Hello again

I started this CoDa debate and I want to continue it now, since I think it strikes at certain basic principles of statistics and data analysis, and how we statisticians project our discipline to practitioners such as geochemists and biochemists.

I want to remind you how this all started, since the debate in this forum has started to diverge into other directions. After Alecos' talk at CoDaWork 2017, I asked a question about the ILR ratio in this diagram, which he presented in his talk, reproduced from a paper by Antonella Buccianti (Buccianti, 2015):

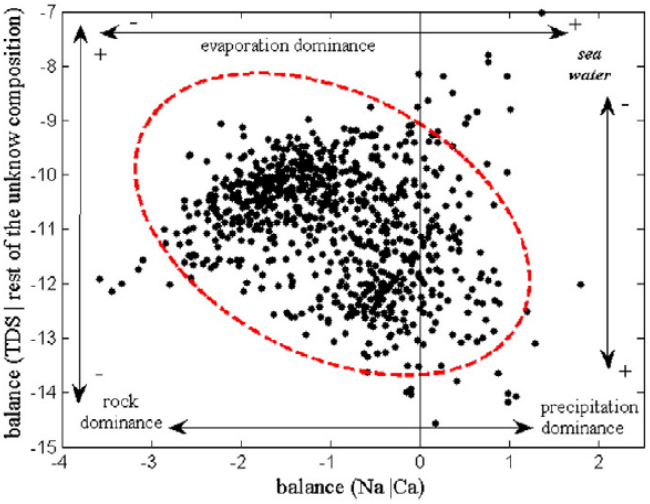
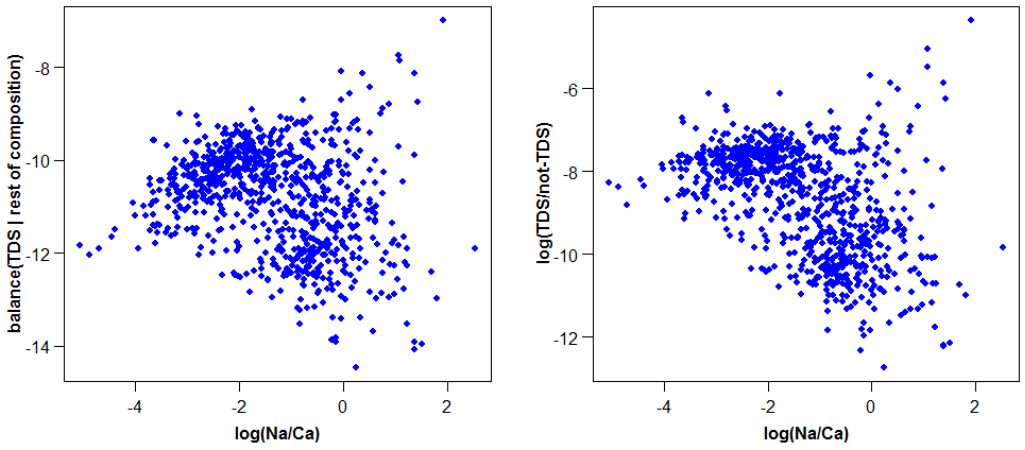


Fig. 6. Gibbs's diagram revised by using isometric log-ratio balances for European stream water samples. The robust confidence ellipse of 97.5% is also reported. Data located outside the ellipse are considered as outliers.

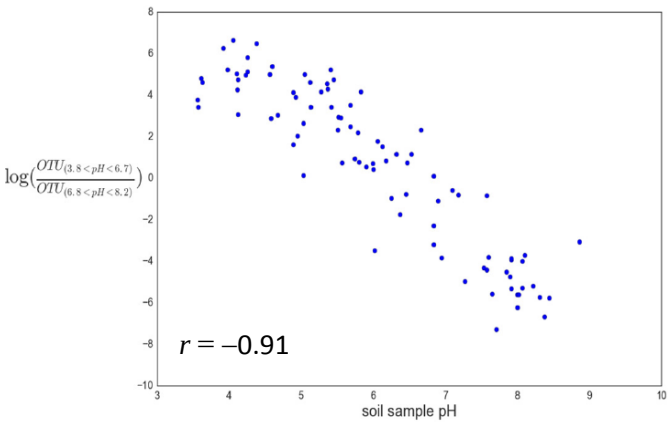
The usual Gibbs diagram would not use an ILR balance on the y-axis but the logarithm of TDS (total dissolved solids). I pointed out that log(TDS) is almost identical to log[TDS/(1-TDS)], because TDS is so small, and that this logratio was much easier to understand. So if one wants to use a logratio, which is recommended, and one with a clearer interpretation than an ILR balance, why not just use the "amalgamation logratio" log[TDS/(1-TDS)], which in this case has an analogy with a "log-odds". The difference between the ILR balance and this amalgamation logratio is not great (see Antonella's on the left, mine on the right):



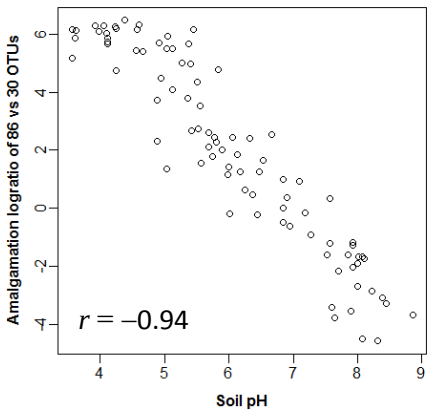
Juanjo immediately replied (at the conference) that my suggestion did not concord with the Aitchison geometry, but I really don't know why anything has to obey the Aitchison geometry if all you want to do is to show a scatterplot of two variables derived from a compositional data set. But Aitchison logratios YES! One wants to define a logratio that is interpretable to the practitioner, and plot it (on the y-axis in this case in order to expose a certain relationship with what's plotted on the x-axis. As many authors in the literature have stated, the ILR is difficult to interpret, so I'm really wondering -- again -- why anyone uses it! Or why CODA workers have inflicted it on poor practitioners. It seems to be only a mathematical nicety, as John Aitchison himself suggested, and nice if one wants to do multivariate analysis of the complete compositional data set, but not necessary at all --- see my CoDaWork 2017 talk "Towards a pragmatic approach to compositional data analysis" online at **youtube.com/CARMEnetwork**.

The previous contribution to this debate was by Jamie Morton, and I have been a real nuisance to him trying to understand his article, getting the data, re-analysing it, and... I find almost the same situation: an ILR (in practical terms, impossible to interpret) which can be easily substituted by an amalgamation logratio (in practical terms, easy to interpret) and does essentially the same job, without having the nice mathematical property that, in my opinion, is not necessary at all to have in the circumstances of the practical problem.

The essential facts are these. Jamie takes a 88x116 data set of 88 soil samples and 116 OTUs (operational taxonomic units, or species here), a subset of a much larger one of 7396 OTUs, in other words a subcomposition - the original data are actually counts and the matrix is quite sparse, lots of zeros. He has pH values for the species (no need to go into how these are computed) and splits the species into low and high pHs, 86 and 30 OTUs respectively. He computes the ILR balances of the 86 vs. 30 OTUs in the prescribed way (geometric mean of 86 OTU counts, divided by geometric mean of 30 OTU counts, then log-transformed, with the normalizing constant of the balance) and plots it against pH of the soil samples, and gets the scatterplot below, on the left, and the correlation = –0.91. I computed the amalgamations (i.e. simple sums) of the parts in the numerator and the denominator respectively, took the logarithm of the ratio of these sums (I could call this an "amalgamation balance" where the word balance is more appropriate), didn't use any normalizing constant and got the scatterplot on the right, correlation –0.94. To add to my scepticism of the ILR balance, I know why the normalizing constant is applied but I see no practical need for it in the present example.



Jamie's plot, y-axis is an ILR balance

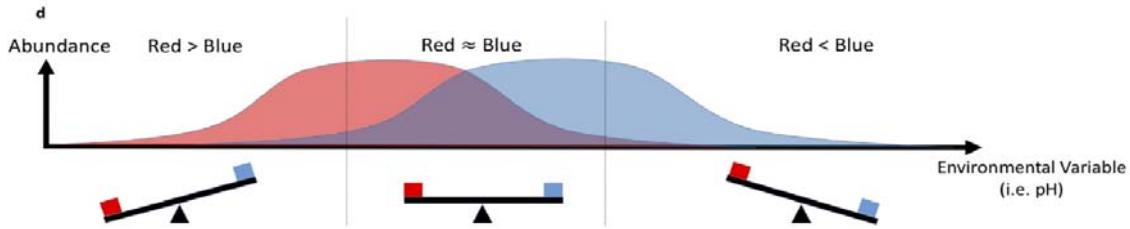


My plot, y-axis is the corresponding amalgamation balance

Now I ask myself: why would anyone prefer the complicated definition of the y-axis on the left over the simple definition of the y-axis on the right, when the point was to show the close relationship of the contrast between the numerator and denominator parts in relationship to the pH? (and the correlation is actually a bit stronger with the simple amalgamation logratio). Maybe I find the reason here, and I quote from Jamie's paper, with his diagrams:

> The balance concept proves to be a very powerful technique for investigating how these groups of organisms change relative to each other as pH increases. Recall the cartoon example in Fig. 1d. If there are two distinct unimodal species distributions, the balance pivots from being weighted by Red individuals in low pH to being weighted by Blue individuals in high pH. The same phenomenon occurs here, except that there are multiple species on the left end of the balance and multiple species on the right end of the balance.

(here is Fig. 1d)



But notice the accent on the concept of "weighting" of a balance, which is not what the ILR balance actually is. The above diagram is more relevant to real weights in the form of accumulated parts, which is what I am doing with the amalgamations, taking the sum of the red parts (species) and "balancing" it with the sum of the blue ones. Look at the vertical axis in the figure above, indicating "Abundance" and the shaded areas indicating integration (i.e. summation) of abundances. It is hard to figure out how geometric means could be "balancing" anything, since their values are not a simple aggregation of parts and depend on the particular mix of constituent parts, so the analogy of an ILR with the above physical balance is a real stretch of the imagination. In fact, it is a misnomer, and has given rise to many people actually thinking that it is something like a ratio of aggregated weights, which is the way I saw it interpreted at CoDaWork 2017, both in the talks and the posters. Speaking to some students at the poster session, it was clear that they thought ILR balances were like the above picture, a balance of integrated "weights" ! It is the amalgamations of parts that form balances.

Here's another defence of the ILR balance by Jamie, but his explanation clearly suggests amalgamations (i.e. equivalent to accumulated relative abundances in this case), not geometric means:

spectrum. A single balance can capture information about the transition from a high relative abundance of Red individuals in low-pH environments to a high relative abundance of Blue individuals in high-pH environments. In low-pH environments, the balance is positive, since there are proportionally more Red individuals than Blue individuals. When the Red and Blue individuals are present in roughly equal proportions, the balance is roughly zero, representing a turning point, transitioning from a Red species-dominated community to a Blue species-dominated community. As the pH increases, the balances become increasingly negative, since there are more Blue individuals than Red individuals. This balance effectively encodes the niche separation of Red and Blue individuals across the pH gradient.

"Proportionally more Red individuals" means more Red individuals summed together, and not the geometric mean of the counts of the Red individuals, which could actually decrease when there are more Red individuals, depending on how they are distributed across the constituent parts. "When Red and Blue species are present in roughly equal proportions" means that their sums are roughly the same, not their geometric means... And so on.

Another small technical issue. In order to compute the ILR, Jamie had to add 1 to every value in the data matrix of counts, due to the many zeros. In my amalgamation logratio, if the sums of the abundance counts in the numerator and denominator are all positive, the problem is solved and nothing has to be added to the data. In this example, there were unfortunately still a few zeros in the sums in the denominator, so I added a single one to the sums in the numerator and the denominator to avoid the zero problem. So I didn't have to increment the original data values themselves, just the amalgamations. Forming amalgamations can alleviate the zero problem, because one might be lucky in having amalgamations that are all strictly positive.

If anyone can give me a reason why -- in these two papers -- either Antonella or Jamie should be using ILR balances of two sets of parts, i.e. the log of the ratio of two geometric means, rather than the log of the ratio of the respective amalgamations, I would like to hear it. I would even venture that the whole ILR business is a "red herring"* being cast at practitioners who don't fully understand the difficulties of this concept, but put their trust in the judgement of theoreticians who claim some benefit in their use. We can make things much easier for compositional data analysts in their various fields, without any loss in practice, only gains in understanding and interpretability.

Michael Greenacre

29 January 2018

* red herring: Something that misleads or distracts from a relevant or important issue.