Part I

INTRODUCTION

1

A short history of compositional data analysis

John Bacon-Shone
Social Sciences Research Centre, The University of Hong Kong, Hong Kong

1.1 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, $\mathcal{S}^D$. The key question is whether standard multivariate analysis, which assumes that the sample space is $\mathbb{R}^D$, 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 $\mathbb{R}^D_+$. This chapter tries to summarize more than a century of progress towards answering this question and draws heavily on the review paper by Aitchison and Egozcue (2005).

1.2 Spurious correlation

The starting point for compositional data analysis is arguably the paper of Pearson (1897), which first identified the problem of '*spurious correlation*' between ratios of variables. It is easy to show 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*

Compositional Data Analysis: Theory and Applications, First Edition. Edited by Vera Pawlowsky-Glahn and Antonella Buccianti.

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.

1.3 Log and log-ratio transforms

The first step towards modern compositional data analysis was arguably the use by McAlister (1879) 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 different metric for data from restricted sample spaces, that takes constraints into account. 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. One of the key texts for the Log-Normal distribution is the book by Aitchison and Brown (1969). However, this only addresses the non-negative constraint of compositional data and does not address the unit sum constraint.

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. This is just the situation that arises with 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. However, it appears that nobody saw the potential for a similar approach for the more general case of compositional data until the first known reference to using the log-ratio transform to solve the constraint problem for compositional (or simplicial) data by Obenchain in a personal communication to Johnson and Kotz (Kotz *et al.* 2000). Indeed, Obenchain contributed to the discussion of the Royal Statistical Society paper by Aitchison (1982), where he stated that he became discouraged by the problem of zero components and thus never attempted to publish his simplex work, even though he had derived many properties of the logistic-normal distribution.

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 log-ratios relative to the last component, so that $\mathbf{y}(\mathbf{x}) = \{\log(x_1/x_D), \ldots, \log(x_{D-1}/x_D)\}$ follows a Multivariate Normal distribution.

Up to that time, the only known tractable distribution on the simplex was the Dirichlet distribution. However, the Dirichlet distribution 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. This makes it impossible to model any reasonable dependence structure for compositional data using the Dirichlet distribution. In contrast, 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 modelling framework. In addition, the Aitchison and Shen (1980) paper showed that the logistic-normal distribution is close to any Dirichlet distribution in terms of the Kullback–Leibler divergence. Later Aitchison (1985) derived a more general distribution that contains both the Dirichlet and logistic-normal distributions, although the potential for using this distribution for testing Dirichlet against logistic-normal distributions within the same class is diminished as these hypotheses are on the boundary of the parameter space. More recently, the generalization of the logistic-normal distribution to the additive logistic skew-normal distribution on the simplex (Mateu-Figueras *et al.* 2005) applies the skew-normal distribution (Azzalini 2005) to log-ratios on the simplex and offers the useful possibility of modelling data where the distribution of $\mathbf{y}(\mathbf{x})$ is not symmetrical. Use of the logistic-normal distribution opens up the full range of linear modelling available for the multivariate Normal distribution in $\mathbb{R}^D$.

1.4 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 modelled in terms of covariances on the log scale and is completely determined by the $D(D-1)/2$ log-ratio variances $\tau_{ij} = \mathrm{Var}\{\log(x_i/x_j)\}$(where $i = 1, \ldots, D-1; j = i+1, \ldots, D$).

However, finding a convenient matrix formulation seems tricky, either yielding formulations that require selecting a specific component as divisor [when using $\mathbf{\Sigma}$, which is the log-ratio covariance matrix for the $D-1$ log-ratios relative to one component as divisor (Aitchison 1986, p. 77)], have a zero diagonal [when using $\mathbf{T}$, which is the variation matrix for all pairs of log-ratios (Aitchison 1986, p. 76)] or are singular [when using $\mathbf{\Gamma}$, which is the centred log-ratio covariance matrix (Aitchison 1986, p. 79)]. 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.

1.5 alr, clr, ilr: which transformation to choose?

One key question for using the log-ratio transformations is choosing the divisor. Most of the literature initially used an arbitrary component as the divisor, known as using alr (additive log-ratio) transformation. This is potentially problematic because the distances between points in the transformed space are not the same for different divisors. However, 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 (centred log-ratio) transformation. As noted above, the disadvantage of this is that the clr covariance matrix is singular, making it difficult to use in some standard statistical procedures without adaption.

A key step forward was recognition that compositions can be represented by their coordinates in the simplex with a suitable orthonormal basis. This suggests an alternative transformation, known as ilr (isometric log-ratio) 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 $\mathbb{R}^D$. One possibility is to use a sequential binary partition of the components (Egozcue and Pawlowsky-Glahn 2005), known as balances, although this alone still does not ensure uniqueness. This approach is explained in detail in Chapter 3. However, despite the mathematical elegance of this approach, it has practical disadvantages in the relative difficulty of choosing the basis when that is not motivated by the statistical question being investigated and also when relating the coordinates back to the original statistical question.

1.6 Principles, perturbations and back to the simplex

At this point, the reader may have concluded that compositional data analysis is entirely 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 log-ratio analysis implied an illogical and arbitrary distance metric. In fact, the log-ratio 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), 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 1992), which states that inferences about subcompositions should be consistent, regardless of whether the inference is based on the subcomposition or the full composition. For $\mathbb{R}^D$, 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. Any meaningful metric for the simplex should satisfy these two principles and the Euclidean metric for $\mathbb{R}^D$ clearly does not satisfy either for compositional data. Aitchison (1986) introduced the idea of perturbation as the analogue to linear operations in $\mathbb{R}^D$, which was further developed in Aitchison and Ng (2005). A perturbation $\mathbf{p} = (p_1, \ldots, p_D)$ is a differential scaling operator that when applied to the composition $\mathbf{x} = (x_1, \ldots, x_D)$ yields the composition

$$\mathbf{X} = \mathbf{p} \oplus \mathbf{x} = \mathcal{C}(p_1 x_1, \ldots, p_D x_D)\,,$$

where $\mathcal{C}$ is the closure operator that scales elements to ensure that we remain in the unit simplex.

The set of perturbations (if restricted to $\mathcal{S}^D$) form a group with an inverse and an identity perturbation $\mathbf{e} = (1/D, \ldots, 1/D)$. As any composition can be expressed as a result of a perturbation on any other composition, the distance between any two compositions must be

expressible in terms of perturbations. Perturbation clearly corresponds to addition in $\mathbb{R}^D$ and we can define powering to correspond to multiplication in $\mathbb{R}^D$ as

$$\mathbf{X} = \mathbf{a} \odot \mathbf{x} = \mathcal{C}(x_1^a, \ldots, x_D^a)\,.$$

The simplicial metric, or Aitchison distance is then given by

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{i=1}^{D} \left[\log \frac{x_i}{\mathrm{g_m}(\mathbf{x})} - \log \frac{y_i}{\mathrm{g_m}(\mathbf{y})} \right]^2 \right\}^{\frac{1}{2}},$$

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the components, which can be shown to satisfy all the usual metric axioms and to depend only on perturbation distance. It is also easy to show that this metric satisfies the two key principles mentioned above.

The centre for a compositional distribution is then

$$\mathrm{Cen}[\mathbf{x}] = \mathcal{C}(\exp\{\mathrm{E}[\log(\mathbf{x})]\}),$$

with the variation matrix, $\mathbf{T}$, as the most convenient measure of variability. In summary, this allows us to transfer the analysis back to the simplex, without the asymmetry of using alr.

1.7 Biplots and singular value decompositions

It is essential to have simple ways to summarize and display multivariate data sets. Fortunately, singular value decompositions and the related graphical tool of the biplot (Gabriel 1971), provide precisely the tools we need for compositional data when adapted to the simplex (Aitchison and Greenacre 2002). The biplot for the simplex is based on a singular value decomposition of the row and column centred log-ratio matrix. It enables us to graphically display which combinations of the log-ratios contain large and small amounts of variability. The former provides a useful simplification of the major contributions to total compositional variability, while the latter identifies any likely linear dependencies amongst the log-ratios.

1.8 Mixtures

One important application of compositional data is as the covariate that determines a mixture. This yields log-contrast models for experiments with fixed mixtures where the dependence is only on the composition (Aitchison and Bacon-Shone 1984).

Compositional data can also occur doubly as the mixture of compositions. In this case, the mixed composition does not stay within the class of logistic-normal distributions, but can often be approximated well by a logistic-normal distribution as shown in Aitchison and Bacon-Shone (1999).

One specific problem where the mixture of compositions occurs is what geologists call the end-member problem (Renner 1993; Weltje 1997). In this case, the key question is which of the end members (of usually known compositions) are being mixed to form the

outcome composition. A full Bayesian analysis of the end-member problem including spatial dependence is found in Palmer and Douglas (2008).

1.9 Discrete compositions

In the discussion of the Royal Statistical Society paper by Aitchison (1982), R.L. Plackett raised the question about how best to model discrete compositions and whether this might provide a solution to the problem of zeros (see below). The first full analysis of discrete data using compositional data models can be found in Billheimer *et al.* (2001), who use the logistic-normal distribution to model the probabilities for a multinomial distribution to allow much more sophisticated modelling of the occurrence data for species. This can be seen as a more sophisticated approach to the multivariate count modelling of Aitchison and Ho (1989), who use a logistic-normal model for the log means of Poisson data. However, the resultant data in both cases were counts, rather than discrete compositions, although in Billheimer's case, it is the relative occurrence that is of interest. As shown in Bacon-Shone (2008), it may be helpful to model discrete compositions even without knowing the total counts, helped by the knowledge that the original counts are non-negative integers.

1.10 Compositional processes

One of the nice consequences of analysing in the simplex, is that it is easy to investigate compositional processes as in Thomas and Aitchison (2005), who examine dependence on time through

$$D\mathbf{x}(t) = \mathcal{C}\left\{\exp\left[\frac{\mathrm{d}}{\mathrm{d}t}\log\mathbf{x}(t)\right]\right\},$$

where $\mathrm{d}/\mathrm{d}t$ denotes differentiation with respect to time.

This approach allows easy investigation and modelling of any processes that can be parameterized, including the possibility of change-points (Bacon-Shone 2011).

1.11 Structural, counting and rounded zeros

Aitchison recognized from the start that there was a need to solve the problem of zeros in compositional data as the log-ratio is undefined in this case. He wrote a much earlier paper (Aitchison 1955) which looked at the related problem of zeros for non-negative data, which presents a similar problem when using Log-Normal, Gamma or Weibull distributions, all of which have zero probability of observing zero. This paper used a conditional approach that separates the zeros from the continuous distribution and was applied to household expenditure data, which is a compositional data problem when classified into different categories of expenditure.

The original approach to compositional zeros in Aitchison (1982) was to simply replace all zeros by a small positive amount less than the detection limit, with the closure operator applied to apply the unit sum constraint and then sensitivity analysis to check the impact of

the replacement value. However, this positive replacement approach potentially distorts the compositional data.

Proper 'in the simplex' approaches were independently proposed by Martín-Fernández *et al.* (2000) and Fry *et al.* (2000). Palarea-Albaladejo *et al.* (2007) used a parametric model to handle the zeros as missing data using the expectation-maximization (EM) algorithm.

As pointed out by Aitchison and Kay (2003) and by Bacon-Shone (2003), zeros can occur for at least three distinct reasons. First, there may be a structural reason why the component must be zero, such as alcohol expenditure components for household expenditure data in families that do not drink alcohol. This situation is best modelled by the conditional approach. Secondly, there may be a zero because of an underlying discrete process (Bacon-Shone 2008), such as expenditure on white goods (i.e. major household appliances) in household expenditure data, where people may go several years between making purchases and may miss capture in the data collection process. This situation is best modelled by modelling the underlying discrete process. Lastly, there may be a limit in the measurement or recording processes, such that very small components are recorded as zero. For this situation, the approaches of Martín-Fernández *et al.* (2000) and Fry *et al.* (2000) mentioned above seem most relevant.

Recently Butler and Glasbey (2008) proposed another modelling approach to compositional data with zeros, using Euclidean projections onto the simplex, with the probability 'outside' the simplex used to model the point probabilities on the boundaries. Unfortunately, this approach does not have a special case of log-ratio analysis and fails the test of the two key principles mentioned above.

A more comprehensive discussion of how to handle zeros in compositions can be found in Chapter 4.

1.12 Conclusion

This brief summary shows how much progress has been made in the last century in finding appropriate analyses for compositional data, much of it in the last 30 years, relying heavily on the insights of John Aitchison.

Acknowledgement

This research has been partially supported by the Research Grants Council, Hong Kong (Grant HKU 700303).

References

Aitchison J 1955 On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* **50**(271), 901–908.

Aitchison J 1982 The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **44**(2), 139–177.

Aitchison J 1985 A general class of distributions on the simplex. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **47**(1), 136–146.

Aitchison J 1986 *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by The Blackburn Press), London (UK).

Aitchison J 1992 On criteria for measures of compositional difference. *Mathematical Geology* **24**(4), 365–379.

Aitchison J and Bacon-Shone J 1984 Log contrast models for experiments with mixtures. *Biometrika* **71**(2), 323–330.

Aitchison J and Bacon-Shone J 1999 Log contrast models for experiments with mixtures. *Biometrika* **86**(2), 351–364.

Aitchison J and Brown JAC 1969 *The Lognormal Distribution with Special Reference to its Uses in Econometrics.* Department of Applied Economics Monograph: 5. Cambridge University Press, Cambridge (UK). 176 p.

Aitchison J and Egozcue JJ 2005 Compositional data analysis: where are we and where should we be heading?. *Mathematical Geology* **37**(7), 829–850.

Aitchison J and Greenacre M 2002 Biplots for compositional data. *Applied Statistics* **51**(4), 375–392.

Aitchison J and Ho C 1989 The multivariate Poisson-log normal distribution. *Biometrika* **76**(4), 643–653.

Aitchison J and Kay J 2003 Possible solution of some essential zero problems in compositional data analysis. In *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop* (ed. Thió-Henestrosa S and Martín-Fernández JA). http://ima.udg.es/Activitats/CoDaWork03/. University of Girona, Girona (Spain). CD-ROM.

Aitchison J and Ng K 2005 The role of perturbation in compositional data analysis. *Statistical Modelling* **5**(2), 173–185.

Aitchison J and Shen SM 1980 Logistic-normal distributions. Some properties and uses. *Biometrika* **67**(2), 261–272.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA and Pawlowsky-Glahn V 2000 Logratio analysis and compositional distance. *Mathematical Geology* **32**(3), 271–275.

Azzalini A 2005 The skew normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**(2), 159–188.

Bacon-Shone J 2003 Modelling structural zeros in compositional data. In *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop* (ed. Thió-Henestrosa S and Martín-Fernández JA). http://ima.udg.es/Activitats/CoDaWork03/. University of Girona, Girona (Spain). CD-ROM.

Bacon-Shone J 2008 Discrete and continuous compositions. In *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop* (ed. Daunis-i Estadella J and Martín-Fernández J), p. http://hdl.handle.net/10256/723. University of Girona, Girona (Spain). 11 p.

Bacon-Shone J 2011 Mixing of compositions at points and along lines. *Computers & Geosciences* **37**(5), 692–695.

Billheimer D, Guttorp P and Fagan W 2001 Statistical interpretation of species composition. *Journal of the American Statistical Association* **96**(456), 1205–1214.

Butler A and Glasbey C 2008 A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **57**(5), 505–520.

Chayes F 1960 On correlation between variables of constant sum. *Journal of Geophysical Research* **65**(12), 4185–4193.

Cox D and Snell E 1989 *Analysis of Binary Data*, 2nd edition. Chapman and Hall/CRC, London (UK). p. 236.

Egozcue JJ and Pawlowsky-Glahn V 2005 Groups of parts and their balances in compositional data analysis. *Mathematical Geology* **37**(7), 795–828.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G and Barceló-Vidal C 2003 Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35**(3), 279–300.

Fry JM, Fry TRL and McLaren KR 2000 Compositional data analysis and zeros in micro data.. *Applied Economics* **32**(8), 953–959.

Gabriel KR 1971 The biplot – graphic display of matrices with application to principal component analysis. *Biometrika* **58**(3), 453–467.

Kotz S, Balakrishnan N and Johnson NL 2000 *Continuous Multivariate Distributions. Volume I, Models and Applications*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, NY (USA). 730 p.

Martín-Fernández JA, Barceló-Vidal C and Pawlowsky-Glahn V 2000 Zero replacement in compositional data sets. In *Studies in Classification, Data Analysis, and Knowledge Organization. Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000)* (ed. Kiers H, Rasson J, Groenen P and Shader M). Springer-Verlag, Berlin (Germany) pp. 155–160.

Mateu-Figueras G, Pawlowsky-Glahn V and Barceló-Vidal C 2005 The additive logistic skew-normal distribution on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* **19**(3), 205–214.

McAlister D 1879 The law of the geometric mean. *Proceedings of the Royal Society of London* **29**, 367–376.

Palarea-Albaladejo J, Martín-Fernández JA and Gómez-García JA 2007 Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology* **39**(7), 625–645.

Palmer MJ and Douglas GB 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), 313–327.

Pearson K 1897 Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **LX**, 489–502.

Rehder S and Zier U 2001 Letter to the Editor: Comment on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J.A. Martín-Fernández and V. Pawlowsky-Glahn. *Mathematical Geology* **33**(7), 845–848.

Renner RM 1993 The resolution of a compositional data set into mixtures of fixed source components. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **42**(4), 615–631.

Tanner J 1949 Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *Journal of Applied Physiology* **2**(1), 1–15.

Thomas CW and Aitchison J 2005 Compositional data analysis of geological variability and process: a case study. *Mathematical Geology* **37**(7), 753–772.

Watson DF and Philip GM 1989 Measures of variability for geological data. *Mathematical Geology* **21**(2), 233–254.

Weltje JG 1997 End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology* **29**(4), 503–549.