Multivariate data and compositions

gg

April 8, 2016

To run this file: Rscript -e "rmarkdown::render('multi_comp.Rmd') ## Types of data

Univariate data have is one variable per sample. This is the typical kind of data generated by biologists. For example, you measure the height of people in a room, divide them into male and female and compare. In this case, we have one variable (height) measured for two groups (male and female). These data are unconstrained, within the bounds of human height, and are in general what traditional statistical tools were developed for.

Multivariate data have more than one variable per sample. Multivariate data can be independent or dependent. So for example, if the variables are truly independent (randomly chosen, not linked) then each may be treated as univariate. For example if we measure height, hair color and handedness of people, then each person is an observation that holds three variables. We can assume that these are (relatively) independent, and all statistical tests univariate and multivariate should be valid and importantly, multiple test corrections would be valid.

However, if the variables are dependent, then we have many unappreciated problems. Unfortunately, the typical high throughput sequencing dataset is multivariate and highly dependent. This dependency is forced upon the data by the sequencing instrument itself.

Let us set up a thought experiment. Let us imagine we have a very simple dataset composed of 100 samples and three taxa. Note that everything about this example generalizes to datasets with more samples and more taxa, it is just difficult to show this with an n-dimensional graphic. In this example, we will generate counts for 100 samples from a single experiment where the taxa abundance can range anywhere between 1 and an arbitrary number of counts. We will assume that we are actually counting the number of molecules belonging to each taxon, and that there is no practical upper limit on the number of molecules.

We will save the special case of 0 for later. Now we have 100 multivariate observations with three variables.

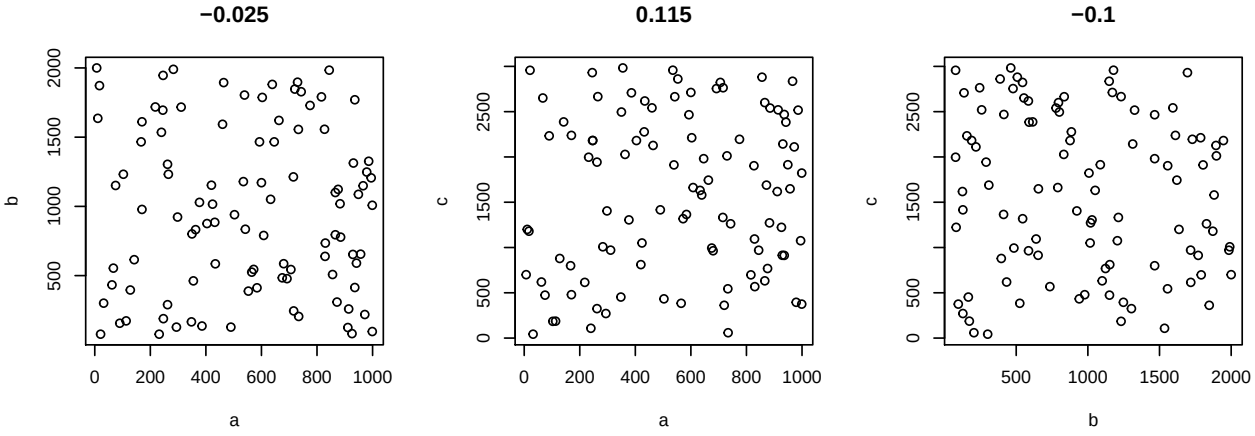

Figure 1: Counts were generated randomly for three taxa labeled a, b and c, and were free to range from 1 to 1000 for a, up to 2000 for b, and up to 3000 for c. One hundred samples were generated. The correlation coefficient of each sample vs. each other is shown above the plots.

The plots show that the variables in each of the samples are randomly placed. Therefore, we can see that the data in each pair of samples are essentially uncorrelated, as we expect for randomly generated data. This is what we would expect for randomly generated data where each point is absolutely independent of each other: in other words this is the best case scenario that would be rarely seen in a biological context. We shall now constrain the data to a constant sum and see how this affects the shape of the data.

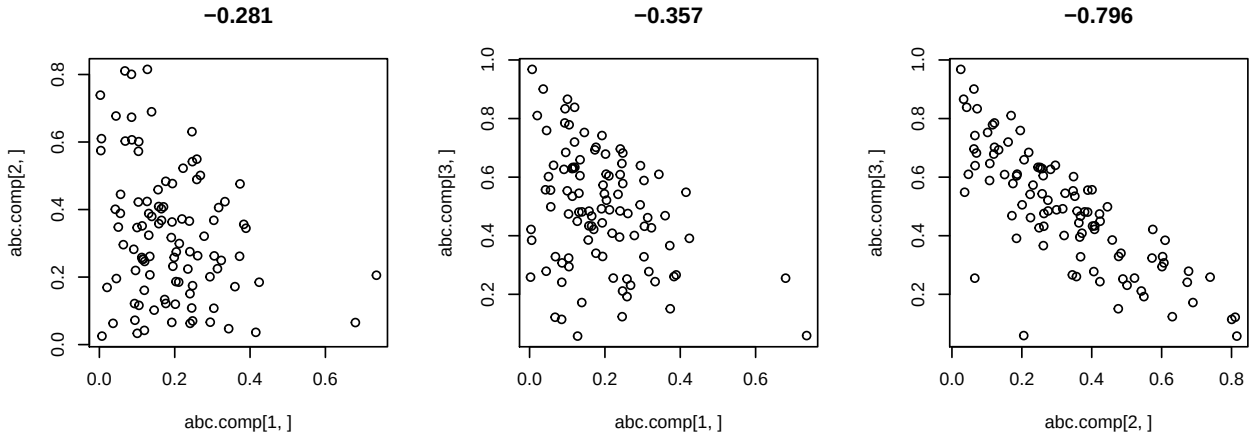

Figure 2: The same data as above were constrained to sum to an arbitrary sum: in this case, the data were converted to proportions. Any arbitrary number (percentage, ppm) has the same effect. The correlation coefficient of each sample vs. each other is shown above the plots. We can see very clearly that the associations between each of the datapoints are now not independent.

The simple act of converting each count to a proportion markedly skews the data. Here we can see that the data now appear much more correlated and constrained. While this is a simple example with three taxa, it is generalizable to any number of samples, and to any number of variables (taxa) in a sample.

Why is this so?

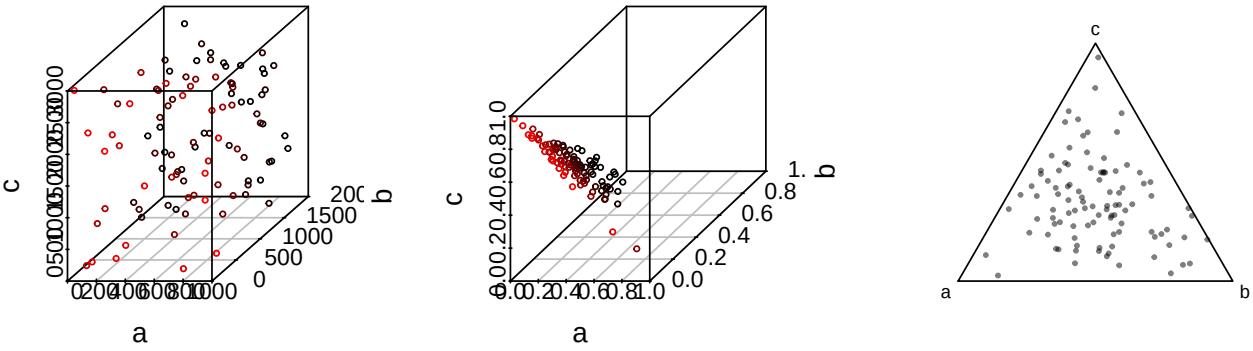

Figure 3: Plotting these data in three dimensions shows the reason for the non-independence. The unconstrained data are on the left, the constrained data are in the middle, and the right side shows the data as if we were looking directly down onto the plane of data seen in the middle box. Data points are coloured from red to black to show their place on the b axis.

The essential problem is one of geometry. When the data are unconstrained, as shown in the box on the left, the data points are scattered uniformly within the box. This is a visualization of multivariate data that are unconstrained and uncorrelated. Essentially, in data of this type, knowing information about one datapoint gives no information about any other datapoint.

When the data are constrained to a constant sum as shown in the middle box, then the datapoints collapse to a flat plane within the box, with limits at the corners. This plane is called a simplex, and occurs because if we know information about all points but one, then we know the information about the last point as well. For example, if the data must sum to 1, then knowing that the previous 10 points have a sum of 0.9, we know that the last point must have a value of 0.1. Thus, we can see intuitively that the data are constrained. We can re-orient our view to look directly perpendicular to the simplex plane and observe how the three dimensions of the data map onto the two dimensional space on the right.

The data are clearly in a very different geometry than are the unconstrained data points. Moving from place to place in the box on the left is an additive process since the difference between points is a linear distance. In the simplex, moving from point to point is a multiplicative process, and we think of differential taxon abundance in terms of being multiples of the original abundance.

In order to analyze such data, we must place ourselves on the simplex, or modify the data to move it from a constrained to an unconstrained simplex. Thus, we need to recast our thinking into a compositional data analysis (CoDa) way of thinking.