Compositionally appropriate linear association in high throughput sequencing data

Greg Gloor

May 17, 2017

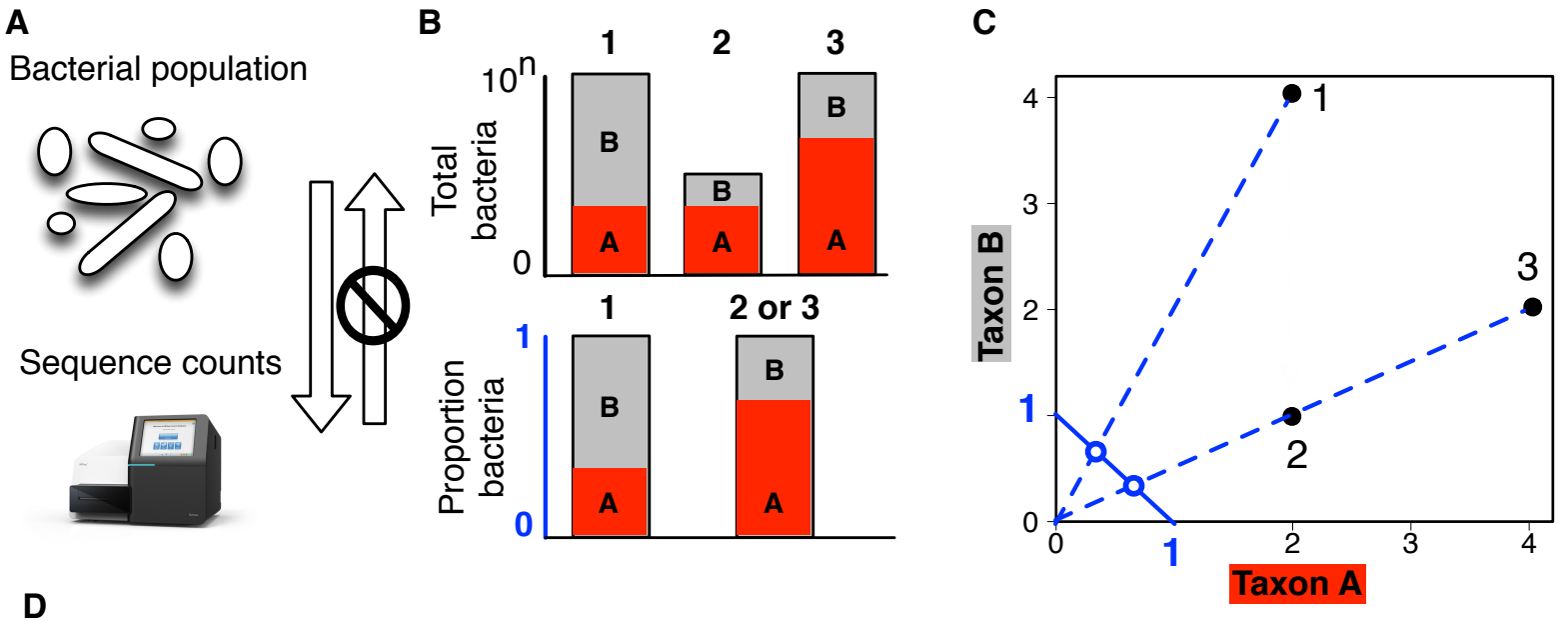
The problem of spurious correlation Pearson, 1897

Spurious correlation arises when data are constrained by a constant denominator, or equivalently, by a constant sum

"If $u = f_1(x, y)$ and $v = f_2(z, y)$ be two functions of the three variables x, y, z, and these variables be selected at random so there exists no correlation between x, z, y, z, or z, x, there will still be found to exist correlation between u and v. Thus a real danger arises when a statistical biologist attributes the correlation between two functions like u and v to organic relationship . . . ." Pearson 1897

This problem exists whenever there is a constant denominator in a dataset: proportion, percentage, ppm, normalized counts, reads/sequencing depth, etc. Sequencing data are constrained by this problem.

Relationship to HTS



D

	Change in population			Perceived change in proportion		
	1 → 2	1 → 3	2 → 3	1 → 2	1 → 3	2 → 3
Taxon A	—	↑	↑	↑	↑	—
Taxon B	↓	↓	↑	↓	↓	—

↑ Increase
↓ Decrease
— No change

Numbers vs. proportions

Numbers

	A	B	C	D	E
S1	10	20	20	50	50
S2	15	40	30	20	200
S3	20	80	10	30	15

Proportions (A-E)

	A	B	C	D	E
S1	.067	.133	.133	.333	.333
S2	.049	.131	.098	.066	.656
S3	.129	.516	.065	.194	.097

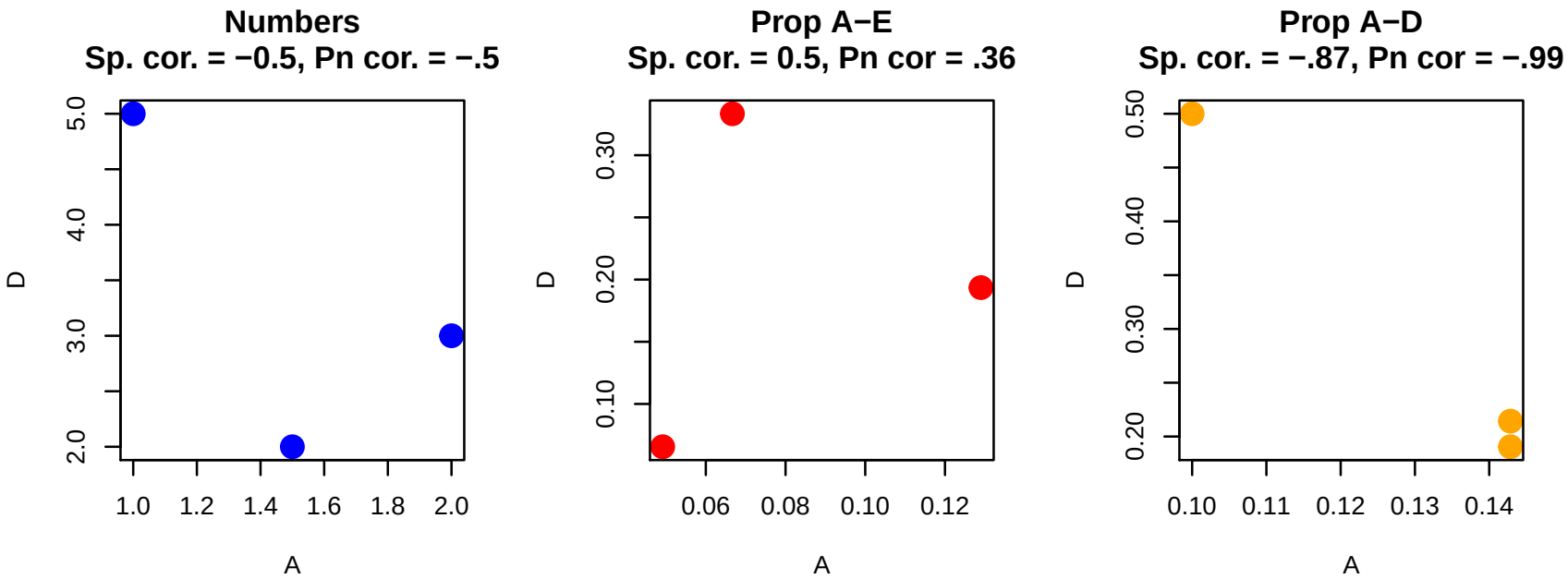
Proportions (A-D)

	A	B	C	D
S1	.100	.200	.200	.500
S2	.143	.381	.286	.190
S3	.143	.571	.071	.214

Spurious correlation in action Aitchison 1986

Fake Correlations! Sad!

Plots of A vs. D in each situation



What can you trust?

Correlation is not stable

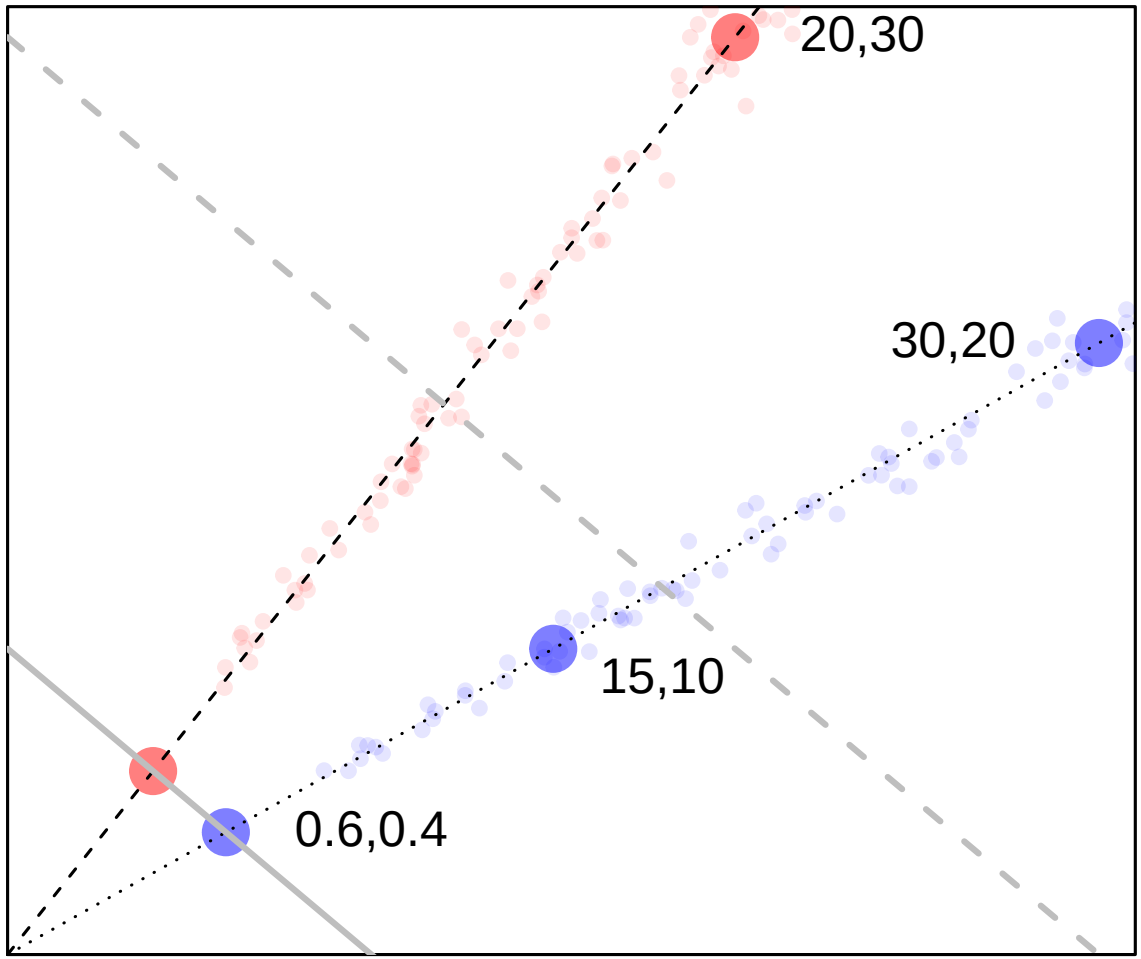
The correlation observed is not the same for the numerical and proportional data

The correlation changes again when the proportional data are subset

this is spurious correlation and is an unpredictable correlation observed between two variables whenever they share a common denominator, whether correlated or not with either or both of the two variables.

We usually think of correlations as linear relationships of the type $y = mx + b$

Unfair blue and unfair red coin Pawlowsky-Glahn, 2015

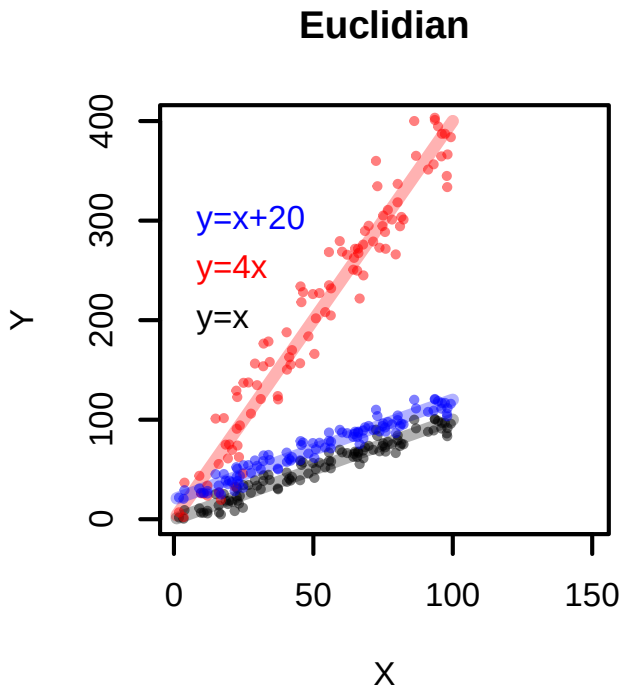


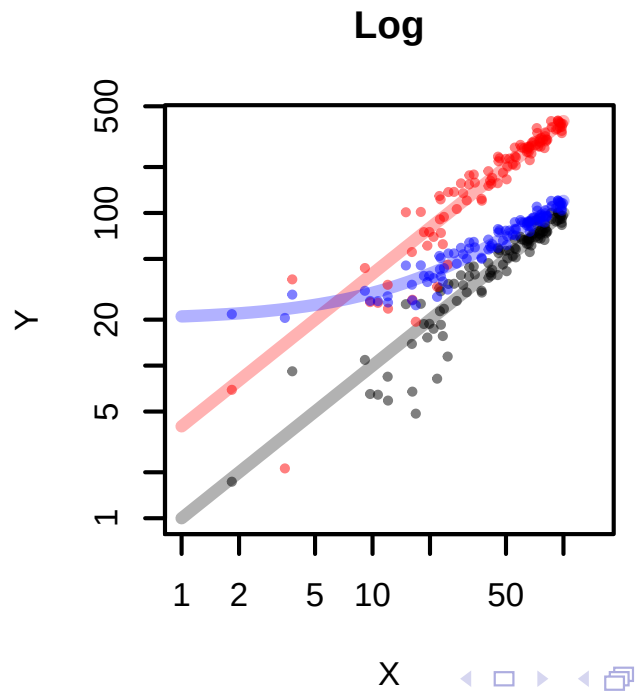
interpretation

1. note that any variables that are correlated must appear on a line projecting from the origin
2. linear relationships on this line are perfectly correlated: "compositionally associated", thus **they have a constant ratio**
3. We can represent the value on the simplex line as a ratio $\frac{H}{T} = \frac{0.6}{0.4} = 1.5$ 4: This can be made symmetrical by taking the logarithm: $log(\frac{0.6}{0.4}) =\sim 0.41$ (blue), since $log(\frac{20}{30}) =\sim -0.41$ (red)
4. Addition and subtractions of logarithms are the natural operations on the simplex.
5. Any simplex is the same as any other: **cannot normalize out of the simplex**

Familiar measures of correlation

- False positive correlations for both -ve and +ve correlation but different reasons
  - negative correlation bias
  - more ways of correlating with an intercept than without
- Slope in Euclidian = intercept in Log-Log
- Intercept in Euclidan = non-linearity of ratios
- Constant ratio relationships have a slope of 1 and are linear





Variance of ratios Aitchison, 1986

So if $y = mx$ is constant then $Var(log(\frac{X}{Y})) = 0$ }

This is not a correlation, it is a measure of lack of association. Zero means associated (constant ratio), any other value means a lack of association, but we need to scale the measure.

Problems:

1. the metric must be scaled
2. the slope must be 1 (in log space), or the intercept must be 0 in Euclidian space
3. we must have a linear line
4. we must account for scatter

The ϕ metric Lovell, 2015

Transform by the centered log-ratio: formally equivalent to calculating all pairs of ratios: makes differences between the parts linearly different Aitchison 1986

The data are still on the simplex, but not constrained to be in (1,0) to (0,1) for a ratio.

```
Z <- c(1,2,4,8,16,32,64)
cZ <- log2(Z) - mean(log2(Z))
```

$cZ = (-3, -2, -1, 0, 1, 2, 3)$

$clr(Z) = log\frac{z_i}{gz}; i = 1 \ldots D; gz = \text{geometric mean of } Z$

$\phi_{xy} = \frac{Var(clr(x) - clr(y))}{Var(clr(x))} = 0$ if the two variables are associated

geometrically: $\phi_{xy} = 1 + m^2 - 2m|r|$

The ρ metric Erb,2016

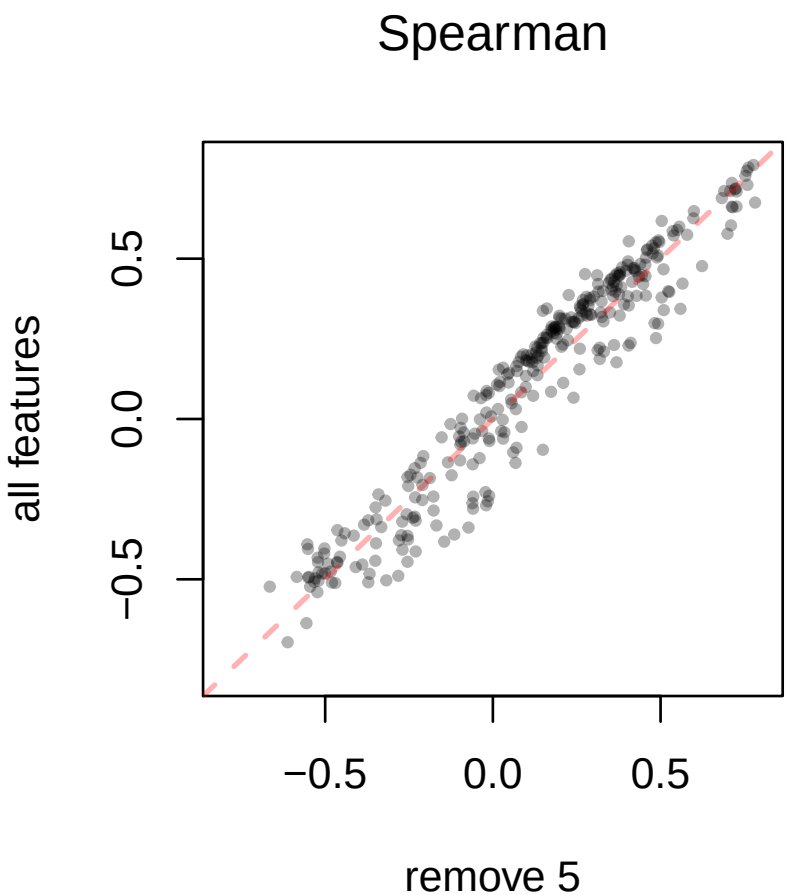
$$\rho_{xy} = \frac{2cov(clr(x),clr(y))}{Var(clr(x))+Var(clr(y))}$$

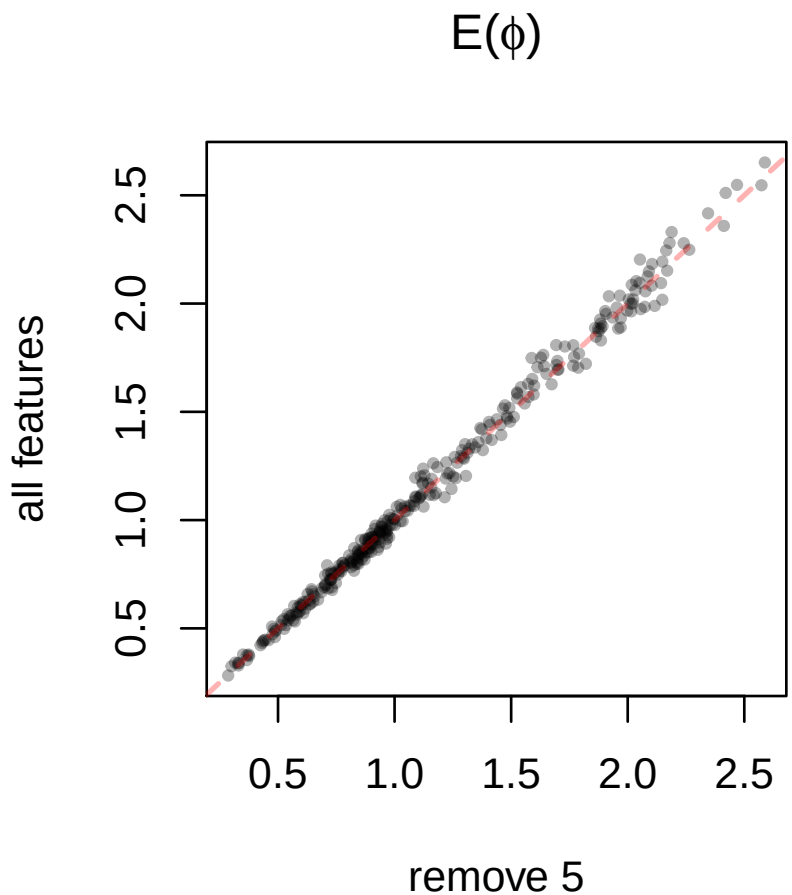
geometrically: $\rho_{xy} = \frac{2r}{m+1/m}$

no neat geomtric interpretation, but ranges from -1 to $+1$.

ϕ is more consistent than Spearman's correlation

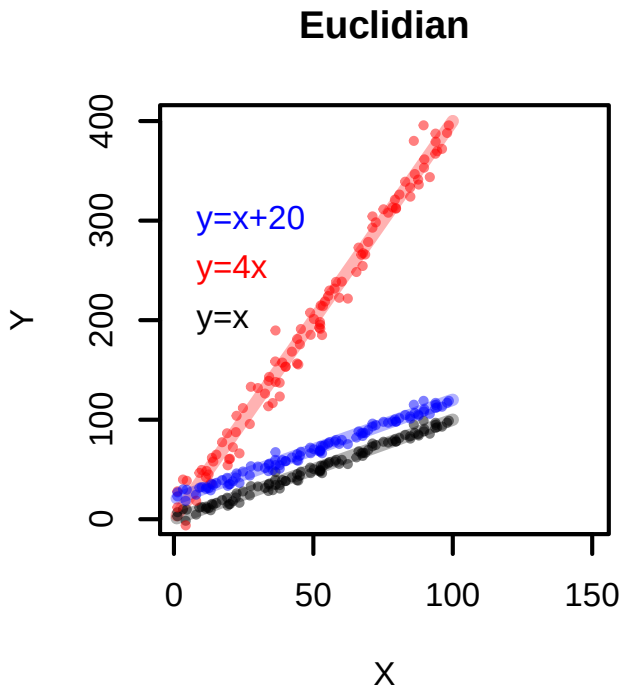
Consistency

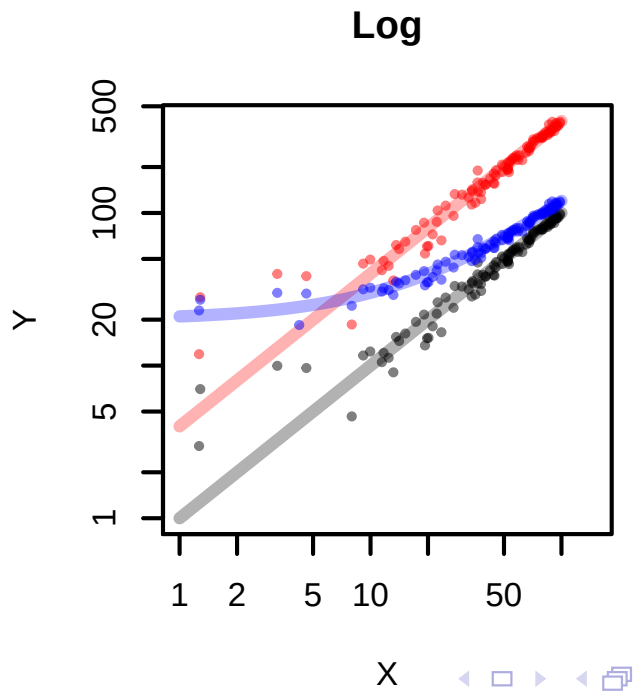




Summary

- compositional data are any data represented by a constant sum
- only ratio information is obtained
- negative correlations are uninterpretable
- any simplex is always equivalent to the unit simplex
- "in the absence of any other information or assumptions, correlation of relative abundances is just wrong" (Lovell)
- We need to calculate two numbers, slope and correlation (Egozcue et al. submitted)





Further Readings and Sources

- Pearson K. 1897. Mathematical contributions to the theory of evolution. – on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London: 60:489
- Aitchison J. 1986. The Statistical Analysis of Compositional Data. Chapman and Hall
- Lovell D. 2015. Proportionality: a valid alternative to correlation for relative data. PLoS Comp Bio. 11:e1004075
- Pawlowsky-Glahn V. 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons
- Erb I. 2016. How should we measure proportionality on relative gene expression data? Theory in Biosci. 135:21
- Egozcue JJ, Pawlowsky-Glahn V, Gloor G. submitted. Linear Association In Compositional Data Analysis