Amalgamation – some thoughts stimulated by the Coda group discussions

August 19th, 2017 **Ron S. Kenett**

Immanuel Kant's work implies that we are equipped with some pure intuitions which allow us to award structure to reality. Kant put space, time and causation on his list but there still is a barrier preventing us from gaining access to the hidden 'things as they are'. Schopenhauer, more or less, agreed with Kant but thought that we had privileged access via our own bodies and minds. With this context in mind, the premise I build on is that applied statistics is about creating information from data. As an applied statistician, my concern is about information quality. For an introduction to information quality and the 8 information quality (InfoQ) dimensions see https://en.wikipedia.org/wiki/Information_Quality_(InfoQ).

Typically, statisticians are addressing questions such as: What is the population?, when are population characteristics (e. g. proportions) relevant?, what is the "correct" variance to attach to a mean or proportion? or when should we standardize (for comparison)? In addressing these issues, I embrace the approach succinctly stated by Terry Speed: "Practice before theory".

This note builds informally on email correspondence and is mainly the result of Michael Greenacre's encouragements to write it and related stimulating discussions. In putting some thoughts together, I will refer to 6 sets of emails listed below. These are listed to serve as reference and context. They however do not represent all the correspondence that went back and forth following CoDaWork17 and, therefore, might look eclectic. The points below are not listed in any prioritized sequence whatsoever.

Point 1: A mapping of CoDa methods to different contextual situations would help clarify what applies where. This requires a good taxonomy to effectively establish such a mapping.

Point 2: Amalgamation is affecting data collection, data organization and data analysis. This is an inevitable issue to deal with and good methods need to be proposed and adopted.

Point 3: CoDa is mainly about generalization. Scale invariance and subcompositional coherence are exactly doing that. Besides providing an analytic method that ensures such properties, proper methods for presenting what can be generalized or not, are needed.

Point 4: The modeling impact of the CLR, ALR and ILR transformations are different. Some more clarifications of these differences seem to be needed.

Point 5: In 2017, CoDa has gained increased acceptance relative to the 1980s. However, relative to its potential, much more can and should be done. An in-depth discussion on the CoDa adoption roadblocks could help facilitate this progress. Obviously, accessible and user friendly software is key to this. The R applications and codapack provide a partial answer, however CoDa is not embedded in standard statistical packages. Moreover, software is not the only impediment to reach wider interest in the research and practitioner's communities. To achieve this, a wide view perspective is needed combining sound mathematical and practical aspects.

The paragraphs below provide the background for these points.

1. The CoDa context

Nonzero measurements that make up a set of components lead to a CoDa context. There is however a difference if the components represent soil characteristics, microbial types, results in an anchored Likert scale reflecting levels of customer satisfaction or election outcomes. In all cases we have the basic elements of CoDa analysis, but they represent different scenarios. In the CoDa literature, these components are typically not evaluated for causality links and methods such as propensity scores or self selection bias assessment are not discussed. Regression models with compositional covariates, ANOVA and mixed models and specialized models such a phylofactorization have been used in various applications. Alternatively, as done by Michael, the compositional vector is treated as responses and possible explanatory variables include Northing and Easting. Pawlowsky-Glahn, Egozcue and Tolosana-Delgado in *Modeling and Analysis of Compositional Data*, Wiley, 2015, provide an overview of CoDa methods.

2. Amalgamation

Amalgamation addresses the heart of Kant's idea of "things as they are". The main approach seems to be aimed at reduction of dimensionality in a way that makes domain specific sense. The CoDa argument against amalgamation is that amalgamation is a non-linear operation in the Aitchison geometry, so things work differently after amalgamation since you are not projecting into a subspace, like when you are dealing with real variables (post CoDaWork17 clarification note by Vera Pawlowsky-Glahn and Juan Jose Egozcue). Amalgamation is related to hierarchical models. For example, in education, you might be interested in the performance of a student in a specific class in time and space, you might want to look at statistics of a class cohort, a school a district or a country. In geochemistry this relates to how you consider and named elements. The challenge is to find the right hierarchical model that helps generate the highest information quality. In fact, this is a moving target since future technology will offer new capabilities. The bottom line is that the data resolution we are using needs to fit the needs of our study, here and now. Data resolution is the first of the eight InfoQ dimensions mentioned above. Amalgamation, as considered in the CoDa community, is different but related to data fusion and data integration, another InfoQ dimensions. When you have separate data sets that you want to analyze with some form of linkage, you can apply different techniques. Such approaches can, for example, handle privacy issues and unsynchronized data stamps. However, again, amalgamation in CoDa is different.

3. Generalization through CoDa transformations

Generalization is another dimension of InfoQ and is a basic property of CoDa. Scale invariance and subcompositional coherence are exactly doing that. At the conceptual level generalization can be provided by stating alternative representations of findings (those typically listed in the discussion section). These alternatives are labeled as having "Meaning Equivalence". In addition, one can list alternatives that look alike but mean something else. These are results not claimed to be reproducible and are labeled as having "Surface Similarity". The boundary between

alternative descriptions with meaning equivalence and alternatives with surface similarity demarks a boundary of meaning (BOM). More on this in the information quality book listed above. Jamies Morton's discussion of ALR and ILR in Set 6 below is doing some of that.

4. The modeling perspective

The note by Justin Silverman of 12/8/17 presents an in depth evaluation of ILR and ALR. Justin presents examples were ALR is shown as "not preserving metric concepts" with a Euclidean distance measured between two points in one ALR basis is not the same as the Euclidean distance measured between two points in another ALR basis. This highlights that any "units" defined in a first ALR transform would not make sense in a second ALR transform. He also expresses difficulties in interpreting ILR and states "I really have no conception of what a geometric mean of a vector with more than 3 parts is". However he shows that different ILR coordinates represent a simple rotation (a special type of projection) that maintains the meaning of our units so that the ILR orthonormal basis coordinates does not have the same problems as the ALR coordinates. Moreover he states: 'By modeling in the ILR I get the benefits of the CLR and the ALR all rolled into one. I can literally treat the ILR transformed data just like it was in real-space with the classical multivariate normal central limit theorem." These different modeling perspectives are very important to consider.

5. CoDa roadblocks

Analysis of data such as proportions, or percentages, can be addressed by the log ratio methodology of compositional data (Aitchison 1986). This is an important issue because features inherent to compositional data, such as scale invariance and the relative scale cause lead statistical analysis of raw compositional data to spurious results ((Pawlowsky-Glahn et al. 2015). Natural principles of compositional data analysis (scale invariance, permutation invariance and subcompositional coherence are followed by the Aitchison geometry on simplex, the sample space of compositions (Egozcue 2009). The three popular CoDa transformations are CLR, centered log ratio coefficients (Aitchison 1986)), ALR, additive log ratio coordinates (Aitchison 1986) and ILR, isometric log ratio coordinates (Egozcue et al. 2003). References to the above are listed in the 2017 excellent paper by Hron et al (Math Geosci, DOI 10.1007/s11004-017-9684-z). Hron et al address the issue of detection limits or imprecision problems of geochemical concentrations and propose orthonormal log ratio coordinates, called *weighted pivot coordinates*, that capture the relevant relative information about an original component and treat the redundant information in a controlled manner. This advance in CoDa methodology comes from practical considerations. These considerations might not be in unison with mathematical considerations. Another topic in the CoDa literature deals with so called "balances". Jamie Morton used this extensively in his paper on microbial niche differentiation. Here again the mathematical and practical considerations appear to be sometimes conflicting.

Set 1:
==========

From: Ron S. Kenett
Sent: 16 August, 2017 5:39 AM
To: 'Justin Silverman' <justin.silverman@duke.edu>; Michael Greenacre <michael.greenacre@gmail.com>
Cc: Alecos Demetriades <alecos.demetriades@gmail.com>; Martin <josepantoni.martin@udg.edu>; Eric Grunsky <egrunsky@gmail.com>; Juan Jose Egozcue <juan.jose.egozcue@upc.edu>; Professor John Bacon-Shone <johnbs@hku.hk>; Vera Pawlowsky <vera.pawlowsky@udg.edu>; Antonella Buccianti <antonella.buccianti@unifi.it>; Blondes, Madalyn <mblondes@usgs.gov>; Alex Washburne <bigalculus@gmail.com>; jamietmorton@gmail.com
Subject: RE: Coda discussion

Justin
I like your meter stick analogy paragraph.
Two sentences in this paragraph stand out:
1. « The power of units is that we can transport them and interpret them in new situations ».
2. « When your units change as a function of another parameter (as in relativistic physics) humans get very confused and lose their ability to intuitively understand the meaning and instead end up just staring at meaningless numbers ».
These seem to be the premises for your comments on ILR, ALR and CLR.

Instead of looking inside the box, let me refer to an approach looking elsewhere.
Transportability, which is a form of what we call “Generalizability” in the context of information quality (InfoQ) dimensions, has been addressed by Judea Pearl using graphical models, more specifically DAGs and Bayesian networks(BN).

If you have a set of data referring to compositional elements, you can build a BN showing how “your units change as a function of another parameter”. In this exercise, you can incorporate expert knowledge of what must be linked (white lists) and what should not (black lists). Amalgamation offers another interesting “twist’ to this. As a first step, graphical models provide what I called “descriptive causality”. As a second step, you can apply counterfactuals and “do calculus” to assess causality relationships. This second step seems however completely outside the CoDa scope.

So, my comments are:

i. Yes we need an intuitive interpretation in order to do applied statistics as opposed to mathematics
ii. Graphical models, together with CoDa methods can help provide this
iii. We need more data driven perspectives. As Terry Speed has said repeatedly : “Practice before theory…”

Thank you for sharing your thoughts – they provide a very useful perspective.
Bst
ron

Set 2:

==========

From: Eric Grunsky [mailto:egrunsky@gmail.com]
Sent: 14 August, 2017 10:44 PM
To: Ron S. Kenett <ron@kpa-group.com>;
Subject: Re: A question of weighting, normalization and variance

Thanks Ron, I think this kind of model building exercise would be useful in the approach that John was taking in his talk and that we have been discussing for some time.
We can look at the variables in terms of uncertainty for each element. This is typically done by the geochemical laboratories prior to the release of the data and is seldom included in the results. If you want, we can obtain data where the uncertainty for each element is available.
Regards, Eric.

From: Ron S. Kenett
Sent: 14 August, 2017 11:05 PM
To: 'Eric Grunsky' <egrunsky@gmail.com>;
Subject: RE: A question of weighting, normalization and variance

Eric

Yes – what you refer to is usually part of the analytic method validation study. In industry one also applies something called Gage Repeatability and Reproducibility (GR&R). One approach is bottom up, from an analysis of the method (something similar to what Michael is driving), the other approach is top down and relies on empirical experiments consisting of repeated measurements by different lab personnel. I do not know if this is well known/trivial/old news. If more details on this are of interest, I can send you more material. This of course relates to the weighted pivots coordinated of Hron et al. (2017).

ron

From: Ron S. Kenett
Sent: 10 August, 2017 9:04 PM
To: 'Michael Greenacre' <michael.greenacre@gmail.com>;
Subject: RE: A question of weighting, normalization and variance

"one can determine weights that equalize the variance contributions of the elements to the total variance" This sounds similar to the idea of oversampling/undersampling in predictive analytics that is to deal with classification of groups with rare events. In that case the weighting is used to improve the predictive performance. It seems that what Michael is considering is undue sensitivity of LRA and his goal is to reduce it. Can this be formalized?

ron

Set 3:

==========
On 16 July 2017 at 23:01, Ron S. Kenett <ron@kpa-group.com> wrote:

Michael

We are used to distinguish Xs from Ys and focus on links between Xs and Ys. The components are considered Xs. You can however study the structure of the Xs. In some sense the boundary gets blurred. Moreover the variance bias tradeoff is seeking to optimize a balance between the two. Hopefully we can discuss these in the future

Bst, ron

From: Michael Greenacre [mailto:michael.greenacre@gmail.com]
Sent: 17 July, 2017 12:08 AM
To: Ron S. Kenett <ron@kpa-group.com>
Subject: Re: Revision of pragmatic paper

Hi again Ron
Your answer makes me think that I'm not explaining myself well.

Perhaps I should not have mentioned the word component, which is a distractor.
But the word "collinearity" suggested the regression context where X's explain a different Y, which is not what I'm doing.

I'm doing something I never did before, I'm looking for some individual Y's to explain all the Y's. The full set of m*(m-1)/2 log-ratios are the Y's, there are no X's in my data.

These are special Y's, loaded with a huge amount of redundancy (exact collinearities).
Then I'm looking for a small subset of the Y's that explain all or almost all of the Y's, to soak up that redundancy in some optimal way, and I'm not trying to predict them.

If I had some explanatory variables X involved, e.g. geography or geological time periods, I could use them to try to explain, or even predict, the geochemical Y's. But I'm not considering predictors here, it's just a simplification of the Y set I'm looking for.

In another situation, I might think of my m*(m-1)/2 log-ratios as candidate X's for another problem, which has its own separate Y's regarded as a response to the compositional data. Then I would have to be careful about this collinearity issue: Y being usually 1-dimensional doesn't give much "room" for looking for explanatory variables that are relatively uncorrelated.
If I can make an analogy with community ecology where I work mostly, the species observed in the oceans are always regarded as responses (Y's), and sometimes there are X's (e.g. environmental variables) that can explain them. There are no exact redundancies in the Y's but I

regularly identify the "most contributing species" to the low-dimensional solution considered the "signal" in the data, to simplify the interpretation for the biologist faced with 200-300 species in the data set..

The compositional data in my view are Y's, they are the products of some processes (which we might have data on, then the problem is more interesting, like having environmental information to explain the fish communities).

It's the same with fatty acids. They are passed on from species to species in the food chain, and they are always regarded as responses (Y's) along the way. If we knew their stomach contents we could try to explain the fatty acid composition of their flesh.
All the best, and sorry if I expressed myself badly, I hope this is clearer (about the topic of "collinearity" raised by Eric)

Michael

From: Ron S. Kenett
Sent: 17 July, 2017 8:41 AM
To: 'Michael Greenacre' <michael.greenacre@gmail.com>
Subject: RE: Revision of pragmatic paper

Michael

With your detailed list below, looks indeed like we have a cacophonic notation. However, what you are doing seems to me more analogous to PCA and variable selection. On a more general level – the simplification you seek is considered by you in absolute terms and derived from domain expertise reasoning. For different goals, different simplifications schemes might apply. What is the implicit goal in the simplification pathway you have travelled? Is the subset of the Ys needed for producing simplified but accurate maps? If you want to study geographical splines, would this work? If you wanted an effective way to compress data, would this be good enough? If you had repeated measurements over time and wanted to study global warming trends, would the same scheme also apply?

ron

Set 4:

==========

From: Ron S. Kenett
Sent: 11 July, 2017 12:08 AM
To: 'Michael Greenacre' <michael.greenacre@gmail.com>
Subject: RE: Revision of pragmatic paper

Michael

The issue with data integration is to combine two different but overlapping data sets (in terms of variables). Our idea on this was to model the data structure, identify links and calibrate the two sets w.r.t the link.

Correspondence analysis might be a neat way to do that. Attached are two papers on this that might motivate you to look into it.

Bst, ron

From: Michael Greenacre [mailto:michael.greenacre@gmail.com]
Sent: 10 July, 2017 11:21 PM
To: Ron S. Kenett <ron@kpa-group.com>
Subject: Re: Revision of pragmatic paper

Very useful paper, thanks Ron. I will certainly have this checklist ready for the next review (actually, one is overdue!)

And I'm glad you think my effort might make the grade, with a little bit of luck!

1. Amalgamation, as considered in the CoDa community, is different but related to data fusion and data integration. Clarifying the similarities and differences would be beneficial and bring in new insights.

Correspondence analysis is a natural environment for aggregations and obeys the principle of distributional equivalence (which regular log-ratio analysis doesn't, but the weighted form does): if two categories (columns, say) have identical distributions across the row categories, then they can be aggregated (and thus their weights simultaneously aggregated too) without affecting the distances between the rows. This is the key principle underlying CA, and the concept is particularly appropriate in linguistic data of counts (Benzécri originally a linguist).

And I'll get to the BN's !!

Best wishes

Michael

Set 5:

==========

From: Ron S. Kenett
Sent: 27 June, 2017 5:41 AM
To: 'Michael Greenacre' <michael.greenacre@gmail.com>
Subject: RE: Fwd: Northern Ireland soil geochemistry - amalgamation

Michael

The way I see this is as follows:

1. The data is geographically located with unique coordinates
2. The locations are classified into AgeBracket classes. This is a one to many mapping since a location can be theoretically classified into more than one class
3. The data points include 46 components that reflect geographically located measurements

To me, the maps shown in Eric's presentation are the key to all this. Ignoring the geography is discounting "neighboring" effects.

To deal with this, I reduced the geographical locations to 5 clusters (instead of the many AgeBracket classes)

It now becomes a problem of predictive analytics where composition data is used to predict cluster membership

The amalgamation issue can now be treated, i.e., how structuring the components affects the predictive performance.

Some of the predictive tools that can be looked at include random forests, K nearest neighbors, neural networks, Bayesian networks, etc…. Seems like this would be a great research roadmap for grad students. Any candidates??

Bst
ron

On 26 June 2017 at 17:27, Ron S. Kenett <ron@kpa-group.com> wrote:

Michael

My comment on this is that it seems that the data used here does not account for the geographical structure. In other words, the coordinates of the measurement sites do not affect the analysis – am I missing something?

ron

Set 6:

==========

From: Jamie Morton [mailto:jamietmorton@gmail.com]
Sent: 18 August, 2017 7:43 PM
To: Ron S. Kenett <ron@kpa-group.com>
Cc: Alex Washburne <bigalculus@gmail.com>; Justin Silverman <justin.silverman@duke.edu>; Michael Greenacre <michael.greenacre@gmail.com>; Alecos Demetriades <alecos.demetriades@gmail.com>; Martin <josepantoni.martin@udg.edu>; Eric Grunsky <egrunsky@gmail.com>; Juan Jose Egozcue <juan.jose.egozcue@upc.edu>; Professor John Bacon-Shone <johnbs@hku.hk>; Vera Pawlowsky <vera.pawlowsky@udg.edu>; Antonella Buccianti <antonella.buccianti@unifi.it>; Blondes, Madalyn <mblondes@usgs.gov>
Subject: Re: Coda discussion

I'll add in my two cents in the discussion.

I think that ALR and ILR both have their practical use cases. We have seen that ALR performs quite well with sample classification in conventional multinomial regression models, and in terms of the graphical models that Michael presented at CoDaWork, ALR does seem to provide nice intuitive explanation (I don't completely grasp the graph partitioning idea that Alex presented, but would definitely be curious to see the details fleshed out).

However, I think there are definitely some questions / scenarios that ALR cannot cope with. Namely, trying to tweak out differentially abundant features between the samples. This is something that I try to highlight in my paper. Identifying exactly which species is changing is a really hard problem -- looking at proportions alone, it is impossible to determine if a species is growing / declining / or even changing between the samples.

There are have been a number of approaches using ALR-like techniques to try infer this with additional assumptions (Sparcc assumes a sparse correlation, and ANCOM assumes few species are changing). However, the crux of the problem lies in the fact that this approaches in the end are ill-define -- you are trying to solve an over-parameterized system of equations (this is acknowledged in the SparCC paper and briefly discussed in my blogpost). And this is not tractable without additional assumptions.

The ILR approach is quite different. Rather than focusing on which exact species are changing, we can relax the problem and instead focus on which partitions are changing. While this is a slightly different question, this is actually a tractable problem.

That being said, there are a ton of really cool things that you can do with the ILR transform. Because the ILR is really just an extended ALR with multiple parts in the numerator and the denominator of the log-ratio, you can represent this transform as a classifier. You can try identify and classify the parts that drive the separation between the groups of interest (this is basically what Javier's talk on optimal balances at CoDaWork was about). And because it forms an orthonormal basis, you can directly map the ILR transform to standard dimensionality reduction techniques (such as PCA, Orthogonal PLS, ...). And what makes this really cool is that at the end of the day, you can basically convert a dimensionality reduction problem in a sorting

problem using the ILR transform. Given a clustering of your parts, you can essentially sort / group the parts followed by the ILR transform to obtain simple latent variables to summarize high-dimensional trends. This is what I what I was aiming for with the hierarchical clustering approach so that we can explicitly test for microbial partitioning due to pH.

In terms of future work -- I think we are just getting started realizing the implications of these techniques. And I'm really excited about the new method advances that have yet to come in the near future.

Best,

Jamie

From: Ron S. Kenett
Sent: 18 August, 2017 8:07 PM
To: 'Jamie Morton' <jamietmorton@gmail.com>
Cc: Alex Washburne <bigalculus@gmail.com>; Justin Silverman <justin.silverman@duke.edu>; Michael Greenacre <michael.greenacre@gmail.com>; Alecos Demetriades <alecos.demetriades@gmail.com>; Martin <josepantoni.martin@udg.edu>; Eric Grunsky <egrunsky@gmail.com>; Juan Jose Egozcue <juan.jose.egozcue@upc.edu>; Professor John Bacon-Shone <johnbs@hku.hk>; Vera Pawlowsky <vera.pawlowsky@udg.edu>; Antonella Buccianti <antonella.buccianti@unifi.it>; Blondes, Madalyn <mblondes@usgs.gov>
Subject: RE: Coda discussion

Jamie

As is now said – really cool. Thank you for sharing this. I also share your enthusiasm.

In terms of your paper. I have suggested an approach to generalize findings. Basically, it involves stating alternative representations of your findings (those listed in the discussion section). These alternatives are labeled to have “Meaning Equivalence”. In addition, the idea is to list alternatives that look alike but mean something else. These are not results you claim to be reproducible. We label these alternatives as having “Surface Similarity”. Actually, in some sense, your ALR, ILR discussion is exactly doing that. Attached is a paper on this.

Bst

ron