0509	-0.7780		3.8603	2.3133	13.0616	21.7117	5.0852	14.9347	1.9637	4.1600	4.9121	17.5721	9.7149	12.2763	11.4433	18.7229	19.9041	17.8071	8.3652
z_04	-0.8631	3.3593	0.5304	1.3084		1.8090	7.4513	9.4110	3.9189	5.1423	1.6504	4.1710	0.3223	6.6961	6.1533	3.0375	2.8563	9.6102	8.3556	6.6800	1.9348
z_06	-3.1268	1.0956	-1.7333	-0.9553	-2.2637		14.6229	16.2590	7.3621	11.2049	0.5968	5.6426	2.8868	12.9343	11.1048	7.8978	7.3015	18.3286	16.0779	13.3078	6.4337
z_02	-1.8151	2.4073	-0.4216	0.3564	-0.9520	1.3117		10.0327	3.1278	5.1128	13.5255	5.6134	5.6367	7.0150	1.6971	6.2545	6.2059	1.2915	4.9821	6.0586	2.9612
z_011	-4.2538	-0.0313	-2.8602	-2.0823	-3.3907	-1.1270	-2.4386		15.3096	1.3050	15.0963	16.7695	8.4486	0.8057	12.9891	2.9054	3.2767	7.0243	1.4614	1.0791	4.4873
z_05	-2.9192	1.3033	-1.5257	-0.7477	-2.0561	0.2076	-1.1041	1.3346		8.7920	7.4154	2.9953	3.1068	11.4445	2.8810	7.7485	7.6507	6.6767	10.7144	10.6453	3.8228
z_08	-2.4270	1.7954	-1.0335	-0.2555	-1.5639	0.6998	-0.6119	1.8268	0.4922		10.2446	9.7584	4.2224	0.3286	7.2375	0.8572	1.0691	3.7336	0.6636	0.3379	1.2666
z_04580	-3.7376	0.4848	-2.3441	-1.5661	-2.8745	-0.6108	-1.9225	0.5162	-0.8184	-1.3106		5.0287	2.8953	12.0534	9.9886	7.6111	6.9510	17.3001	14.9904	12.3890	5.8190
z_09	-3.0010	1.2214	-1.6075	-0.8295	-2.1379	0.1258	-1.1859	1.2528	-0.0818	-0.5740	0.7366		3.5367	12.4328	2.5923	8.9460	7.4588	9.9128	12.5643	11.8477	4.5175
z_014	-1.8245	2.3980	-0.4309	0.3470	-0.9614	1.3023	-0.0093	2.4293	1.0947	0.6025	1.9131	1.1765		5.8349	4.6409	2.5727	2.1761	7.7275	6.8202	5.4929	1.2248
z_020	-4.2727	-0.0503	-2.8792	-2.1012	-3.4096	-1.1459	-2.4576	-0.0190	-1.3535	-1.8457	-0.5351	-1.2717	-2.4483		9.5813	1.1781	1.4605	4.9189	0.8363	0.4248	2.3573
z_028	-4.8391	-0.6167	-3.4456	-2.6676	-3.9760	-1.7123	-3.0240	-0.5853	-1.9199	-2.4121	-1.1015	-1.8381	-3.0146	-0.5664		7.8886	7.0293	4.7184	7.8767	8.2652	3.4707
z_024	-3.5156	0.7068	-2.1221	-1.3441	-2.6525	-0.3888	-1.7005	0.7382	-0.5964	-1.0886	0.2220	-0.5146	-1.6912	0.7571	1.3235		0.4585	5.3531	2.4240	1.5196	1.1327
z_016	-3.0632	1.1593	-1.6696	-0.8917	-2.2000	0.0637	-1.2480	1.1906	-0.1440	-0.6362	0.6745	-0.0621	-1.2387	1.2096	1.7759	0.4525		5.9039	2.7206	1.6927	1.0021
z_015	-3.1259	1.0965	-1.7324	-0.9544	-2.2628	0.0009	-1.3108	1.1279	-0.2067	-0.6989	0.6117	-0.1249	-1.3014	1.1468	1.7132	0.3897	-0.0627		2.9383	4.4272	3.5878
z_026	-4.4574	-0.2350	-3.0639	-2.2859	-3.5943	-1.3306	-2.6423	-0.2036	-1.5382	-2.0304	-0.7198	-1.4564	-2.6329	-0.1847	0.3817	-0.9418	-1.3943	-1.3315		0.3252	2.6755
z_025	-4.4192	-0.1968	-3.0257	-2.2477	-3.5561	-1.2924	-2.6041	-0.1654	-1.5000	-1.9922	-0.6816	-1.4182	-2.5947	-0.1465	0.4199	-0.9036	-1.3561	-1.2933	0.0382		2.1614

Mean ln(Xi/Xj)

68.3084 Total Variance

CoDa-covariance-biplot — CoDa-form-biplot

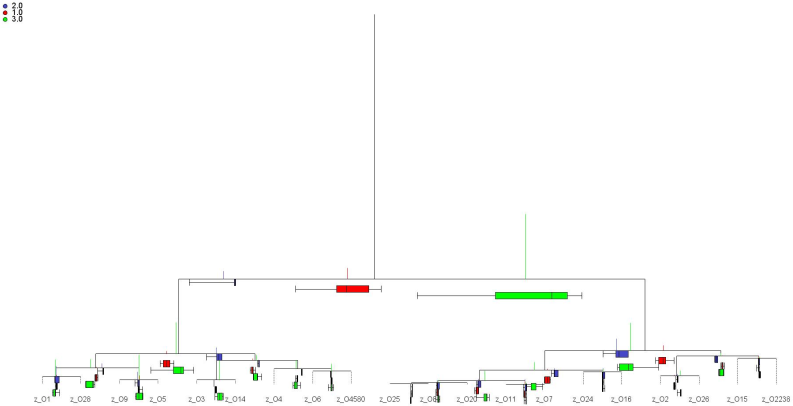




reflects relationships between parts — reflects distances between samples

proportion of explained variance: 0.9080

CoDa-dendrogram — visual ANOVA for each balance



$$b_i = \sqrt{\frac{r \cdot s}{r+s}} \ln \frac{\left(\prod_{i=1}^{r} x_i\right)^{1/r}}{\left(\prod_{j=1}^{s} x_j\right)^{1/s}}$$

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).
- **Mateu-Figueras G, Pawlowsky-Glahn V and Egozcue JJ (2011)**: The principle of working on coordinates. In Pawlowsky-Glahn, V. and Buccianti A. (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).