

Simple biplot

99

19 July, 2016

To run this file: `Rscript -e "rmarkdown::render('make-interpret_biplot.Rmd')"` # The compositional biplot

When analyzing and interpreting compositional data, it is important to remember that we are examining the variance in the ratios of the underlying data, and not directly examining abundance. The first tool that we will use is the compositional biplot. This is generated by the following set of steps:

1. remove essential 0 values
2. perform any additional filtering (sparsity, minimal abundance, minimum sample count, etc)
3. adjust remaining 0 values with the `zCompositions` package
4. perform the clr transform on the data
5. conduct a singular value decomposition using `prcomp`
6. display the results in a principle component plot

Let us see how this works in principle. We will make a sample dataset that has 30 samples and only eight OTUs. Samples will be in two groups. The first 20 will and the last 20 will differ. OTU A will be more abundant in the first 20 samples, and less abundant in the last 10, OTUs C and D will be the opposite, with OTU D having a greater difference between groups than OTU C. OTUs B, and E to H will be highly variable, but the variation will be associated with samples randomly.

Principle component plots display the projection (shadow) of your multi-dimensional data onto a lower number of dimensions, and here the maximum number of dimensions possible is seven. The compositional biplot displays the results of your experiment in a semi-quantitative way.

From this plot we can determine the following:

1. The first two principle components explain about 80% of the variance in the dataset. This is very good. The greater the variance explained, the greater the confidence one can have in the projection. If we thought of this as a shadow of the data, then most OTUs and samples would have fairly distinct locations.
2. The samples (in black) partition into two groups: samples 1:20 on the left and samples 21-30 on the right with a clear separation between them. They are separated on PC1, which indicates a lack of confounding effects on the split between samples.
3. The length and direction of the arrows (OTU locations) is proportional to the standard deviation of the OTU in the dataset. So we can see that OTU A is highly OTU along the same direction as are samples 1-20. We can interpret this as OTU A is more abundant in these samples than in samples 21-30. Likewise for the other OTUs.
4. OTUs C,D are very close together; this is referred to as having a short link. The length of a link is proportional to the variance in their ratios. In other words, the variance of the ratios of these two OTUs will be fairly constant.

The first dimension is the one that displays the largest amount of variation in the data: for example, if your data had four OTUs, it would have 3 dimensions (like a rugby ball) and the first dimension would represent the long axis.

The principle components are arranged in decreasing order of the variance explained. If you have a strong effect driven by a single experimental OTU, then you will have a large amount of variation explained on the first axis, and a much smaller

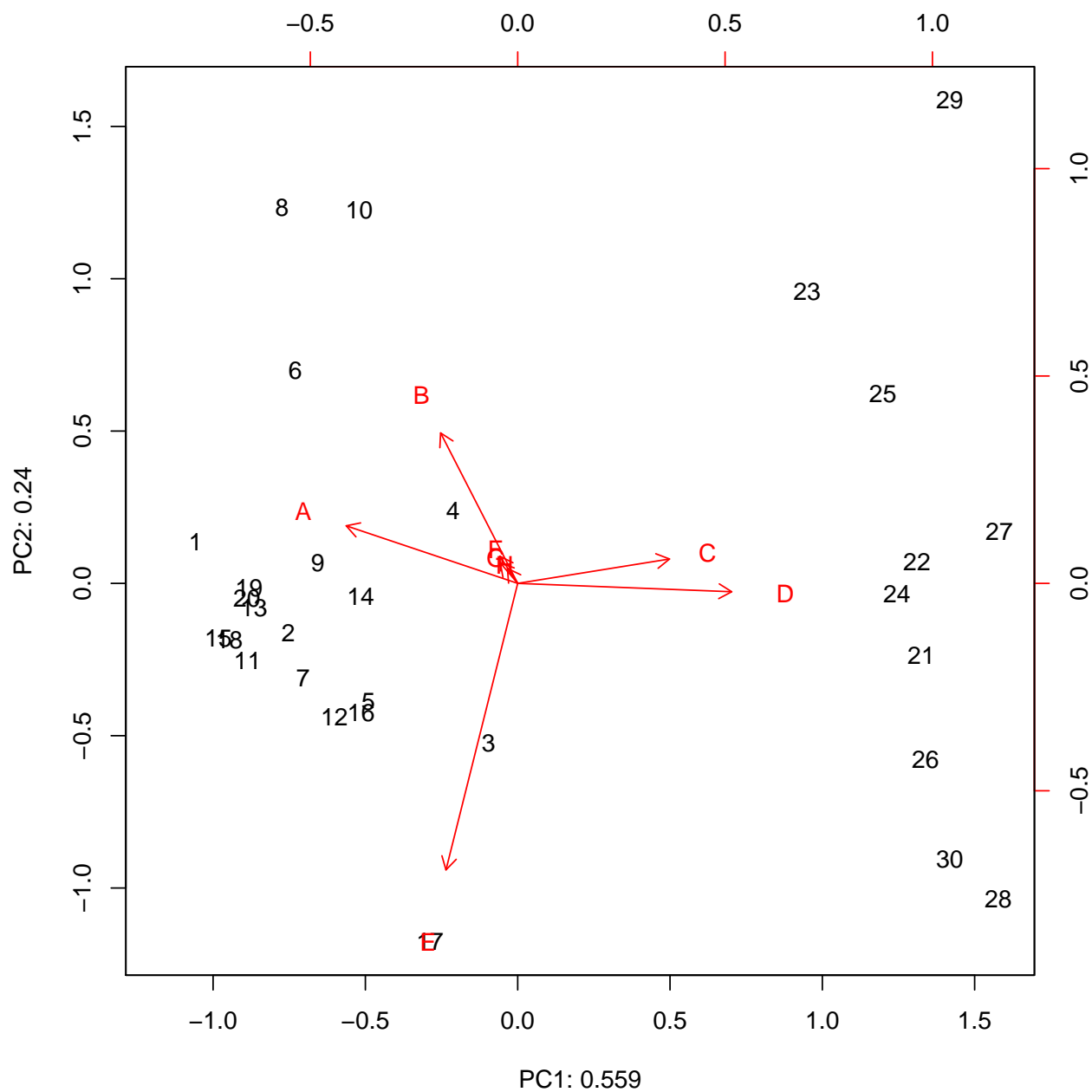


Figure 1: Counts were generated randomly for eight OTUs labeled a_h and were free to range from the values indicated for each OTU. Thirty samples were generated, and OTUs A,C and D were differentially abundant in the first twenty and last 10 samples. The others were of widely different abundances, but with totally random change between samples.

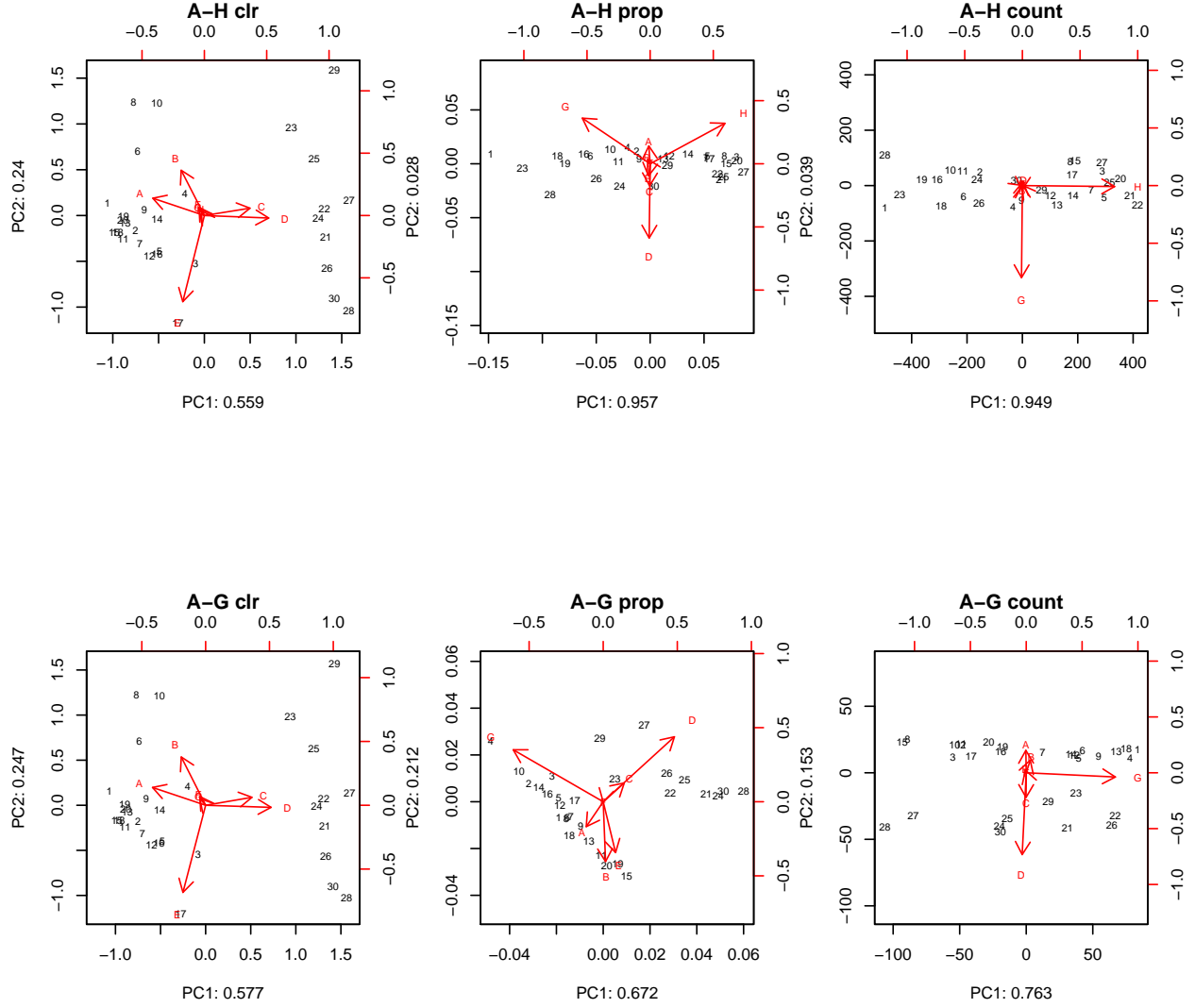


Figure 2: The advantage of compositions is that the results are largely invariant to subsetting. That is, we get essentially the same answer with all (A-H), and with a subset of the data (A-G).