Rejoinder to Juanjo & Vera and some more comments

Michael Greenacre, Assos, Turkey, 2 August 2017

Thanks for the detailed response, which contains some very interesting details, which I would like to comment on. I appreciated especially the pedagogical section on orthogonality/ uncorrelation/ etc... I also realized that this debate that I started is practically the same one started by John Aitchison with his "lesser Goilbird" example 14 years ago, so history is repeating itself. Clearly, the matter has not been resolved yet.

1. Complexities of isometric log-ratios

Your initial remarks misrepresented what I said, so I would like to repeat first what I said, referring to the CodaWork 2017 conference:

" I was surprised to see so many papers using ILR transformation, based on ratios of geometric means. As a univariate concept, this is extremely difficult, if not impossible, to explain to a practitioner. I found that some presentations at the conference talked about a balance as if it were a ratio between amalgamations of parts in the numerator and denominator, which it isn't."

Note that I was referring to **ratios** of geometric means, to which you replied:

" But the first thing we want state is that we do not consider geologists, or any other practitioner, unable to understand something like a geometric mean."

You continue a sentence later, contrasting geologists, whom you maintain understand balances, with "statisticians and mathematicians", which is the group I belong to:

"Not so strange to us is that those who raise their voice stating that geometric means (or even logarithms) are complicated, while arithmetic means are easy, are mainly statisticians and mathematicians, but that is the world we have to live in."

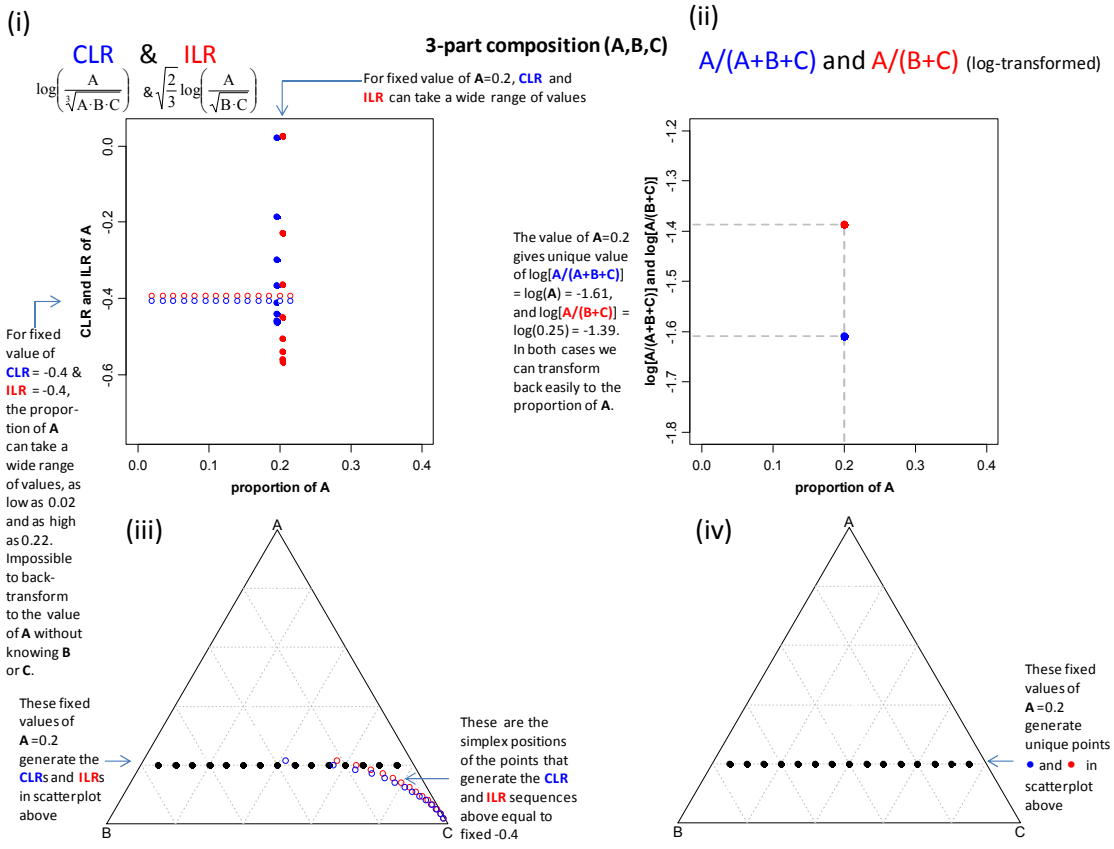
So I raise my voice again, and repeat that a log **ratio** of two geometric means, which is what most balances are, is a very difficult concept to understand. Furthermore, I hazard the opinion, suggested by the way several oral and poster presenters at CodaWork 2017 interpreted their results that involved balances, that most practitioners do not actually know what they are dealing with – an isometric log-ratio is a very complex univariate concept. Personally, I have no problems with logarithms, and I have no problems with geometric means (I have often double-centred a log-transformed data matrix) – it's the **log-ratio of two geometric means**, intended as a substantive univariate measure, that I am concerned with.

It is unfortunate, in my opinion, that the term "balance" has been assigned to the log-ratio of geometric means --- it suggests the stoichiometric concept of balancing, which a "balance" clearly isn't. I think practitioners think they are dealing with a ratio of amalgamations in some sense, i.e. in a "ratio of means" sense at least, but unfortunately for them it is one serious level more complex because the means are geometric. On the next page I give some very simple examples of just how confusing a "balance" can be.

Suppose you have a 3-part composition [A B C] where A+B+C=1. There was indeed a poster at CodaWork about occupations and time at work divided up into A = Physical Activity (PA), B = Sedentary, C = Standing. On the poster there was exactly an ILR of the form A over the geometric mean of B and C. Males and females were compared on this ILR, but what is the substantive meaning of this comparison? If the ratio had been A/(B+C) [ratio of PA to sitting&standing] on a log-scale it would be perfectly understandable. For example, if A=0.2, then B+C=0.8 and the ratio is 1/4 (four times more sitting&standing than PA) – see the red dot • in Fig. (ii, top right) below at height log(1/4) = –0.139. And if I were comparing males and females on this log-ratio, the result would be perfectly comprehensible, leading to the estimated multiplicative difference on this ratio between the two genders. The blue dot • in Fig. (ii) is a trivial log-ratio, A/(A+B+C) = 0.2, put in to compare with the CLR in Fig. (i).

So now see the analogous CLR and ILR in Fig.(i, top left), where sums are replaced by geometric means. The blue and red solid dots in a vertical stripe are all positioned at A=0.2, just separated a bit to see them better. By varying B and C, while keeping their total=0.8, you can get CLRs slightly positive to values as low as –0.45 (for CLR) and –0.6 (for ILR). So if I'm finding a difference between men and women using that ILR, I really do not know what that difference means, it depends on A, B and C in a very complex way (and this is a very simple example of an ILR!). Likewise, if I fixed the CLR and ILR at a value –0.4, this could correspond to values of A=Physical Activity as low as 0.02 and as high as 0.22 (the open red and blue circles shown horizontally).

The simplex plots in Figs (iii) and (iv) show where the closed and open dots are situated in triangular coordinates. The row of black dots are values corresponding to A=0.2 (20% PA), varying B and C. They just generate the single dots in (ii), but they generate the vertical sets of solid dots in (i). And when CLR = ILR =–0.4, the various possibilities for A, B and C correspond to the two curved sets of open dots in Fig.(iii). If anyone can enlighten me how and why I would ever use this CLR and ILR as univariate statistics in a study such as this, please let me know!



To summarize, I see no reason why a researcher studying the composition of the three components physical activity, standing and sitting, would be interested in converting to ILR ("balance") coordinates, and then compare males and females on these variables, rather than form some ratios that make substantive sense, for example the ratio of physical activity to the other two activities combined (i.e. amalgamated, summed), and estimate multiplicative difference between these ratios for males and females. The important aspect is the ratio, and the log-transformation. For purposes of inference, distribution-free tests or log-normal based ones (with appropriate checking of assumptions) can be used.

2. Buccianti (2015) and the "Aitchison geometry"

You dismiss my comments about the scatterplot in Antonella's paper by saying:

"We do not want to start a discussion on the appropriateness of the alternative Gibbs diagram and cluster analysis presented by A. Buccianti (2015)"

But this scatterplot was exactly what triggered my question after Alecos' talk, where he reproduced the scatterplot, and then this whole debate. I maintain that the original Gibbs diagram, which everyone can understand, has been unnecessarily replaced in this paper by a scatterplot that has been made overly complex with the mere objective of using "balances" to satisfy certain mathematical requirements that are unnecessary for the geological objective. I would ask why a geologist would want to make a ratio of the **geometric mean** of the dissolved solid components with the **amalgamation** (not the geometric mean in this case, but it's still called a balance) of the rest of the components (there are 64 parts in this data set):

$$balance(TDS) = \sqrt{\frac{D-1}{D}} log \frac{(HCO_3 \times Cl \times SO_4 \times K \times Na \times Mg \times Ca \times SiO_2)^{1/D-1}}{\sum(all\ other\ parts\ up\ to\ 10^6)}$$

The simple and understandable alternative is the log-ratio of the amalgamation of the 8 parts in the numerator and the (present) amalgamation of the rest in the denominator, i.e. 10^6 minus the numerator, in the format of an odds ratio. That is, an "amalgamation balance".

As far as the "Aitchison geometry" being held up as the gold standard that everyone has to follow, you say about amalgamations:

"The problem with amalgamation is that you change your sample space. You move from the D-part simplex to the $(D - m + 1)$-part simplex when amalgamating m parts. And amalgamation is a non-linear operation in the Aitchison geometry, so things work differently after amalgamation. You are not projecting into a subspace, like when you are dealing with real variables. A particular issue is that amalgamation changes the order of differences between observations."

A practitioner wants to form an amalgamation for substantive reasons, and should be allowed to, uninhibited by mathematical niceties. We methodologists have to satisfy practical needs, and not expect practitioners to strait-jacket their ideas into restrictive mathematical theory. If I am studying proportions of time sleeping, working, going to the movies, relaxing with my

friends, eating in restaurants, reading, etc..., and I want to form an amalgamation of leisure time, I will ADD those parts together and not form a geometric mean, and that's it. I really don't know what the geometric mean of my leisure activities is supposed to be measuring. If I am studying beverage intake proportions: water, soft drinks, fruit juice, beer, wine, whisky, etc... and I want to form a category of alcoholic beverages for purposes of my study, I will ADD those ones together, and not form a geometric mean of them. What is being measured by the geometric mean of the alcoholic beverages I consume? I am not interested in the "Aitchison geometry" of the full composition, the geometry (and sample space) that I am interested in is the one that includes the amalgamations.

I similarly made the point, that in some CodaWork2017 talks, pre-coded amalgamations were presented as parts, and no-one made a single criticism of that (which I agree with). But you say:

"Another issue is to amalgamate first, before you start any analysis. This is like stating that you are unable to distinguish between the amalgamated parts, or that you have no interest in their separation. But you need to be aware that an analysis based on un-amalgamated parts might give results that are not consistent with those obtained using amalgamated parts."

To which I say: I am not interested in the analysis of the un-amalgamated parts. If I were, I would un-amalgamate them and I am sure the results would change, by definition, since the data are different now. Again you seem to hold up the "Aitchison geometry" as some form of gold standard, unrelated to the needs of the practitioner. This comes out again in your comment:

"One disappointing feature of *aggregation balances* is that they cannot be (rectilinear) coordinates of a vector space with perturbation and powering as operations. In fact, in the Aitchison geometry of the simplex, an aggregate balance does not appear as a straight line."

It's a straight line in my geometry, where aggregations (amalgamations) are admitted from the start. I reiterate: If I was pre-defining some aggregations (=amalgamations) and ratios of aggregations (aggregation balances), then my geometry becomes that of the aggregations considered as parts, and the only interest I might have in referring what I am doing to the space of the un-aggregated parts is to check how much variance I am explaining (see my talk at `https://youtu.be/49S76ufk1Ng`) or how close to the "Aitchison geometry" I am (also see my talk). Any difference between what I am doing and the "Aitchison geometry" is what I am clearly not interested in for the moment!

As a final remark here, the idea of weighting the parts, which I keep harping on at every CodaWork, makes the amalgamation of parts in (weighted) log-ratio analysis obey the *principle of distributional equivalence*, which is the founding principle of correspondence analysis. This has been proven in Greenacre and Lewi (2009)[1]. What this means is that if two parts are in exactly the same relative proportions across samples (e.g. beer and wine, where beer is always drunk, say, twice as much as wine across all respondents), then they can be **amalgamated** into a category "beer&wine", reducing the number of parts by one, without the (weighted) log-ratio

[1] Greenacre MJ, Lewi PJ (2009) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification* 26: 29-64

distances between respondents being affected. This principle is not satisfied by regular unweighted log-ratio analysis, which is what almost everyone is doing at the moment. If two such parts are amalgamated, which are clearly equivalent, the regular log-ratio distances between respondents are changed, so it's not the same "Aitchison geometry".
Similarly, just split a part (column of data set) *X* into two parts in any way you like, say 2/3 of *X* into *X1* and 1/3 into *X2*, so you have one extra part, but have added no extra information to the data set. But the log-ratio distances will change between the samples (rows), so the "Aitchison geometry" has been altered. The distances won't change, however, if you use weighted log-ratio analysis and corresponding distances.

3. The lesser Goilbirds

Not having read the texts you cite from the Proceedings of CodaWork 2003, I was really amazed to see that yes, I am just repeating history. This is what you write about John's comments then (my highlighting in red):

"J. Aitchison (2003) stated that the ilr treatment, being mathematically sound, was generally not necessary or even useless. In certain cases, he proposed the use of amalgamations and the associated log-contrasts as a more intuitive and practical way of dealing with those problems. We agree with J. Aitchison's (2003) statement
My complaint is not that such structure (referred to ilr and orthonormal basis in the simplex) *is unimportant, but that we must not let pure mathematical ideas drive us into making statistical modelling more complicated than it is necessary ...*"

Well, that's exactly it! I am sorry to bring this all up again after 14 years, as I agree 100% with the above. So then why are "balances" still being promoted and amalgamations rejected? Again you now harp on:

"when working with compositional data which are assumed to be adapted to the Aitchison geometry of the simplex, the (apparent) simplicity of amalgamation hides its non-linear character in that geometry."

Just not interesting, I'm afraid. The fallback on "non-linear character in that geometry" is not a good enough reason, especially if a practitioner is not at all concerned about the "Aitchison geometry" but just wants to understand the structure of the data set at hand in a way that respects the Aitchison **principles**. In any case, it would not be the first time that variables of "non-linear character" have been created in a study, it's happening all the time in regression modelling without any problems at all.

In summary, I think John was right. The time spent by the *lesser Goilbirds* is an excellent example, and the active and passive times should obviously be computed by amalgamation (summation) of their respective parts. Combining their parts using square roots or geometric means makes no substantive sense in the context of *lesser Goilbirds'* behaviour, and furthermore makes the resultant logratios uninterpretable for a behavioural ornithologist.