Compositional data - an overview

John Bacon-Shone[1] & Eric Grunsky[2]

[1]Social Sciences Research Centre, The University of Hong Kong, Pokfulam Road, Hong Kong, johnbs@hku.hk

[2]Geological Survey of Canada, Natural Resources Canada, 601 Booth St., Ottawa ON K1A 0E8 CANADA, egrunsky@nrcan.gc.ca

Abstract

Compositional data are data where the elements of the composition are non-negative and sum to unity. The key question is what is the appropriate analysis for data from this restricted sample space. We start by summarizing more than a century of progress towards answering this question.

Aitchison(1986) provides a framework appropriate for data that satisfies sub-compositional coherence, i.e., where conclusions about a sub-composition should be the same based on the full composition or the sub-composition alone. However, not all compositional data satisfies this principle and it is helpful to consider the complete cycle of processes that yield any specific dataset and hence the appropriate analysis for data generated in this manner.

Key Words: compositional data analysis, sub-compositional coherence, multivariate data analysis

1. Introduction

Compositional data are data where the elements of the composition are non-negative and sum to unity. While the data can be generated directly (e.g. probabilities), they often arise from non-negative data (such as counts, area, volume, weights, expenditures) that have been scaled by the total of the components. Geometrically, compositional data with D components has a sample space of the regular unit D-simplex

The key question is whether standard multivariate analysis, which assumes that the sample space is R^D, is appropriate for data from this restricted sample space and if not, what is the appropriate analysis? Ironically, most multivariate data are non-negative and hence already have a sample space with a restriction to R^{D+}, making standard multivariate analysis unsuitable.

We first summarize more than a century of progress towards answering this question, drawing heavily on the review papers by (Bacon-Shone 2011) and (Aitchison and Egozcue 2005).

2. History

The starting point for compositional data analysis is the paper of (Pearson 1897), which first identified the problem of "spurious correlation" between ratios of variables, showing that if X,Y and Z are uncorrelated, then X/Z and Y/Z will not be uncorrelated. Pearson then looked at how to adjust the correlations to take into account the "spurious correlation" caused by the scaling. However, this ignores the implicit constraint that scaling only makes sense if the scaling variable is either strictly positive or strictly negative. In short, this approach ignores the range of the data and does not assist in understanding the process by which the data are generated. (Tanner 1949) made the essential point that a log transform of the data may avoid the problem and that checking whether the original or log transformed data follow a Normal distribution may provide some guidance as to whether a transform is needed. (Chayes 1960) later made the explicit connection between Pearson's work and compositional data and showed that some of the correlations between components of the composition must be negative because of the unit sum constraint. However, he

was unable to propose a means to model such data in a way that removed the effect of the constraint.

The first step towards modern compositional data analysis was McAlister (McAlister 1879)'s use of Log-Normal distributions to model data that are constrained to lie in positive Real space. Interestingly, he proposed this as the law of the geometric mean (versus the Normal distribution as the law of the arithmetic mean) and pointed out the lack of practical value for variance of a variable that must be positive, which can be seen in retrospect as recognition of the need for a non-Euclidean metric for data from restricted sample spaces. Instead, he emphasized the meaning of the cumulative distribution. This is by no means the only way to model data on the positive real line and competes with, for example, the gamma and Weibull distributions. It is equivalent to taking a log transform of the data, so that the non-negative constraint is removed, and then assuming a Normal distribution. However, this only addresses the non-negative constraint of compositional data and does not address the unit sum constraint.

3. Logratio Transforms

The simplest meaningful example of a composition is with just two components, so the unit- sum constraint implies that the second component is just one minus the first component, such as probabilities for a binary outcome. (Cox and Snell 1989) use the logit or logistic transformation of the probability in this case, which enables the use of regression models for the logit transformed probabilities. The first public introduction of the properties of the logistic-normal distribution can be found in (Aitchison and Shen 1980). This distribution is written in terms of logratios relative to the last component, so that

$$y(x) = \{\log(x_1/x_d), \ldots \log(x_{D-1}/x_D)\}$$

follows a Multivariate Normal distribution. Unlike the Dirichlet distribution, which has some very restrictive properties, such as complete subcompositional independence, i.e. for each possible partition of the composition, the set of all its subcompositions must be independent, the logistic-normal distribution yields a distribution on the interior of the simplex that does not require these inflexible properties, but instead they become testable linear hypotheses on the covariance matrix within a broad flexible modeling framework. Use of the logistic-normal distribution opens up the full range of linear modeling available for the multivariate Normal distribution in R^D.

4. Subcompositional Dependence

As mentioned above, the logistic–normal distribution has the ability to model useful dependence structures. In his seminal book, (Aitchison 1986) developed this idea, showing that the covariance structure can be modeled in terms of covariances on the log scale and is completely determined by the $D(D-1)/2$ logratio variances

$$\tau ij = \text{var}\{\log(x_i/x_j)\}(i=1,\ldots D-1; j=i+1,..D).$$

However, finding a convenient matrix formulation seems tricky, yielding formulations that either require selecting a specific component as divisor (when using Sigma, which is the logratio covariance matrix for the D-1 log-ratios relative to one component as divisor), have a zero diagonal (when using T, which is the variation matrix for all pairs of logratios) or are singular (when using Gamma, which is the centred logratio covariance matrix). However, it turns out that there are simple linear relationships between these alternative formulations, so it is feasible to choose whichever formulation is simplest to use in any specific context. Indeed, as shown in (Aitchison 1986) and further developed in (Aitchison et al. 2000) linear statistical methods with compositional data as the dependent variable are invariant to the choice of divisor as the implicit linear transformations between different representations cancel out in any F ratio of quadratic or bilinear forms, so this is a conceptual rather than practical problem. One way of avoiding this problem of choosing a divisor is to divide by the geometric mean, known as the clr (centered logratio) transformation. The disadvantage of this is that the centred logratio covariance matrix is singular, making it difficult to use some standard statistical procedures without adaption. However, compositions can be represented by their coordinates in the simplex with a suitable orthonormal basis. This suggests an alternative transformation

known as ilr (isometric logratio transformations) (Egozcue et al. 2003), which avoids the arbitrariness of alr and the singularity of clr. Thus ilr has significant conceptual advantages, but unfortunately, there is no clear "simplest" or canonical basis, unlike R^D. One possibility is to use a sequential binary partition of the components (Egozcue and Pawlowsky-Glahn 2006), known as balances, although this alone still does not ensure uniqueness. Hence, despite the mathematical elegance of this approach, it has practical disadvantages in the relative difficulty of choosing the basis and relating the coordinates back to the original statistical questions.

5. Principles as a starting point for compositional data

Compositional data analysis may appear as a pragmatic approach to avoiding the unit sum constraint, that may have mathematical weaknesses. Indeed, mathematical geologists, typified by (Rehder and Zier 2001) argued that logratio analysis implied an illogical and arbitrary distance metric. In fact, the logratio approach can be derived entirely from a few key principles, which enable the derivation of the entire mathematical framework including an appropriate distance metric on the simplex. As explained in (Aitchison et al. 2000) et al., it should be obvious that compositional data analysis can only make meaningful statements about ratios of components, i.e. the first principle is scale invariance. This should be obvious in that compositional data is unit-free, but some geologists, such as (Watson and Philip 1989), did not find this obvious. The second key principle is subcompositional coherence (Aitchison 1991), which states that inferences about subcompositions should be consistent, regardless of whether the inference is based on the subcomposition or the full composition. For R^D, this would translate into the self-evident principle that inference about a subset of variables should be the same regardless of whether we base the inference on the subset of variables or the full set.

6. Limitations of the log-ratio approach

While the log-ratio approach provides a powerful toolkit for many compositional datasets, it is important to be aware of the limitations of this approach. The two key problems are that firstly, some compositional data has zero components, which are inconsistent with the log-ratio approach and secondly, there may be linear constraints on the simplex, which become non-linear in log-ratio terms. There is a literature on how to avoid the problem of zeroes, but we will instead consider a process modelling approach to identifying an appropriate analysis. Other approaches include (Watson and Philip 1989) and (Stanley 1990) who map compositions onto the positive orthant of the hypersphere and (Butler and Glasbey 2008) who use the Tobit approach of modelling the zero boundaries as the censored probabilities. These approaches can be helpful in that the Butler model can handle compositional data with many zeros, but they also yield their own problems, such as what possible process would generate data on the hypersphere (in the case of Watson) and a model that yields no meaningful inferences for ratios of components (in the case of Butler). Arguably, all statistical models should relate to an underlying process that generated the data we observe, so we now consider this approach.

7. Random processes as a starting point for compositional data

As George Box has famously said, "all models are wrong, but some are useful". I teach my students that to a mathematician, numbers are abstract entities, while to a statistician, numbers always have context – we are trying to understand how they were generated and that requires understanding what sort of process may have generated the data. A good statistical model not only matches the data well, but must also be interpretable.

In practice, we often know quite a lot about how data might have been generated and that knowledge can make a dramatic difference in how precise our inference can be.

Good statisticians understand the importance of understanding the underlying random processes that generate statistical data. For example, we know that the sum of additive random data rapidly converges to a Normal distribution under very weak conditions. However, the mathematical beauty of the Central Limit Theorem often encourages assuming a Normal error process, even when that is logically inconsistent with

the data (e.g. whenever there are constraints or integer data). As the saying goes, when you only have a hammer, everything looks like a nail!

Compositional data is usually the outcome of a number of processes. The most obvious process is closure, which turns non-negative data into a composition. While we often treat this process as implicit in modeling compositional data on the simplex, there are times when explicit inclusion is important.

For example, if our underlying process follows a multivariate log-Normal distribution, it is easy to see that applying closure does not affect the log ratios, which must all follow a Normal distribution, which is equivalent to the standard log-ratio approach and hence the principles must all apply. Of course, if we have a different underlying process distribution, we cannot be sure that this will still hold. Indeed, we may be ignoring important information if we try and model the composition directly, both in terms of the dependence structure of the data and also in terms of having structural zeros, where structural zeros are where we know that a zero component truly means that a component is absent, rather than being simply below the detection limit for that component. Structural zeros can also reflect mixing of different groups (household expenditure with some teetotallers mixed with drinkers), which lead naturally to a probabilistic process determining whether or not a component is present in the composition. For censored zeros, we can often model this approximately taking into account the implicit interval censoring normally involved in the recording process. For example, if we record proportions to three decimal places, then a proportion between 0 and 0.0005 will be recorded as zero. If we embed the interval censoring in our statistical model, we have a mechanism to handle censored zeros ((Leung and Bacon-Shone 2013)). However, if it is important to recognise that detection limits may not be accounted for precisely if we only have the compositional data, as closure removes critical information if the detection limits are with reference to the raw data rather than the composition.

For fixed mixtures of complete compositions, the resultant compositions often look like unitary compositions, but this may not be making best use of all our information. Palmer (Palmer and Douglas 2008) and Tolosana-Delgado (Tolosana-Delgado, von Eynatten and Karius 2011) have looked at end-member mixing, which makes use of some of our knowledge about the mixing process. However, work we have done elsewhere (Grunsky and Bacon-Shone) shows that linear mixing may take place at the level of specific elementary oxides. We may then end up with integer constraints on the compositions when expressed in molar terms, reflecting elemental substitutions. This gives the surprising result of having important linear dependence on the simplex, in addition to the unit constraint, which clearly cannot be accounted for if we insist on the principle of subcompositional coherence for all subcompositions, rather than after accounting for the constraints. This suggests again that understanding the process rather than blind applications of mathematical principles is essential, if we are to produce meaningful inference.

There are also situations (like Aitchison's time budget examples) where amalgamation of some compositional components are key elements of any sensible model. Again, this may mean that any meaningful model should use a combination of amalgamation and log-ratios, which is inconsistent with subcompositional coherence for some subcompositions.

8. Conclusion

Mathematical elegance is beautiful, but understanding how to model a data set has a beauty too! We must ensure that our analysis fits the problem by building models of the process that generated the data, not the other way around.

References

Aitchison, J and SM Shen. 1980. "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika* 67(2):261.

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*, Vol. 416: Chapman and Hall.

Aitchison, J. 1991. "Delusions of Uniqueness and Ineluctability." *Mathematical Geology* 23(2):275-77.

Aitchison, J., C. Barcelo-Vidal, J. A. Martin-Fernandez and V. Pawlowsky-Glahn. 2000. "Logratio Analysis and Compositional Distance." *Mathematical Geology* 32(3):271-75.

Aitchison, J. and J. J. Egozcue. 2005. "Compositional Data Analysis: Where Are We and Where Should We Be Heading?". *Mathematical Geology* 37(7):829-50.

Bacon-Shone, John. 2011. "A Short History of Compositional Data Analysis." Pp. 3-11 in *Compositional Data Analysis: Theory and Applications*: John Wiley and Sons.

Butler, A and C Glasbey. 2008. "A Latent Gaussian Model for Compositional Data with Zeros." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(5):505-20.

Chayes, F. 1960. "On Correlation between Variables of Constant Sum." *Journal of Geophysical Research* 65(12):4185-93.

Cox, DR and EJ Snell. 1989. *Analysis of Binary Data*: Chapman & Hall/CRC.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barcelo-Vidal. 2003. "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology* 35(3):279-300.

Egozcue, JJ and V Pawlowsky-Glahn. 2006. "Simplicial Geometry for Compositional Data." *Geological Society London Special Publications* 264(1):145.

Grunsky, Eric and John Bacon-Shone. 2011. "The Stoichiometry of Mineral Compositions." Paper presented at the CODAWORK 2011, Girona, Spain.

Leung, T.C and John Bacon-Shone. 2013. "Compositional Data Analysis and the Zero Problem: Comparison of Additive and Multiplicative Replacements with Interval Censoring." Paper presented at the CoDaWork 2013, Vorau, Austria.

McAlister, D. 1879. "The Law of the Geometric Mean." *Proceedings of the Royal Society of London* 29:367-76.

Palmer, M. J. and G. B. Douglas. 2008. "A Bayesian Statistical Model for End Member Analysis of Sediment Geochemistry, Incorporating Spatial Dependences." *Journal of the Royal Statistical Society. Series C: Applied Statistics* 57(3):313-27.

Pearson, K. 1897. "On a Form of Spurious Correlation Which May Arise When Indices Are Used, Etc." *Proceedings of the Royal Society* 60:489-98.

Rehder, S and U Zier. 2001. "Letter to the Editor: Comment on 'Logratio Analysis and Compositional Distance' by J. Aitchison, C. Barcelo-Vidal, Ja Martìn-Fernandez, and V. Pawlowsky-Glahn." *Mathematical Geology* 33(7):845-48.

Stanley, Clifford R. 1990. "Descriptive Statistics For n-Dimensional Closed Arrays: A Spherical Coordinate Approach." *Mathematical Geology* 22(8):933-56.

Tanner, JM. 1949. "Fallacy of Per-Weight and Per-Surface Area Standards, and Their Relation to Spurious Correlation." *Journal of Applied Physiology* 2(1):1.

Tolosana-Delgado, R, H von Eynatten and V Karius. 2011. "Constructing Modal Mineralogy from Geochemical Composition: A Geometric-Bayesian Approach." *Computers & Geosciences* 37(5):677-91.

Watson, DF and GM Philip. 1989. "Measures of Variability for Geological Data." *Mathematical Geology* 21(2):233-54.