Correlation in compositional data

Greg Gloor

May 5, 2017

About this document

This document is an .Rmd document and is located at

github.com/ggloor/ in
compositions/background_reading/Correlation_composition.Rmd

The problem of spurious correlation

Spurious correlation arises when data are constrained by a constant denominator

"if the ratio of two absolute measurements on the same or different organs be taken it is convenient to term this ratio an index.

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 statisital 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, etc.

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

Outline

1) demonstration of the problem

2) an intuitive introduction to compositional data

3) deriving the list of pathologies

4) solution: examining ratios

5) phi and rho as partial solutions

Set up a small dataset

Assume first that we are dealing with numbers, three samples, five features. Calculate the correlations from the whole, a subset, the whole converted to proportions per sample, and a subset converted to proportions per sample.

```
s1 <- c(1,2,2,5,5)
s2 <- c(1.5,4,3,2,20)
s3 <- c(2,8,1,3,1.5)
s <- rbind(s1,s2,s3)
colnames(s) <- c("A", "B", "C", "D", "E")

cor.s <- cor(s, method="spearman")

s.p <- t(apply(s, 1, function(x) x/sum(x)))
cor.s.p <- cor(s.p, method="spearman")

s.p2 <- t(apply(s[,1:4], 1, function(x) x/sum(x)))
cor.s.p2 <- cor(s.p, method="spearman")
```

Numbers vs. proportions

What do you notice about the effect of converting each sample to a proportion?

```
print(s)
```

```
##      A B C D    E
## s1 1.0 2 2 5  5.0
## s2 1.5 4 3 2 20.0
## s3 2.0 8 1 3  1.5
```

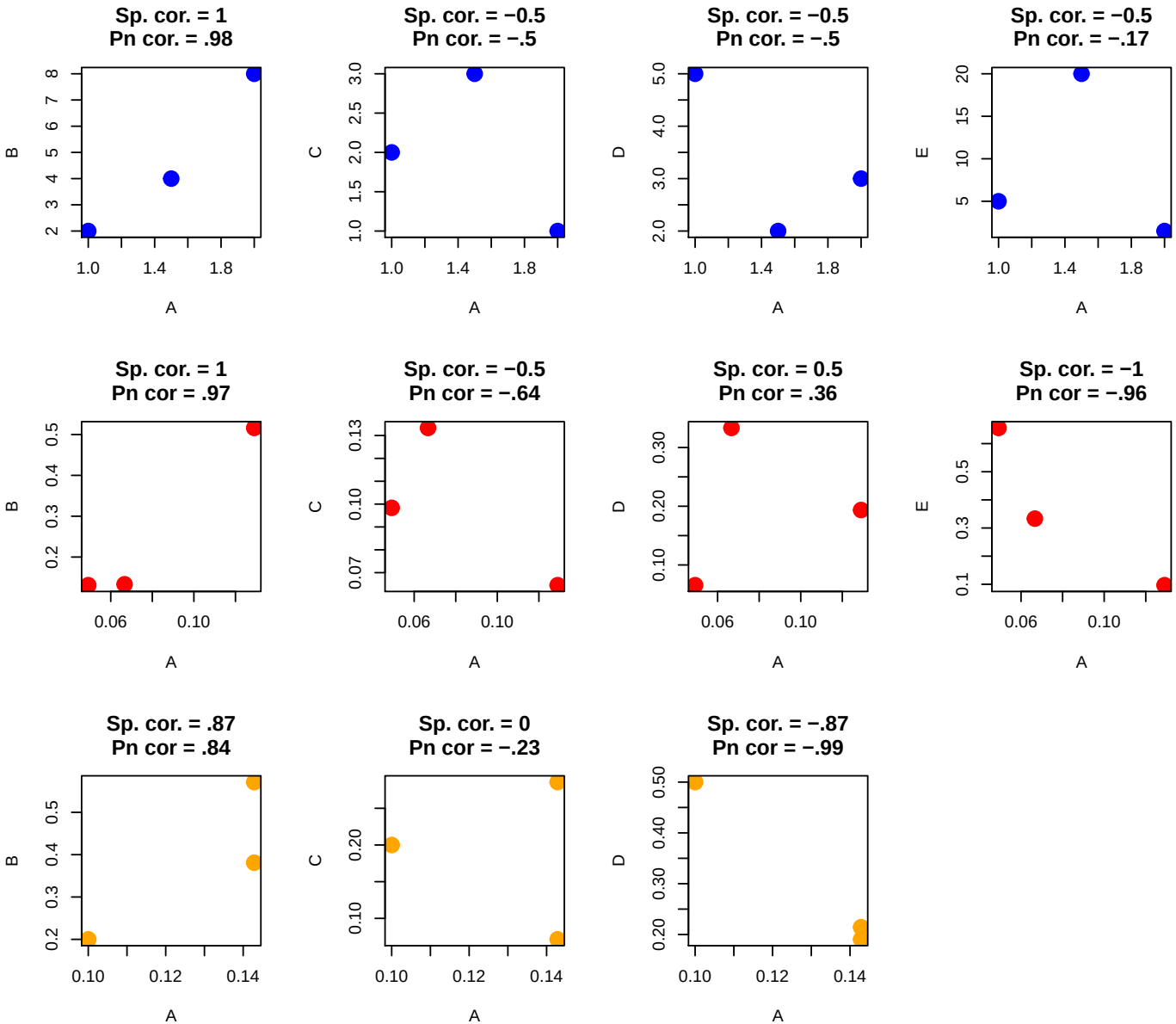
```
print(round(s.p, 3))
```

```
##        A     B     C     D     E
## s1 0.067 0.133 0.133 0.333 0.333
## s2 0.049 0.131 0.098 0.066 0.656
## s3 0.129 0.516 0.065 0.194 0.097
```

```
print(round(s.p2,3))
```

```
##        A     B     C     D
## s1 0.100 0.200 0.200 0.500
## s2 0.143 0.381 0.286 0.190
## s3 0.143 0.571 0.071 0.214
```

Spurious correlation in action



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 $y = m \times x + b$, and the correlation coefficient is a standardized covariance relationship. But correlation is defined in terms of covariance as we shall see later.

univariate thought experiment:

50% of 50%?

- how many $ are you paying?
- what proportion are you paying?
- how did you arrive at that value?

Lessons?

1. not adding or subtracting
2. can't derive underlying number

bivariate thought experiment:

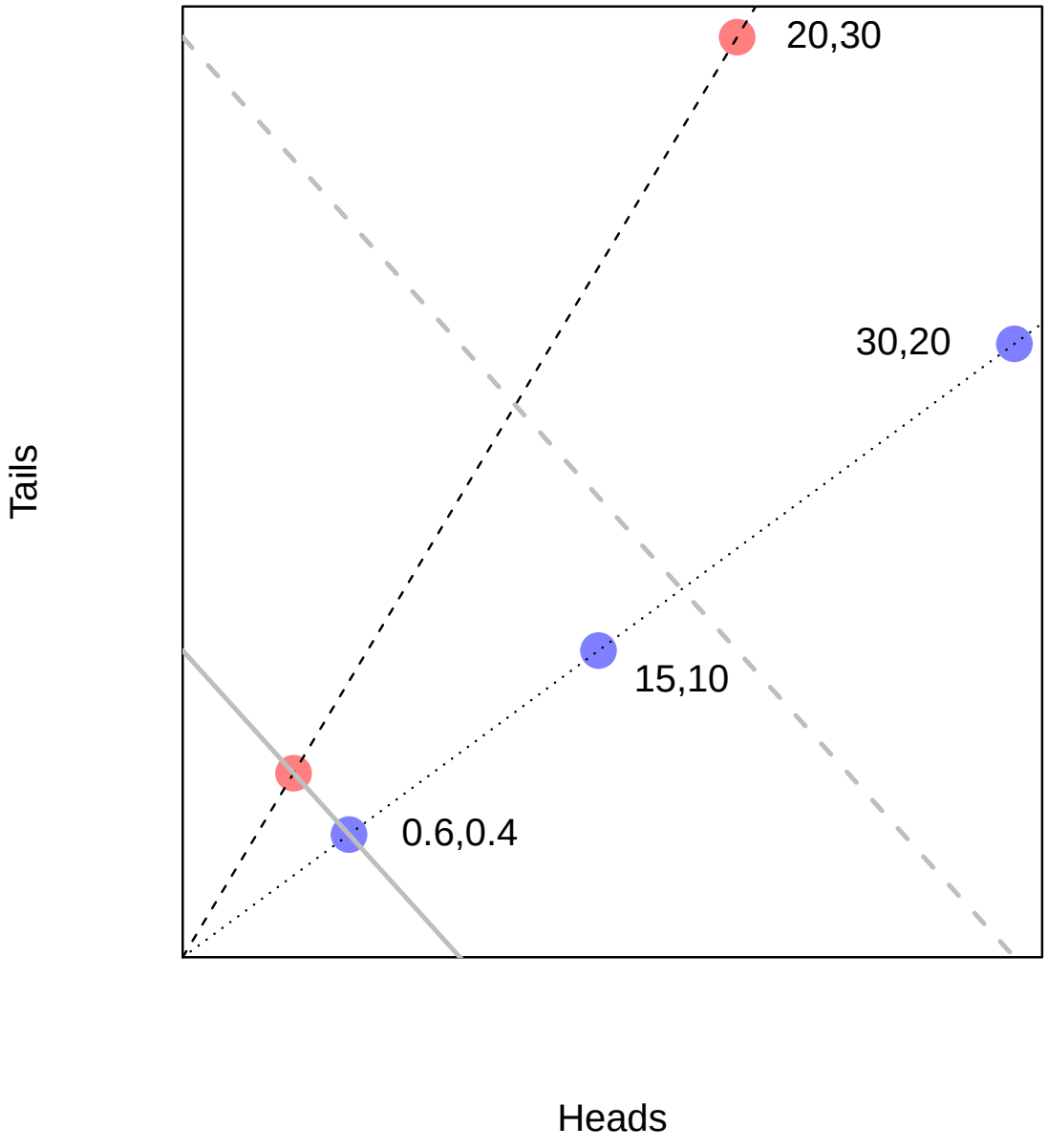
Unfair coin: 60% heads

- How many tails?
- How many flips?
- What information do we really have?
- What is the correlation between heads and tails?

Lessons?

1. not adding or subtracting
2. can't derive underlying number
3. only have ratio
4. working on D-1 simplex
5. negative correlation bias: negative correlation is meaningless
6. these properties are generalizable to as many variables as is necessary

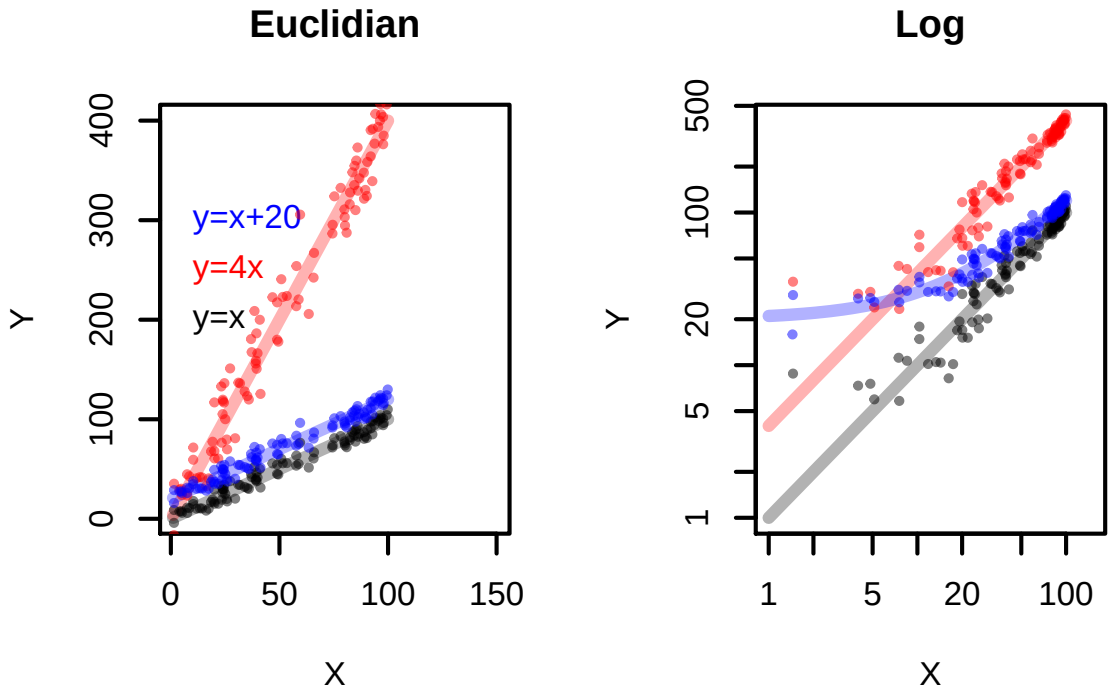
A geometric intuition (Pawlowsky-Glahn)



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"
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

Correlations plotted



1. Constant ratio relationships cannot have an intercept, becomes a non-linear in log-space
2. Constant ratio relationships can have a slope $\neq 1$, these become the intercept in log-space
3. Familiar measures of correlation do not require an intercept of 0
4. False positive correlations for both positive and negative correlation but different reasons

recasting correlation as ratios between parts

$$Var(X) = \langle (X - \mu_x)^2 \rangle$$

$$Cov(XY) = \langle (X - \mu_x)(Y - \mu_y) \rangle$$

$$cor(XY) = \frac{Cov(XY)}{\sqrt{Var(X)Var(Y)}}$$

Note that Spearman's correlation is the same measure on ranks

This will identify linear relationships, but does not account for slope or intercept

Now we know we are looking for linear relationships of ratios and how these manifest on linear or logarithmic plots.

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

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)
3. we must have a linear line
4. we must account for scatter

The ϕ metric (Lovell)

We first transform our data 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))
```

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

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

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

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

The ρ metric (Erb)

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

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

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

Summary

- compositional data are any data represented by a constant sum
- only ratio information is obtained
- simplex and only simplex operations
- 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)
- Why not calculate two numbers, slope and correlation (Egozcue pc)