# Finding "the balance" between general theory and particular practice of CoDa

Alex Washburne, Montana State University

August 6, 2017

Michael Greencare raised what I believe to be a very good point. Balances are unintuitive and often difficult to explain. While they have properties that are useful in theory - orthogonality and isometry - their prescription as a universal tool across disciplines is, in my opinion, a bit premature. That said, I also use ILR balances for a particular problem in ecology. However, in line with the folk at Berkeley, I am also aware of alternatives to the ILR transform (e.g. the log-odds ratio) that are especially appropriate under particular parameteric assumptions about our data (e.g. multinomial data).

## First: Why my work (phylofactorization) has used the ILR, why it's hard to explain my results and why I use it anyways

### Life "multiplies and divides" - the algebraic geometric structure of our data

I arrived at the ILR transform somewhat ignorant of its existence. I was approaching the problem as a practicing mathematical biologist trying to develop a means of analyzing community-ecological datasets in light of the evolutionary tree (a sequential binary partition). Some of these community-ecological datasets are compositional, but others are not. For instance, it's not clear that a dataset on the numbers and types of birds observed in a trip to the Amazon, across a range of environmental meta-data and habitat types, is compositional.

However, while these data may or may not be compositional, they are biological, and the first principles of biology are often taken for granted: life "multiplies and divides". Populations fluctuate geometrically, and so models such as geometric brownian motions are far superior to capturing the noise in real ecological data. These data are best analyzed in log-space, after which the ordinary operations of arithmetic means and subtraction will suffice. Thus, given a vector of abundances, $\boldsymbol{x}$, (and barring particular parameteric assumptions discussed below) one can and should analyze $\log(\boldsymbol{x})$ when analyzing these

particular data, in practice. When doing so, the arithmetic mean of a sub-group (e.g. all hawks & owls) $R$,

$$\frac{1}{r}\sum_{i\in R}\log(x_i) = \log\left(g(\boldsymbol{x}_R)\right)$$

where $g(\boldsymbol{x}_R)$ is the geometric mean of our group, $R$. Thus, for me, the geometric means are not a result of a screwy amalgamation that we have to do because it has better properties than an amalgamation we'd prefer, but rather they are the consequence of a very natural averaging in log-space due to the algebraic-geometric structure of many biological data. To write this differently, the modern practice for modelling communities using a dynamical system looks like

$$x_i(t) = a_t x_i(t-1)$$

where $a_t \sim f(\theta)$ is a random variable and positive real number (the essence of our first principle that living things "multiply and divide"). Because of this modern practice, our geometric mean gives us the average fold-change and thus logarithms are the "natural" way of analyzing these biological data.

This algebraic-geometric structure need not hold in other fields, or even other kinds of datasets within the same field. Ultimately, the appropriateness of geometric averaging vs. arithmetic amalgamating will be a particular discussion within a field to understand the algebraic geometry of the data and, for me, the underlying processes by which the data are formed (life & death, migration and removal, reaction and flux correspond to $\times/\div$, $+/-$, and a mix, respectively) are a good place to start. In biology, an alternative algebraic-geometric structure exists in studies of evolution - the Wright Fisher Process, which I've shown in Washburne et al. (2016) to be most appropriately analyzed with amalgamations and differences (in arcsines, not logs).

**ILR balances are two-sample t-statistics**

I wanted to infer differences - which groups of species, $R$ and $S$, are most different? In log-space, I preferred to look at differences of means. However, I wanted to ensure that these differences of means didn't bias towards small/large groups. If all species had the same log-variance, $\sigma^2$, and I just took the raw differences of means of two groups, the variance would depend on the sizes of the groups:

$$\mathbb{V}\left[\log(g(\boldsymbol{x}_R)) - \log(g(\boldsymbol{x}_S))\right] = \sigma^2\left(\frac{1}{r}+\frac{1}{s}\right).$$

Thus, I wanted a variance-stabilizing transformation so that a measure of difference across a dataset (e.g. the variance in the difference of means, which tells us about the negative covariance between the groups), I needed to divide by the standard deviation expected under the model in which all the parts' variances are equal. Thus, I obtained my key statistic to be used for analysis:

$$y = \sqrt{\frac{rs}{r+s}} \log\left(\frac{g(\boldsymbol{x}_R)}{g(\boldsymbol{x}_S)}\right).$$

I noted that this is just a two-sample t-statistic for data assumed to be log-normal. Thus, this statistic will work well with data which are either log-normal or clr-normal. Their compositionality (or not) is irrelevant - multiply our vector of abundances by any constant, whether it's the inverse of the geometric means of the parts (used in the clr transform) or the inverse of the system size (used to make our abundances into compositions), and that constant cancels out. We can see again how this relates to the algebraic geometric structure of our data - in other data for which we're taking arithmetic means and differences, the addition of a constant to all parts is just a translation, and our differences-of-means are translation invariant.

**ILR bases as hierarchical regression - controlling for previously inferred partitions**

Regardless whether we use arithmetic means of logs (geometric means) or arithmetic means / amalgamations of raw data, my next question was: how do I control for a previous inference? For instance, suppose I find two groups, $R$ and $S$, whose ILR transform $y$ maximizes my objective function. How do I look for other groups, controlling for the one I just found?

I decided that I would control for groups I have already found by partitioning my data - separating my data into two groups, $R$ and $S$, already identified, and then repeating the analysis within each group. In my submission to CoDa work 2017, I illustrate that phylofactorization is a graph partitioning algorithm and also a form of hierarchical regression. Here's why that is not a bad thing. Below I've tabulated a mock dataset of an imagined (and highly exaggerated) set of counts of Lizard, Snake, Hawk, Owl and Mouse:

| Species | Environment A | Environment A | Environment B | Environment B |
|---|---|---|---|---|
| Lizard | 1 | 1 | 1 | 1 |
| Snake | 1 | 1 | 1 | 1 |
| Hawk | 1 | 1 | $10^5$ | $10^5$ |
| Owl | 1 | 1 | $10^5$ | $10^5$ |
| Mouse | 1 | 1 | 1 | 1 |

We can immediately see that Hawks and Owls are hyper-abundant in environment B. Are these data compositional? Possibly - possibly there is very little effor going into sampling Environment A (in the microbiome, the differences in sequencing depth, i.e. effort, are often not controlled for in compositional analyses). Suppose I told you that each column was a composition (we could re-scale the counts in Environment A so that this illustration holds and the column sums are constant), and suppose we analyzed this system using clr-coordinates. Converting each column to proportions, the geometric mean of the first two columns will become $1/5$ and the last two columns $\frac{3^{\frac{1}{5}}}{10}$. The geometric mean changed, and that would cause the CLR coordinates to all change, and so a

multiple-regression analysis of these data would reveal that all species change when, as we see, Lizards, Snakes and Mice do not change relative to one-another and Hawks and Owls do not change relative to one-another.

Thus, if our first ILR coordinate separates {Hawk,Owl} from {Lizard, Snake, Mouse}, our second ILR coordinates should either compare Hawk:Owl, or Lizard:{Snake,Mouse}, or Snake:{Lizard,Mouse} or Mouse:{Snake,Lizard}. Doing so will control for a previously inferred shift in abundances, {Hawk,Owl}:{Lizard,Snake,Mouse}, already accounted for.

This hierarchical nature of the ILR transform makes it very challenging to explain. For downstream ILR balances, I have to remind people which group is in the numerator & which is in the denominator. I agree with Dr. Greencare that this is challenging to explain. I usually explain it in terms of hierarchical regression and controlling for ratios of groups we've already identified as important. If we looked at the clr transform for each part, or even for the log-odds ratio for each part, we would see significant differences for each part in Environment B. However, by controlling for the first split {Hawk,Owl}:{Lizard,Snake,Mouse}, we can correctly describe these data, were they compositional, as one-dimensional (their changes can be described as changes in a single ILR coordinate). Crucially, this is also why I build my ILR balances from the root onward for each dataset - by looking at the ILR transform as a means of clustering or graph-partitioning, our splits can carry some intuition.

### Mini-discussion

Thus, I arrived at the ILR transform not from theoretical principles of orthogonality, sub-compositional coherence and isometry. Rather, I arrived at the ILR transform through (1) the algebraic-geometric structure of biological data, (2) a need for a standardized difference of means (e.g. a two-sample t-statistic), and (3) a desire to control for previously identified partitions. These three properties give a practitioner reason to use the ILR, and justify the clunkiness of downstream balances.

## Parametric Assumptions in the ILR

In connecting the ILR transform to the two-sample t-statistic, I briefly noted above that the ILR is appropriate for log-normal data. Data which are not log-normal may have different standard errors and their two-sample t-statistics are only asymptotically normal. For other data, non-log-normal data, we may wish to use other approaches.

For instance, consider a dataset whose columns, $\boldsymbol{n}$, are drawn from a multinomial distribution $\boldsymbol{n} \sim Mult(N, \boldsymbol{p})$. We are well aware that there exist maximum-likelihood techniques for performing regression on $\boldsymbol{n}$, most noteably the generalized linear model which considers regression on the logit or log-odds ratio. Multinomial random variables are stable to amalgamation (in that an amalgamation of a strict subset of the parts will, along with the remaining parts or

their amalgamation, yield a multinomial random variable). I would never look at isometric log-ratios of data assumed to be multinomial, even though they are compositional.

For another case, consider the Wright-Fisher process in biology, whose dynamics $\boldsymbol{X}_t \in \Delta^{D-1}\ \forall t$ are given by the stochastic differential equation

$$d\boldsymbol{X}_t = \lambda(\boldsymbol{p} - \boldsymbol{X}_t)dt + \sigma(\boldsymbol{X}_t)d\boldsymbol{W}_t$$

where $\sigma \in \mathbb{R}^{D\times D}$ and

$$\left[\frac{1}{2}\sigma\sigma^T\right] = \begin{cases} -X_t^i X_t^j & i \neq j \\ X_t^i(1 - X_t^i) & i = j \end{cases}.$$

Such process can be analyzed by amalgating parts and looking at acrsines of differences of amalgamated groups (see Washburne et al. 2016 - Novel covariance-based neutrality test reveals asymmetries ....). I would never analyze a system assumed to be Wright Fisher and infer selection intensities and competitive asymmetries using isometric log-ratios.

In general practice, there are myriad compositional processes - including many urn process, but many others as well - which are of relevance to the particlar field. I do not recommend these researchers always use the ILR transform. In fact, even for the microbiome, for which I developed phylofactorization, I see the ILR trasnform as an approximate tool, barring more rigorous understanding of the distribution of our sequence-count data (with hopes that they can lead to generalized linear modelling of these data). The reason is: not all processes in nature are subcompositionally coherent. Sometimes, hidden variables outside of the subcomposition carry relevant information, and so one naturally gets a distorted lens with screwy distances when looking only at subcompositions.

However, even if we have different algebraic-geometric structure in our data, some difficulties of the ILR transform will remain. In particular, there may often be a need to look at non-overlapping subcompositions, in which case the partitions and the general idea behind the ILR transform, as well as the difficulty of articulating downstream non-overlapping subcompositions (a split of a previous subcomposition), will remain.

## Final Remarks

I want to thank all of you for this discussion. Dr. Greencare - I think you are wise to be thinking outside of the ILR box for your fatty acid problem. Only use the ILR transform if it is both justified (in the assumptions about the algebraic-geometric structure of your data) and interpretable. At the same time, I argue there is merit in downstream balances, however difficult they may be to explain in an elevator speech, and the merit stems from the hierarchical regression I've articulated above.

For all the rest, I think there is much to be done. There are myriad compositional processes, and I believe the field of compositional data analysis will

become more popular if it is able to listen to and work with the particular concerns of practitioners in the field. After all, as the prof. from UC Berkeley said, in all of statistics, compositions are not far beneath the surface. We don't analyze multinomial data with ILR transforms, but they are parameterized by a compositional vector, $\boldsymbol{p}$, which can be factored in a manner similar to the ILR balances. We don't analyze the Wright Fisher process with ILR transforms, but it is indisputably compositional and we do look at differnces of amalgamated subcompositions. While these processes don't have the strict assumptions of scale invariance, subcompositional coherence, and whatnot, they do have assumptions appropriate for the field and often the hypothetico-deductive progress of science is advanced in a parametric fashion, one model at a time. Sometimes, the prevailing theory does not have the nice mathematical properties we want. That's okay - compositional data analysis will be stronger and more broadly applicable if it accomodates the myriad algebraic-geometric structures on the simplex, with the ILR always remaining as an invaluable tool for data assumed to be log-normal or logistic-normal.