AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2022 American Psychological Association
ISSN: 1082-989X

2024, Vol. 29, No. 2, 362–378
https://doi.org/10.1037/met0000464

Compositional Data Analysis Tutorial

Michael Smithson[1] and Stephen B. Broomell[2]
[1] Research School of Psychology, The Australian National University
[2] Department of Social and Decision Sciences, Carnegie-Mellon University

Abstract

This article presents techniques for dealing with a form of dependency in data arising when numerical data sum to a constant for individual cases, that is, "compositional" or "ipsative" data. Examples are percentages that sum to 100, and hours in a day that sum to 24. Ipsative scales fell out of fashion in psychology during the 1960s and 1970s due to a lack of methods for analyzing them. However, ipsative scales have merits, and compositional data commonly occur in psychological research. Moreover, as we demonstrate, sometimes converting data to a compositional form yields insights not otherwise accessible. Fortunately, there are sound methods for analyzing compositional data. We seek to enable researchers to analyze compositional data by presenting appropriate techniques and illustrating their application to real data. First, we elaborate the technical details of compositional data and discuss both established and new approaches to their analysis. We then present applications of these methods to real social science datasets (data and code using R are available in a supplementary document). We conclude with a discussion of the state of the art in compositional data analysis and remaining unsolved problems. A brief guide to available software resources is provided in the first section of the supplementary document.

Translational Abstract

Psychological researchers sometimes must deal with numerical data that has a constant sum for each case in the sample. For instance, the amounts of time out of a 24-hr day that a person devotes to sleep, eating, work, recreation, and all other activities must sum to 24 hr. Likewise, the percentages of a person's income allocated to food, rent, clothing, transportation, all other expenses, and savings must sum to 100%. These are known as "compositional data" in some disciplines, and traditionally as "ipsative data" in psychology. Researchers in psychology during the past several decades have had difficulties in analyzing compositional data because of the constant-sum requirement, and as a result, tended to avoid this kind of data. Fortunately, straightforward techniques for analyzing compositional data have been developed since the 1980s and software resources are available for them. We elaborate these techniques and demonstrate their application to real data. We also discuss the state of the art in compositional data analysis, including unsolved problems and new approaches. This article has two goals: enabling researchers to analyze compositional data, and persuading them that analyzing data from a compositional standpoint can be useful.

Keywords: compositional data, ipsative data, log-ratio, beta regression, copula

Supplemental materials: https://doi.org/10.1037/met0000464.supp

Dependencies among observations and measurements play a large role in determining what statistical analyses should be used to obtain accurate inferences in the social sciences. For example, hierarchical modeling techniques appropriately handle dependencies between observations derived from clustered sampling techniques (e.g., sampling students clustered by classrooms that are clustered by schools). Without accounting for these dependencies, inference can easily go astray, for example, through misestimation of standard errors.

This article discusses techniques for dealing with a form of dependency in data arising when numerical variables sum to a constant for individual cases. Continuous variables of this kind are commonplace. Examples include percentages that sum to 100, probabilities that sum to 1, and hours in a day that sum to 24. The categories or bins across which the data sum to a constant are called *parts*, and the collection of categories is called a *composition*. Such data are referred to as *compositional* data.

The dependency stems from the fact that increasing the value of one part requires taking value away from at least one of the others. The analysis of this type of compositional data has received relatively little attention in the social sciences, despite the fact that it is often encountered in the research questions we ask and the data we collect. We therefore provide an introduction and tutorial for analyzing compositional data, with a focus on regression models that treat compositional variables as dependent variables.

Compositional data were first encountered in psychology with the collection of what was called "ipsative" data. Ipsative measures were

This article was published Online First January 31, 2022.

Michael Smithson https://orcid.org/0000-0003-4455-2192

We thank Michael Platow, Yiyun Shou, Nicole Tan, and Guangyu Zhu for their helpful questions and comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Michael Smithson, Research School of Psychology, The Australian National University, Canberra, A.C.T. 2601, Australia. Email: Michael.Smithson@anu.edu.au

employed in a wide variety of psychological testing and measurement contexts, ranging from personality (e.g., Gordon, 1951) to attitudes and preferences (e.g., Sisson, 1948). Early taxonomies of psychological measurement methods included variants of ipsative scales, as in Cattell's (1944) typology. Rationales for using ipsative scales were similar to those for using forced-choice items rather than Likert-scaled items. Ipsative ratings have been claimed to be more objective (Sisson, 1948), to yield greater differentiation among ratings, and to diminish response biases such as halo effects, faking good or impression management (Bowen et al., 2002), and social desirability (Cunningham et al., 1977).

At the time, the dependencies due to the compositional nature of ipsative data could not be correctly analyzed, and the recognition of a lack of appropriate statistical approaches killed the practice of collecting such data in the 1960s–1970s. Clemans (1966) and Hicks (1970), among others, declared that ipsative data could not be mathematically transformed to nonipsative forms, and therefore ipsative measurement should be avoided where possible. A thorough and fairly recent review of ipsative data measurement and applications in psychology is available (van Eijnatten et al., 2015), so we do not repeat that here.

Current work in the social sciences nevertheless still involves compositional data. Sometimes data are collected without realizing their compositional nature (hereafter, unavoidably compositional). In other circumstances, the data are not strictly compositional, but new insights could be gained by reframing them as a composition (optionally compositional).

In some cases, social science researchers are already collecting unavoidably compositional data, but are focused on analyzing only one of the components (which can also make their compositional nature less salient). However, analyzing a single component does not completely remove issues of compositional dependency, and overlooks aspects of these data that can only be revealed by analyzing the entire composition. For example, research in judgment and decision making under uncertainty elicits intervals from participants using a fixed response scale (e.g., Budescu et al., 2014; Juslin et al., 2007; Soll & Klayman, 2004). The focus of these projects is on the elicited interval, which is typically analyzed by representing the interval by its midpoint and range, which are each independently analyzed. However, when intervals are collected using a fixed response scale, the areas of the scale that fall below, within, and above the interval sum to a constant. This generates a dependency such that intervals with midpoints toward either edge of the scale must be narrow (otherwise the midpoint could not be close to the edge), but intervals with midpoints in the middle of the scale are not constrained in this way (and as a result, are more likely to be wider). As we demonstrate later with this example, intervals can be analyzed directly by treating them as a partition of the response scale, and their width and location understood by jointly analyzing the partitions of the response scale generated by the intervals as a composition.

In other cases, researchers may wish to recast their data as compositions in order to draw different insights that could not be obtained in any other way. Specifically, any variable that involves summing variables that share a metric, where the sum is unrestricted, can be optionally treated compositionally. For example, research in hospital emergency department patient processing involves measuring patients' lengths-of-stay. Such data is typically analyzed using conventional regression models to understand factors that lengthen or shorten length of stay. However, the length of stay itself can be treated as an exposure variable that generates proportions of the length-of-stay consumed by different activities of interest (e.g., triage, nursing care, and time spent with doctors). As we demonstrate later on, the compositional form of these data can be used to evaluate whether the proportion of time spent with doctors is at the expense of other uses of time, such as the proportion of time spent in nursing care.

We seek to enable researchers to engage in compositional data analysis, opening the door for collecting new forms of data and enabling researchers to analyze their data from a different perspective that may better suit their research goals. In the following, we first introduce the technical details of compositional data analysis and discuss two approaches to analyzing compositional data. Along the way, we discuss their assumptions, comparative strengths, and weaknesses. We then present applications of these methods to real social science data-sets. The data-sets and code are available as supplementary materials to this article (https://osf.io/9kndf/). We conclude the article with a discussion of available resources for compositional data analysis and remaining unsolved problems on this topic.

Compositional Data Live in the Simplex

Compositional data analysis was first introduced by Lewi (1976) for the study of biology and later expanded and popularized by Aitchison (1982) for the study of geology. This approach first defines the geometry in which compositional data live and then outlines methods for transforming compositional data to a Euclidean geometry for analysis and inference.

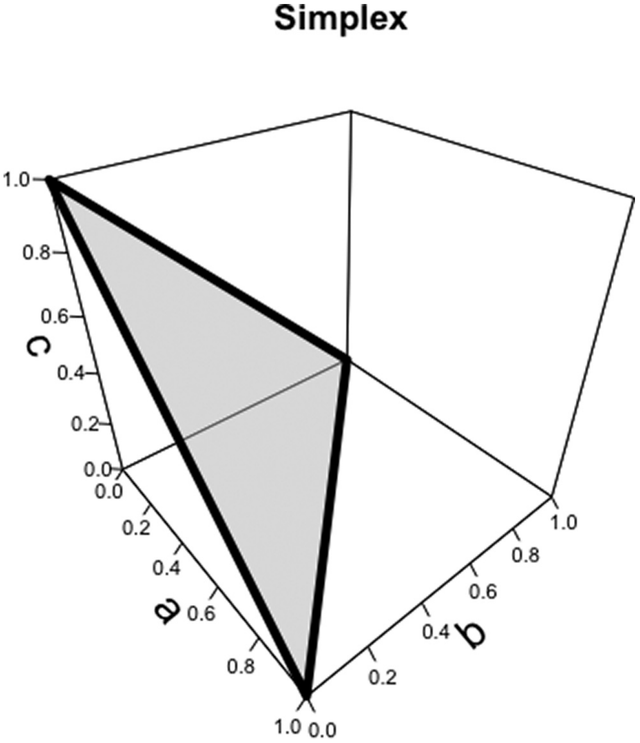
We define a composition based on the amount of each part that is present. Assume we have three parts of a whole represented by the set $\{a, b, c\}$ such that

$$a > 0; b > 0; c > 0; \text{ and } a + b + c = d.$$

If we allow a, b, and c to take on any value greater than 0 while constraining them to sum the constant d, we get the mathematical definition for a simplex. Dividing all parts of the composition by their sum, d, transforms any simplex into a standardized simplex where all parts are constrained to sum to 1. For example, the hours spent in a day dedicated to sleeping, working, and relaxing (assuming all activities belong to one of these categories) forms a composition that resides on the simplex defined by the sum constraint of 24 hr. The representation of these data is the same on a standardized simplex obtained by the closure operation that divides all parts by their sum, making the composition sum to 1. In the following, we will work in the standardized simplex.

Figure 1 displays the standardized simplex as the subspace that is defined by a surface where the coordinate axes a, b, and c are constrained to sum to 1, taking on a shape that is triangular. A composition of three numbers $\{a, b, c\}$ is a single point on this triangular surface. Plots of compositions in the simplex can easily be produced in R (using the compositions package), and are referred to as ternary plots (we present some examples below). It is easy to visualize the simplex based on two and three parts, but the simplex can have higher dimensions for compositions with an arbitrary number of parts K and—while our approach to data analysis can remain the same—compositional data with many parts are difficult to visualize in the simplex.

Figure 1
Three-Part Composition Surface in the Simplex



Note. The gray triangular surface represents the simplex based on three parts. Compositions can only exist on this gray surface because of their constraint to have a common sum across the parts.

Within the geometry of the simplex, operations such as addition and multiplication take on a different meaning. Two compositions can be added together, but the sum of two compositions results in perturbing the relative value of each part of the composition. Compositions with K parts are represented as vectors $\mathbf{x} = \{x_1, x_2, \ldots, x_K\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_K\}$, and perturbation of composition $\mathbf{x}$ with composition $\mathbf{y}$ is performed as

$$\mathbf{x} + \mathbf{y} = \left\{\left\{\frac{x_1 * y_1}{\sum_{k=1}^{K} x_k * y_k}, \frac{x_2 * y_2}{\sum_{k=1}^{K} x_k * y_k}, \ldots, \frac{x_K * y_K}{\sum_{k=1}^{K} x_k * y_k}\right\}\right\}.$$

Perturbation moves the value associated with one part of the composition to another relative to the magnitude of the product between the individual parts of each composition. Therefore, the composition $\mathbf{x}$ with $x_k = 1/K$ for all k is the additive identity in compositional geometry, meaning adding this uniform composition to any composition $\mathbf{y}$ yields $\mathbf{y}$.

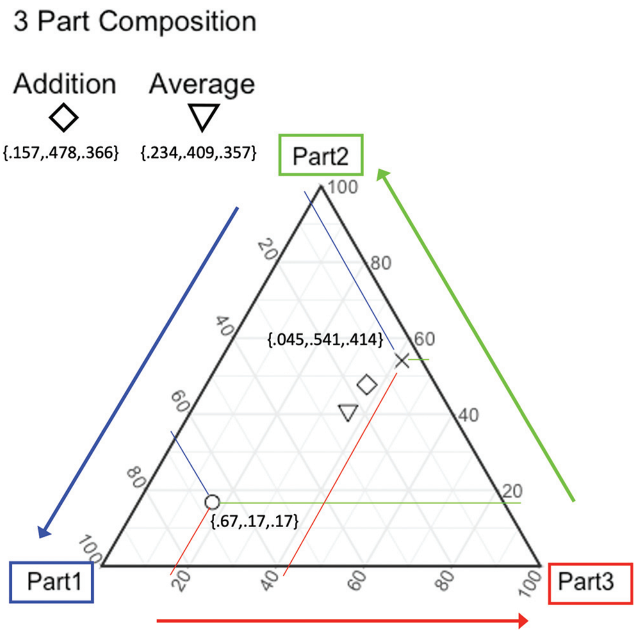
Adding a composition to itself represents a multiple of a composition. We can therefore multiply a composition by a scalar, but this product results in raising each part of the composition to the power of the scalar. Given composition $\mathbf{x}$ and scalar a, the product is performed as

$$\mathbf{x} * a = \left\{\left\{\frac{x_1^a}{\sum_{k=1}^{K} x_k^a}, \frac{x_2^a}{\sum_{k=1}^{K} x_k^a}, \ldots, \frac{x_K^a}{\sum_{k=1}^{K} x_k^a}\right\}\right\}.$$

Raising the parts to a power also moves the value associated with one part to another, with the shift based on the magnitude of each part. When $a > 1$, the product shifts value so that the difference between parts gets larger: larger parts get larger and smaller parts get smaller. When $a < 1$, the product shifts value so the difference between parts gets smaller: smaller parts get larger and larger parts get smaller. The multiplicative identity is the scalar $a = 1$ or the uniform composition where $x_k = 1/K$ for all k.

To illustrate these operations, we will use some data from a study consisting of recorded information on patients admitted to the emergency department of the University of Alberta Hospital between midnight January 23, 1999 and midnight January 29, 1999 (Yoon et al., 2003). The study focused on the length of time required for patients to be processed, and here we will consider time-interval proportions for three stages: registration-triage, nursing assessment, and physician-disposition. Suppose that a patient spent 4.5% of their time in registration-triage, 54.1% of their time under nursing assessment, and finally 41.4% of their time with a physician. Figure 2 displays this composition as an "x" in the simplex.

Figure 2
Example of Compositional Addition and Average



Note. Each composition represents proportion of time spent in three stages of emergency care. Hypothetical patient marked with an "x" can be transformed to units of "tasks" using addition (marked with a diamond) or averaged with another hypothetical patient (marked with a triangle). The mean and average are computed on the same two compositions to show the difference between these two operations. See the online article for the color version of this figure.

Suppose there are 12 tasks in the registration-triage stage, but only three tasks in the nursing-assessment stage and three tasks in the physician-disposition stage. An efficiency auditor wants to know what the patient's time composition is when it is converted to time per task within each of the stages. The number of tasks composition is {2/3, 1/6, 1/6} and so we can obtain this conversion by adding the composition {.045, .541, .414} to get the composition {.157, .478, .366}. Figure 2 displays the result of this sum as the diamond that falls in-between the two compositions being added together.

As a result of these geometric definitions for addition and multiplication within the simplex, the uniform composition represents the barycenter of the simplex. This point is the center of mass, and represents a lack of change for addition and multiplication. Additionally, the mean of a set of compositions is computed as the geometric mean of each part across the set, normed to respect the sum constraint. For instance, suppose we want to know the mean composition of two patient time-intervals in the three stages, and the first patient's composition is {.045, .541, .414} while the second patient's composition is {.67, .17, .17}. The geometric means of the three parts are {.173, .300, .263} and their sum is .736, so the normed mean is {.173, .300, .263}/.736 = {.235, .409, .357}. Figure 2 displays this mean as the triangle that falls in-between the two compositions being averaged while being closer to the barycenter than their sum (the diamond). The geometric mean is related to taking the arithmetic mean of logarithms, leading to an easy way of interpreting the log of compositional data, which is discussed in the next section.

Understanding the geometry of the simplex allows for meaningful manipulations and summaries of compositional data. However, working with compositional data within the simplex has limitations. Visualization becomes difficult for compositions with more than three parts and multivariate analysis is intractable. Compositional data analysis therefore requires transformations from the simplex to Euclidean space (and optionally transforming back to the simplex). The next two sections present such transformations.

Log-Ratio Transform Method

Suppose we have a composition consisting of K parts, with a sample size of N. As before, we will assume that $0 \leq y_{ki} \leq 1$, for $k = 1, \ldots, K$ and $i = 1, \ldots, N$, such that for each i they sum to 1 across the k, so these y_{ki} are on the simplex. The log-ratio transform method (Aitchison, 1982) maps data from the simplex to an unrestricted Euclidean space by taking the logarithm of the ratio of pairs of them, that is, $\lambda_{jki} = \log(y_{ji}/y_{ki})$. Log-ratios transform the ratios to the real line because they can have any value on the real line. When $y_{ji}/y_{ki} > 1$ then $\lambda_{jki} > 0$, whereas when $y_{ji}/y_{ki} < 1$ then $\lambda_{jki} < 0$. Moreover, when the compositional nature of the data is incorporated in the log-ratio transforms, they have an inverse transformation to recover the original probabilities.

Additive Log-Ratios

There are several ways of forming the log-ratio transform. The simplest is the "additive" log-ratio (ALR) transformation, for which each of the compositional parts is compared against a "referent" part (e.g., the K^{th}):

$$\lambda_{kK} = \log(y_k/y_K) = \log(y_k) - \log(y_K) \tag{1}$$

The ALR is analogous to the use of dummy variables for dealing with categorical variables in linear regression, log-linear models, and logistic regression via base-group coding or what also is known as a reference category. Here, the K^{th} part of the composition is acting as the base group or reference against which all of the others are being compared. The λ_{kK} can be thought of as the log-odds of being in the k^{th} part relative to being in the K^{th} part, as in loglinear models and logistic regression. It is worth bearing in mind that when the data are counts, compositional analysis becomes identical to multinomial logistic regression.

To recover the original y_k, for $k < K$ we put

$$y_k = \exp(\lambda_{kK})/\left(1 + \sum\nolimits_{j=1}^{K-1} \exp(\lambda_{jK})\right), \tag{2}$$

and taking into account the sum-constraint,

$$y_K = 1/\left(1 + \sum\nolimits_{j=1}^{K-1} \exp(\lambda_{jK})\right). \tag{3}$$

To illustrate what we have so far, we revisit the emergency department of the University of Alberta Hospital data discussed in the previous section (Yoon et al., 2003) where a patient spent 4.5% of their time in registration-triage, 54.1% of their time under nursing assessment, and 41.4% of their time with a physician. If we use the physician-disposition stage as the referent part of the composition, the log-ratios are $\lambda_{13} = \log(.045/.414) = -2.219$ and $\lambda_{23} = \log(.541/.414) = 0.268$. The first log-ratio is negative because much less time was spent in registration-triage than with a physician, whereas the second log-ratio is positive because more time was spent under nursing assessment than in consultation with the physician.

Now, given $\lambda_{13} = -2.219$ and $\lambda_{23} = .268$, we can reproduce the original three time-proportions by using Equations 2 and 3. First, we have

$$\begin{aligned} y_1 &= \exp(\lambda_{13})/(1 + \exp(\lambda_{13}) + \exp(\lambda_{23})) \\ &= \exp(-2.219)/(1 + \exp(-2.219) + \exp(0.268)) = 0.045. \end{aligned}$$

Next,

$$\begin{aligned} y_2 &= \exp(\lambda_{23})/(1 + \exp(\lambda_{13}) + \exp(\lambda_{23})) \\ &= \exp(0.268)/(1 + \exp(-2.219) + \exp(0.268)) = 0.541. \end{aligned}$$

Finally,

$$\begin{aligned} y_3 &= 1/(1 + \exp(\lambda_{13}) + \exp(\lambda_{23})) \\ &= 1/(1 + \exp(-2.219) + \exp(0.268)) = 0.414. \end{aligned}$$

Log-ratio transformations also can be applied to "subcompositions" and "amalgamations." A subcomposition is a subset of the parts in a composition, treated as though they are a composition in their own right. For example, we might want to consider the subcomposition consisting of the last two stages in the emergency patient process. The original data in our example are .541 of the total time under nursing assessment and .414 of the time with a physician. Converting these into a subcomposition requires rescaling them by dividing by their total, so that we now have y_1 = .541/(.541 + .414) = .566 and y_2 = .414/(.541 + .414) = .434. Log-ratios are "subcomposition-coherent" because they are identical in both the subcomposition and the original composition: $\log(0.566/0.434) = \log(0.541/0.414) = 0.268$.

Amalgamations are combinations of compositional parts. For example, given our three-part compositional structure representing the three different stages in processing a patient, we might want to compare the combination of the first two stages with the final one, because ideally the patient should be spending most of their time in consultation with a physician, that is, in the third stage. In our example, the resulting amalgamation would be a two-part composition with y_1 = .045 + .541 = .586 and y_2 = .414. The resulting log-ratio would be log(.586/.414) = .347.

The ALR transform's λ_{kK} are considered as continuous random variables on the real line, and therefore analyzable with conventional statistical methods for such variables, such as linear regression. To develop an understanding of this approach, let us consider emergency patients at two different triage levels. Table 1 shows the average proportions of time spent by patients in each of the three stages at different triage levels. We can see that patients at Level 5 spend proportionately more time in the first two stages and less in the third stage than patients at Level 4, but how would the ALR approach capture this?

Using the physician-disposition stage as the referent part of the composition, the ALR transform for the Triage 4 patients yields the log-ratios

$$\lambda_{13(4)} = \log(0.170/0.641) = -1.324$$

and

$$\lambda_{23(4)} = \log(0.189/0.641) = -1.221.$$

For the Triage 5 patients, the corresponding log-ratios are

$$\lambda_{13(5)} = \log(0.229/0.543) = -0.861$$

and

$$\lambda_{23(5)} = \log(0.228/0.543) = -0.866.$$

The complete set of log-ratios is displayed in Table 2.

Table 1
Patients in Triage Levels

Stage	Triage 1−2	Triage 3	Triage 4	Triage 5
Registration triage	0.027	0.062	0.170	0.229
Nursing assessment	0.025	0.093	0.189	0.228
Physician-disposition	0.948	0.845	0.641	0.543
Arithmetic mean	0.333	0.333	0.333	0.333
Geometric mean	0.086	0.170	0.274	0.305

Table 2
Log-Ratios for Triage Levels

Level	λ_{12}	λ_{13}	λ_{23}
Triage 1−2	0.086	−3.564	−3.650
Triage 3	−0.410	−2.612	−2.202
Triage 4	−0.103	−1.324	−1.221
Triage 5	0.005	−0.861	−0.866

Taking the differences between the pairs of log-ratios gives us the gains for each of the first two stages from Triage 4 to Triage 5 patients:

$$\lambda_{13(5)} - \lambda_{13(4)} = -0.861 + 1.324 = 0.463$$

and

$$\lambda_{23(5)} - \lambda_{23(4)} = -0.866 + 1.221 = 0.354.$$

Because we can think of these differences as differences between log-odds, we also can exponentiate them to get an effect-size equivalent to odds-ratios. Thus, the patients at Triage 5 level have exp(.463) = 1.588 times greater "odds" than Triage 4 level patients of spending time in the registration stage relative to the physician stage. Likewise, Triage 5 patients have exp(.354) = 1.525 times greater "odds" than Triage 4 patients of spending time in the nursing-assessment stage relative to the physician stage.

Log-Ratio Distances

The ALR linear regression approach illustrated in the preceding subsection may be suited to assessing the effect of a variable on specific parts of a composition, but what if we want a summary statistic that measures the overall difference between two compositions? We can analyze Euclidean distances between compositions in log-ratio space. Assessing distances can complement regression analyses.

The Euclidean distance between two K-part compositions is

$$D_{ij} = \frac{1}{K}\sqrt{\sum\sum\nolimits_{k<k'}\left[\log\frac{y_{ki}}{y_{k'i}} - \log\frac{y_{kj}}{y_{k'j}}\right]^2}. \tag{4}$$

Here, i and j refer to different compositions. Despite the convention of dividing by K, the D_{ij} are unbounded. Nevertheless, they obey the laws of Euclidean distances and can be useful for interpreting the relationships among compositions.

For example, consider the log-ratio distances among the four compositions whose log-ratios are displayed in Table 2. At first glance it might appear that the Triage 1–2 and Triage 3 compositions are closer together than the Triage 3 and Triage 4 compositions, but the Triage 1–2 to Triage 3 log-ratio distance is .601 whereas the Triage 3 to Triage 4 log-ratio distance is .549. From Equation 4 we have

$$D_{13} = \frac{1}{3}\sqrt{(0.086+0.410)^2 + (-3.564+2.612)^2 + (-3.650+2.202)^2} = 0.601$$

and

$$D_{34} = \frac{1}{3}\sqrt{(-0.410+0.103)^2 + (-2.612+1.324)^2 + (-2.202+1.221)^2} = 0.594.$$

Perhaps even less obvious is the fact that the Triage 4 to Triage 5 distance, .198, is considerably less than either of those distances.

Log-Ratio Variances and Covariances

How can we analyze the dependencies among the composition parts? Given that log-ratio transforms produce a collection of variables in Euclidean space, it also can produce a variance-covariance matrix and from that we may compute correlations. It turns out that log-ratio covariances can be analyzed solely in terms of log-ratio variances, so a matrix of log-ratio variances contains the same information as a correlation matrix. This may seem a rather counterintuitive result, so we provide a brief derivation of it here.

We shall denote the variance of the log-ratios by $\tau_{jk} = \text{var}(\lambda_{jk})$, and the covariance of two log-ratios by $\sigma_{jk|lm} = \text{cov}(\lambda_{jk}, \lambda_{lm})$. The following relationships are going to be needed to get to our final result. These all follow from the properties of logarithms of ratios.

$$\lambda_{jk} = -\lambda_{kj}, \tag{5}$$

$$\tau_{jk} = \tau_{kj}, \tag{6}$$

and

$$\lambda_{jl} + \lambda_{lk} = \lambda_{jk}. \tag{7}$$

The next equation follows from the preceding three equations, but requires a bit of algebra to show this:

$$\sigma_{jk|lm} = \sigma_{jl|lm} - \sigma_{kl|lm}. \tag{8}$$

We start by writing the full formula for the difference between covariances on the right-hand side of Equation 8:

$$\sigma_{jl|lm} - \sigma_{kl|lm} = \frac{\sum_{i=1}^{N}\left(\lambda_{jli} - \overline{\lambda}_{jl}\right)\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right)}{N-1} - \frac{\sum_{i=1}^{N}\left(\lambda_{kli} - \overline{\lambda}_{kl}\right)\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right)}{N-1}.$$

Rearranging the terms in the sums via distributivity gives us

$$\begin{aligned}&\left(\lambda_{jli} - \overline{\lambda}_{jl}\right)\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right) - \left(\lambda_{kli} - \overline{\lambda}_{kl}\right)\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right) \\ &= \left[\left(\lambda_{jli} - \overline{\lambda}_{jl}\right) - \left(\lambda_{kli} - \overline{\lambda}_{kl}\right)\right]\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right),\end{aligned}$$

and from Equations 5 and 7 we can rewrite the expression in square brackets as

$$\begin{aligned}\left(\lambda_{jli} - \overline{\lambda}_{jl}\right) - \left(\lambda_{kli} - \overline{\lambda}_{kl}\right) &= \left(\lambda_{jli} - \overline{\lambda}_{jl}\right) + \left(\lambda_{lki} - \overline{\lambda}_{lk}\right) \\ &= \lambda_{jki} - \overline{\lambda}_{jk.}\end{aligned}$$

Substituting the right-most expression for the square-brackets term, and then inserting the result into the numerator sums above, we get the result in Equation 8,

$$\sigma_{jl|lm} - \sigma_{kl|lm} = \frac{\sum_{i=1}^{N}\left(\lambda_{jki} - \overline{\lambda}_{jk}\right)\left(\lambda_{lmi} - \overline{\lambda}_{lm}\right)}{N-1} = \sigma_{jk|lm}.$$

The final piece of the puzzle comes from figuring out what $\sigma_{jl|lm}$ and $\sigma_{kl|lm}$ consist of. Consider the variance of $\lambda_{jl} - \lambda_{ml}$ (this actually is not a digression). From Equations 6 and 7 we know that $\lambda_{jl} + \lambda_{lm} = \lambda_{jl} - \lambda_{ml} = \lambda_{jm}$, so clearly the variance of the latter difference must be τ_{jm}. But we also know that the variance of the difference between two random variables is the sum of their variances minus twice their covariance, so we have $\tau_{jm} = \tau_{jl} + \tau_{lm} - 2\sigma_{jl|ml} = \tau_{jl} + \tau_{lm} + 2\sigma_{jl|lm}$, and rearranging this gives us

$$\sigma_{jl|lm} = \frac{1}{2}\left(\tau_{jm} - \tau_{jl} - \tau_{lm}\right). \tag{9}$$

An identical argument yields

$$\sigma_{kl|lm} = \frac{1}{2}\left(\tau_{km} - \tau_{kl} - \tau_{lm}\right). \tag{10}$$

Substituting Equations 9 and 10 into Equation 8 gives a general formula for the covariance of any two log-ratios in terms of log-ratio variances:

$$\begin{aligned}\sigma_{jk|lm} = \sigma_{jl|lm} - \sigma_{kl|lm} &= \frac{1}{2}\left(\tau_{jm} + \tau_{jl} - \tau_{lm}\right) - \frac{1}{2}\left(\tau_{km} - \tau_{kl} - \tau_{lm}\right) \\ &= \frac{1}{2}\left(\tau_{jm} + \tau_{kl} - \tau_{jl} - \tau_{km}\right).\end{aligned} \tag{11}$$

Usually we are not interested in the entire set of variances and covariances, but instead the subset that is relevant to a particular log-ratio transform. For the ALR, the log-ratios of interest clearly include the λ_{kK}, but inspection of Equation 11 reveals that others need to be included as well. For any j and l not equal to K, Equation 11 yields

$$\sigma_{jK|lK} = \frac{1}{2}\left(\tau_{jK} + \tau_{Kl} - \tau_{jl} - \tau_{KK}\right) = \frac{1}{2}\left(\tau_{jK} + \tau_{Kl} - \tau_{jl}\right).$$

Thus, any relevant covariance $\sigma_{jK|lK}$ in an ALR is positive if $\tau_{jl} < \tau_{Kl} + \tau_{jK}$ and negative if $\tau_{jl} > \tau_{Kl} + \tau_{jK}$.

We will use the patient data in Table 2 as an illustration. Table 3 shows the correlation and variance-covariance matrices for these data.

Table 3
Correlation and Variance-Covariance Matrices for Patient Data

Correlation matrix	λ_{12}	λ_{13}	λ_{23}
λ_{12}	1		
λ_{13}	0.019	1	
λ_{23}	−0.156	0.985	1
Variance-covariance matrix			
	λ_{12}	λ_{13}	λ_{23}
λ_{12}	0.047		
λ_{13}	0.005	1.514	
λ_{23}	−0.042	1.509	1.552

In the correlation matrix we can see that the log-ratios λ_{13} and λ_{23} are strongly correlated (.985). This reflects the fact that the amounts of time spent in the registration and nursing stages relative to the physician stage are very similar for each of the four patients. Given this similarity, we should also expect that the variance of the λ_{12} log-ratios should be small because the ratios y_{1i}/y_{2i} generally will be close to 1 (they are 1.090, .663, .902, 1.005), and indeed the relevant variance in Table 3 is .047. This connection between the correlation and variances of the log-ratios gives us an intuitive clue to why the formula in Equation 11 works.

From Equation 11 the covariance for the log-ratios λ_{13} and λ_{23} is

$$\sigma_{13|23} = (\tau_{13} + \tau_{23} - \tau_{12} - \tau_{33})/2.$$

Now, $\tau_{33} = 0$ because $y_{3i}/y_{3i} = 1$, so the formula reduces to

$$\sigma_{13|23} = (\tau_{13} + \tau_{23} - \tau_{12})/2 = (1.514 + 1.552 - 0.047)/2 \\ = 1.509.$$

Alternative Log-Ratios

The ALR has some limitations. Chief among these is the requirement to choose one of the composition parts as the "base" against which to compare the other parts. In our example this works reasonably well because we would like to know how much time emergency department patients are spending in the first two stages relative to the third (consultation with physician) stage. However, in many situations the choice of a base part would be arbitrary.

As mentioned earlier, the additive log-ratio transform is analogous to base-group coding for dummy variables in regression or ANOVA. A popular alternative in the compositional data literature therefore is analogous to effects-coding for dummy variables, and it is the symmetrical or "centered" log-ratio transformation (CLR):

$$\lambda_{ki} = \log \frac{y_{ki}}{\left(\prod_{j=1}^{K} y_{ji}\right)^{1/K}} = \log(y_{ki}) - \frac{1}{K}\sum_{j=1}^{K} \log(y_{ji}) \qquad (12)$$

This transform compares each part of the composition with its geometric mean, which in the log scale is just the deviation of each $\log(y_{ki})$ from their arithmetic mean. Note that this does not permit all parts in the composition to be included in a regression, because they are perfectly collinear (i.e., they sum to 0). This is akin to the sum-restrictions on coefficients in log-linear models and multinomial logistic regression.

Table 1 displays the arithmetic and geometric means of the composition parts for the four different triage levels. The arithmetic means are identical but the geometric means differ, thereby demonstrating the need to be cautious about assuming that these two types of means will behave similarly. The geometric mean is sensitive to dispersion as well as location in compositional data, just as the mean for a log-normal distribution in the original scale contains a term for the variance.

From Equation 12 and the data in Table 1 we can compute the centered log-ratios displayed in Table 4. For instance,

$$\lambda_{1c1} = \log(0.027) \\ - (1/3)\left[\log(0.027) + \log(0.025) + \log(0.948)\right] \\ = -1.159.$$

CLR transforms are useful for addressing the same kinds of questions as effects-coding in linear regression or ANOVA. The trends in Table 4 are similar to those in the ALR, that is, increased relative amounts of time spent in the two earlier stages of the patient process and therefore less spent in the third stage as we move from Triage Level 1–2 to Level 5, but here these proportions are evaluated relative to their geometric means, which also are increasing with triage level.

A third type of transform is the "isometric" log-ratio transformation (ILR). The ILR transform uses an orthonormal basis in *K-1*-dimensional Euclidean space to form a basis for the compositions in the corresponding *K*-dimensional simplex. The application of an orthonormal basis has utility for researchers interested in applying techniques such as factor analysis to compositional data. There are many ILR transforms, and we present only one popular example here.

Our example treats the composition parts as an ordered list and compares each part to the geometric mean of the remaining parts further down the list. Returning to our emergency-treatment stages example, suppose we are interested in modeling this composition as if it is a sequence of "hurdles." First, what is the proportion of time spent at the first (registration-triage) stage, relative to the proportions in the other stages? Next, what proportion of time is spent in the second stage relative to the proportion of time in the final stage? The transform required to do this uses what are known as "pivot" log-ratios:

Table 4
Centered Log-Ratios for Triage Levels

Level	λ_{1c}	λ_{2c}	λ_{3c}
Triage 1–2	−1.159	−1.245	2.405
Triage 3	−1.008	−0.597	1.605
Triage 4	−0.476	−0.373	0.848
Triage 5	−0.285	−0.291	0.576

$$\lambda_{ki} = \sqrt{\frac{K-k}{K-k+1}}\log\frac{y_{ki}}{\left(\prod_{j=k+1}^{K} y_{ji}\right)^{1/(K-k)}} = \sqrt{\frac{K-k}{K-k+1}}\left(\log(y_{ki}) - \frac{1}{K-k}\sum_{j=k+1}^{K}\log(y_{ji})\right) \tag{13}$$

for $j = 1, \ldots, K - 1$.

Applying this transform to our example data, we may write our two log-ratios as $\lambda_1 = \sqrt{2/3}\log\left(y_1/\sqrt{y_2 y_3}\right)$ and $\lambda_2 = \sqrt{1/2}\log(y_2/y_3)$. These are the log-ratios in the first two columns of Table 5. These show that the proportion of time spent in the first (registration) stage relative to the geometric mean of the other two stages increases as we move up in Triage levels, and they also reprise our earlier observation that the relative amount of time in the second stage relative to the third stage also increases with Triage level.

For an amalgamation model closely related to this ILR transform, we can examine the log-ratios $\lambda_1 = \log(y_1/(y_2 + y_3))$ and $\lambda_2 = \log(y_2/y_3)$. These log-ratios are displayed in the third and fourth columns of Table 5. These are simpler than their ILR counterparts, and some authors (Greenacre, 2018) argue that this kind of model is easier to interpret than an ILR. In this example both transforms agree with one another about the effects of moving up triage levels, but we will see an example later on where the ILR and amalgamation approaches disagree. We will compare these approaches in greater depth later in this article.

How should researchers choose among the various kinds of log-ratio transformations? The choice depends mainly on the questions that the researcher wishes to answer and the purposes served by the analyses. The ALR is well-suited to situations where the researcher wants to use one part of the composition as a "base" so that the other parts' proportions are considered relative to the base part. The CLR considers parts of the composition relative to their geometric mean, so it is suitable for settings in which the parts are being compared with each other by identifying their positions relative to an appropriate measure of central tendency. ILRs are more specialized, in the sense that they can be custom-built for specific purposes by choosing an appropriate basis. Interpretability is an important pragmatic consideration. ALRs generally are the most easily interpreted among the three types of log-ratio transforms, and ILRs often are the most difficult (Greenacre, 2018).

Probability-Ratio Method

Smithson (2019) has presented an alternative to the log-ratio approach to modeling compositions, which uses probability-ratios instead of odds-ratios. The two approaches are strongly connected. Every log-ratio model has a corresponding probability-ratio model. To see this, we set $\nu_{ki} = \exp(\lambda_{ki})$, where λ represents a log-ratio, as before. Then the corresponding probability-ratio is

$$\gamma_{jk} = y_j/(y_j + y_k) = \nu_{jk}/(\nu_{jk} + 1) \tag{14}$$

Table 5
ILR and Amalgamation Log-Ratios for Triage Levels

Level	ILR	Log-ratios	Stick-b.	Log-ratios
	λ_1	λ_2	λ_1	λ_2
Triage 1–2	−0.617	−1.121	−1.559	−1.585
Triage 3	−0.536	−0.676	−1.180	−0.956
Triage 4	−0.253	−0.375	−0.687	−0.530
Triage 5	−0.152	−0.266	−0.526	−0.376

Note. ILR = isometric log-ratio.

The marginal distributions for the probability-ratios corresponding to the Gaussian marginals in the log-ratio approach are the logit-normal distribution, and we will demonstrate later that these yield equivalent results in regression models to those obtained in the log-ratio framework. To illustrate the connection between the two frameworks, Table 6 displays the probability-ratios that correspond to the log-ratios in Table 2. For instance, the Triage 1–2 entry for γ_{12} may be derived from its counterpart in Table 2 via Equation 14:

$$\nu_{12} = \exp(0.086) = 1.090, \text{ and}$$
$$\gamma_{12} = \nu_1/(\nu_{12} + 1) = 1.090/(1.090 + 1) = 0.521.$$

Because they are monotonically related, probability-ratios and log-ratios often tell the same story. Nonetheless, probability-ratios can be easier to interpret than their log-ratio counterparts. Consider the λ_{13} and λ_{23} columns in Table 2 versus the γ_{13} and γ_{23} columns in Table 6. Both sets of columns are assessing the relative portion of time spent in the first two stages of patient care against the third (physician consultation) stage, and both show that this relative portion increases with triage level. However, the γ_{jk} entries have a more direct interpretation; they are the portion of time in each earlier stage out of the sum of it and the third stage.

What are the uses and advantages of the probability-ratio framework? Its uses are twofold. First, it can be a valuable supplement to log-ratio analyses by providing an alternative, but entirely compatible, viewpoint of the data and analyses. Second, it can expand the toolbox of compositional data analysis techniques in ways that go beyond log-ratio methods but still are compatible with them.

A major advantage of the probability-ratio method is that any distribution whose support is the unit interval can be used to model the marginal distributions. The most popular distribution of this kind is the beta distribution (Ferrari & Cribari-Neto, 2004; Smithson & Verkuilen, 2006). Others include the Kumaraswamy distribution and the cdf-quantile family (Smithson & Shou, 2017). All of these distributions are more flexible than the logit-normal distribution and considerably augment the compositional analysis toolbox.

In this article we will provide examples that employ beta distributions. Readers familiar with beta regression may have wondered why the log-ratio compositional analysis literature makes no mention of them other than in connection with the Dirichlet regression approach, which log-ratio advocates regard as inferior. We do not deal with the Dirichlet approach here, but nevertheless the probability-ratio framework connects beta regression with compositional data analysis more flexibly and straightforwardly than the Dirichlet approach, and in a way that is compatible with and extends the log-ratio framework.

Moreover, as we will show later, modeling heteroscedasticity in the probability-ratio approach is more straightforward than in the log-ratio approach. A typical probability-ratio regression model

Table 6
Probability-Ratios for Triage Levels

Level	γ_{12}	γ_{13}	γ_{23}
Triage 1–2	0.521	0.028	0.025
Triage 3	0.399	0.068	0.100
Triage 4	0.474	0.210	0.228
Triage 5	0.501	0.297	0.296

has two submodels: One for location (e.g., means) and another for dispersion. While heteroscedastic linear regression methods have existed for some time, they have been neglected in psychological research. For doubly bounded random variables, however (and therefore for compositional data), modeling dispersion has assumed a more prominent role because location and dispersion are not independent for such variables. Therefore, as shown in the beta regression literature (Smithson & Verkuilen, 2006), ignoring dispersion can result in a misspecified model for location.

Because probability-ratios are in the unit hyper-cube, distances between them in a $K - 1$ dimensional hyper-cube are bounded between 0 and $\sqrt{K-1}$. Thus, unlike log-ratio distances, these can be normed to sit between 0 and 1 by dividing by $\sqrt{K-1}$.

One drawback to the probability-ratio framework is that the dependencies between probability-ratios do not have the neat properties enjoyed by their counterparts in the log-ratio framework. Nonetheless, conventional methods for modeling these dependencies are available for probability-ratios, such as multilevel models (Verkuilen & Smithson, 2012) and copulas (Shou & Smithson, 2019).

Copulas are a very flexible method for dealing with dependencies. It is beyond the scope of this article to provide a detailed account of them but accessible introductions are available (Nelsen, 2006; Yan, 2007). Copulas are multivariate cumulative distribution functions (cdfs) with uniform marginal distributions. A continuous multivariate cdf potentially can be modeled with a copula by converting the marginal distributions to uniform distributions by applying their quantile functions. This strips away the variables' marginal distributions, thereby leaving a remaining dependency structure that is no longer contaminated by artifact from the original marginals. This dependency structure can then be fitted to a copula.

Because copulas enable the dependency structure of a multivariate distribution to be modeled separately from the structures of the marginal distributions, a multivariate model using copulas can have different marginal distributions for each variable (e.g., a bivariate model with a beta distribution for one variable and a logit-normal distribution for the other). Moreover, copulas capture monotonic relationships between variables so they are not limited to linear relationships. The result is considerably greater flexibility than the conventional multivariate normal regression model.

The copula model can be estimated in two stages. The marginal distributions are modeled and their parameter estimates are fed to their respective quantile functions, whose output then yields a multivariate distribution with uniform marginals. A copula model then is estimated using this multivariate distribution as input. An example will be presented later that includes a copula model, but we do not focus on the copula models here and relegate details of that part of the analysis to the supplement.

The Zeros Problem

When a composition part contains no mass, or all of the mass, this results in zeros and/or ones in the data set. None of the compositional analysis methods discussed so far is able to deal directly with zeros or ones in the data. Log-ratios are undefined if the ratios have zeros in either of the arguments, and even when probability-ratios are 0 or 1 the distributions available for modeling probability-ratios do not have defined densities at those values because their support is on (0,1).

Four ways of dealing with zeroes have been proposed throughout the literature on compositional analysis, and choosing among them appears to be a matter of judgment. First, occasionally it may be feasible to create amalgamations that combine different parts of the composition so that zeros are absorbed. This approach is not feasible if there are ones in the data or if the zeros are dispersed across too many composition parts. Instead, it is viable under conditions similar to those which suit collapsing categories on a nominal variable, that is, when one or more categories have consistently sparse data.

The next three ways of dealing with zeros and ones hinge on a distinction between what the compositional data literature calls "essential" and "rounded" zeros (Martín-Fernández et al., 2003). This distinction is not the same as the distinction between "structured" and "observed" zeros in the log-linear and logistic regression literature. Both essential and rounded zeros are observed zeros. Essential zeros are regarded as true scores, whereas rounded zeros are censored scores in the sense that they have been recorded as zeros but their real value falls above zero within some criterion threshold. For instance, if we are analyzing the proportions of waking time that people spend in various activities and one of these is playing sports, a typical adult sample will include a substantial number of zeros in the "sports" part of the activities composition. These are essential zeros, and it is plausible that the variables predicting these zeros (who does not play any sport) will not be the same as the variables predicting nonzero values (the proportion of time sports players spend on sport). This example suggests that in some settings, essential zeros may be qualitatively different from the nonzeros.

The distinction between essential and rounded zeros is not always clear and may require researchers to make a judgment call. For example, patients arriving at the emergency department with immediately life-threatening conditions are rushed through the registration-triage and nursing-assessment stages and straight on to treatment. The amounts of time for them in the first two stages, then, are recorded as zeros. On the one hand, all patients do need to be registered and assessed, so it could be argued that these are rounded zeros. On the other hand, one could argue that immediately life-threatening conditions are qualitatively different from other patient conditions and these patients are treated differently from the others, so the zeros could be regarded as essential zeros. Later on we compare these two ways of treating zeros in the emergency department data.

The compositional data literature considers two ways of dealing with rounded zeros: rescaling and replacement. The rescaling method shifts the zeros and ones away from 0 and 1 via an appropriate shrinkage transform that is applied to both the zero and nonzero data. This is a tradition widely practiced in other techniques dealing with proportions, such as categorical data analysis and beta regression (e.g., Verkuilen & Smithson, 2012) and recommended by the pioneer of compositional data analysis (Aitchison, 1986). After transforming, the composition is "reclosed," that is, its transformed values are divided by their sum so that the resulting values sum to 1. Researchers opting for this approach will want to perform sensitivity analysis to ensure that minor alterations of the transformation do not destabilize their models.

The replacement method for handling rounded zeros, related to the rescaling approach described above, is to replace the zeros with a small positive value. Two versions of this method have been described in the literature. In the "simple replacement" version, the composition is reclosed after the replacements have been done, whereas in the "multiplicative" version the nonzero parts are multiplied by $1 - J\epsilon$, where ϵ is the value replacing each zero and J is the number of zeros being replaced (Martín-Fernández et al., 2003).

The main advantage that both replacement methods have over the rescaling method is that the ratios of the original nonzero values are preserved. That is, $y_j/y_k = y'_j/y'_k$, where y denotes the original values of the nonzero composition parts and y' denotes their values after substitution and reclosing. When the y are probabilities, this property is known as the "reverse Bayesian" principle, which holds that the ratios between probabilities of states should not change if new states are added to the sample space (Karni & Vierø, 2013). An additional advantage enjoyed by the multiplicative replacement method is that the covariance structure of the nonzero parts is preserved, which is not true of the simple replacement method (Martín-Fernández et al., 2003).

The final method of dealing with zeros and ones is a hurdle model. Hurdle models for linear regression were introduced by Cragg (1971). A hurdle model treats zeros (and ones) as essential or true scores, and potentially qualitatively distinct from values in the (0,1) interval. The zeros are represented by a binary variable recording in each case whether each part has a zero or nonzero value, and the ones are represented by another binary variable recording whether there is a one or not. These binary variables are then modeled by separate logistic regressions. The values in (0,1) are then modeled via log-ratios or probability-ratios.

For clarity, we briefly describe a hurdle model for the log-ratio framework here. An example with more detail, using the emergency department data, is presented later in this article. Suppose there are zeros in the compositional data. Then the hurdle model of the distribution density is

$$f\left(\lambda_{jk}, \mu, \sigma, \pi_0\right) = \begin{cases} \pi_0, \text{ for } y_{jk} = 0 \\ \phi_t\left(\lambda_{jk}, \mu, \sigma, \pi_0\right) \text{ for } y_{jk} \neq 0 \end{cases} \tag{15}$$

where $\phi_t\left(\lambda_{jk}, \mu, \sigma, \pi_0\right)$ is the normal density function truncated from below at the cdf value $\Phi\left(\lambda_{jk}, \mu, \sigma\right) = \pi_0$. If the data also contain ones, then the hurdle model requires separate logistic regressions to estimate the conditional probabilities of zeros and ones, π_0 and π_1, respectively. The normal distribution for the remaining cases is truncated below by $\Phi\left(\lambda_{jk}, \mu, \sigma\right) = \pi_0$ and above by $\Phi\left(\lambda_{jk}, \mu, \sigma\right) = 1 - \pi_1$.

Hurdle models also are available for the probability-ratio framework. A class of hurdle models for beta distribution models has been presented by Ospina and Ferrari (2012), and Smithson and Shou (2019) describe a hurdle model for cdf-quantile distributions. Hurdle models are a reasonable choice when there are a lot of zeros and/or ones, and especially if there is justification for modeling them separately from the rest of the data.

Examples

Emergency Department Patient Time-Intervals: Log-Ratios and Probability-Ratios

In this example we compare the log-ratio and probability-ratio approaches, with the latter using the logit-normal distribution for the marginals to demonstrate that it yields the same estimates as the log-ratio Gaussian model does. The data are from the Yoon et al. (2003) emergency department study. The predictors of the patient length-of-stay composition we will consider include initial triage level, whether the patient arrived by ambulance, services or interventions (lab, x-ray, computed tomography, ultrasound, nuclear medicine, specialty consultation), and length of stay (LOS).

The original data comprised 894 cases whose times were recorded for five stages of emergency department assessment and treatment: registration, triage assessment, nursing assessment, physician assessment, and disposition decision. The cases included a substantial number of zeros. There were 696 cases with no time devoted to the decision stage, so we formed an amalgamation by adding the decision stage times to the physician stage times. There also were substantial numbers of zeroes in the registration stage (182) and in the triage stage (256). However, only 60 of the cases had zeros in both of these stages and both stages were considered as initial patient assessments, so again we formed an amalgamation by combining the registration and triage stages.

The resulting composition has three parts: registration-triage, nursing assessment, and physician-disposition. There are only 104 cases with zeros, so for now these are set aside for separate analysis. The 790 cases for which no zeroes occur in any of the three stages are included in the following compositional analysis.

Denoting the proportions of LOS devoted to the registration-triage, nursing assessment, and physician-disposition stages by y_r, y_n, and y_p, respectively, for the log-ratio analysis we use $\log(y_r/y_p)$ and $\log(y_n/y_p)$, while for the probability-ratio analysis we use $y_r/(y_r + y_p)$ and $y_n/(y_n + y_p)$.

Both pairs of variables are moderately positively correlated (r = .666 for the log-ratios and.655 for the probability-ratios), and the relationships appear to be linear in the log scale. This pattern suggests that the trade-off between the physician-disposition stage and the two earlier stages is spread evenly across those two stages, and the predictors of both pair members should be similar.

Indeed, the regression models for each of the pair members produce qualitatively similar results in the mean-response (location) submodels, with identical predictors whose coefficients have the same signs and similar magnitudes. Table 7 shows the coefficients for the final logit-normal regression model of the probability-ratios. In the supplemental materials we show that a heteroscedastic normal distribution log-ratio regression model with a dispersion submodel produces identical coefficients and standard errors as the probability-ratio model presented in Table 7. Likewise, a conventional homoscedastic normal regression model of the log-ratios gives the same coefficients as a homoscedastic version of the logit-normal probability-ratio model.

Table 7
Three-Part Composition Probability-Ratio Logit-Normal Model

Location submodel: Coefficient	Estimate	Std. error	t value	p
Intercept	−0.062	0.079		
Stage	−0.110	0.110	−1.002	.317
LOS	−0.140	0.008	−16.896	<.001
Triage12	−1.607	0.165	−9.725	<.001
Triage3	−0.527	0.065	−8.051	<.001
Ambulance	−0.831	0.100	−8.327	<.001
Stage × Ambulance	0.460	0.140	3.287	.001
Intervention	−1.001	0.097	−10.331	<.001
Stage × Intervention	0.310	0.132	2.354	.019
Dispersion submodel:				
Coefficient	Estimate	Std. error	t value	p
Intercept	0.279	0.034		
Stage	0.034	0.036	0.951	.342
Triage3	−0.130	0.039	−3.299	.001
Intervention	−0.162	0.038	−4.240	<.001

Note. LOS = length of stay.

Likelihood-ratio tests and Akaike information criterion (AIC) values suggest that the best location submodel has LOS, the first three triage levels (Triage1–2 and Triage3), ambulance delivery, and intervention (any combination of the x-ray, lab, and other interventions). Stage is a dummy variable coded 0 for the first probability-ratio (registration-triage stage) and 1 for the second (nursing-assessment stage). The model accounts for about 46% of the variance in the logit-probability-ratios.

The location model indicates that the LOS and triage effects do not discernibly differ between the two probability-ratios. The negative coefficients for LOS and the triage variables indicate that the relative amounts of time devoted to the earlier stages of emergency department treatment are lower for longer stays and more severe triage levels. Note that length of stay effect has not been mediated by any of the other predictors, suggesting that there is something else about length of stay besides triage levels and interventions that contributes to the tradeoffs being examined here. The effects of arrival by ambulance and interventions (in the form of special testing or diagnostic procedures) are moderated by stage. Both variables predict lower proportions of time spent in the first two stages, but not as much lower for the nursing-assessment stage. These effects probably have to do with the seriousness of the case, but they are independent of the triage-level effects.

The best dispersion submodel produces negative coefficients for Triage Level 3 and any intervention. The absence of moderator effects due to stage indicates that predictor effects on dispersion do not differ between the two log-ratios or probability-ratios. The Triage Level 3 has less variance than the other levels, and procedural interventions reduce variance too, regardless of which of the two earlier stages the patient is in. These effects may indicate that procedures in these situations are more standardized than under other conditions.

The original study (Yoon et al., 2003), as mentioned before, treated LOS as the dependent variable. Intermediate Triage Levels 3 and 4 predicted longer LOS than the other triage levels. Likewise, interventions and specialty consultation were associated with longer LOS. Our findings complement the original study by providing contrasting focuses on the more serious triage levels and the additional factor of arrival by ambulance.

We discussed our findings with the lead author of the study, who regarded our findings as both plausible and interesting and noted that to a large extent they probably can be accounted for by the seriousness of the patient's condition (Dr. Phillip Yoon, personal communication, August 5th, 2019). Those in the lower triage-levels or arriving by ambulance often are critically ill or injured and their diagnoses are clear from the start, so the earlier assessment stages can be skipped in order to get them into treatment quickly. They also are more likely to require specialist interventions or consultations, which in turn take more time. Those in the higher triage levels (Levels 4–5) are less urgent and may be more difficult to diagnose, so proportionately more time is needed for the initial assessment stages. The LOS effect may be due to the longer LOSs in the intermediate triage stages, where cases are not as urgent as for the lower triage levels.

Emergency Department Example Continued: Dealing With Zeros

As mentioned earlier, after combining stages there are 104 zeros in the emergency department time-interval data. The preceding subsection presented an analysis of the cases with no zeros. Here, we compare two of the methods for dealing with zeros and ones: multiplicative replacement and a hurdle model. Earlier we observed that arguments could be made for treating the emergency department zeros as either rounded or essential. A resolution of these arguments may begin by examining the distribution of the zeros, with the goal of assessing whether they appear to be qualitatively different from the nonzeros. We already have seen that most of the predictors of the nonzero cases stem from the seriousness of the patients' injuries or disorders. If the pattern of zeros suggests a similar explanation for them then that might favor replacement over a hurdle model.

Table 8 shows the distributions of zeros and nonzeros by triage level. Both the registration and nursing-assessment stages have higher percentages of zeros for Triage Level 1–2 (the most serious level) than for the other levels. In the registration stage, the Level 3 percentage is somewhat higher than those for Levels 4 and 5. Another indication that the seriousness of the patient's condition may be contributing to zeros is that out of 211 ambulance arrivals, 18.4% were zeros, whereas out of 662 walk-ins, only 3.1% were zeros. Both the effects of triage level and ambulance arrival appear to be explicable in terms of the seriousness of the patient's condition.

This preliminary examination suggests that the zeros are not qualitatively distinct from the nonzeros. Nevertheless, for purposes of illustration and completeness, we present the logistic regression in a hurdle model that treats the zeros as distinct from the nonzeros. Table 9 displays the best model for the zeros. The "stage" variable is coded 0 for the registration stage and 1 for the nursing-assessment stage. As foreshadowed, Triage Level 1–2 predicts a larger number of zeros, as does Triage Level 3 in the registration stage. It turns out that the ambulance effect occurs solely in the registration stage. Finally, the stage effect reflects a greater number of zeros for patients in Triage Levels 4 and 5 in the nursing-assessment stage.

The variables predicting the zeros are a subset of those predicting the nonzero values, which suggests that the zeros and nonzeros may not constitute distinct subsets of the data. To test this possibility, we use the multiplicative replacement method, substituting .01 for the zeros and employing the relevant J factor for the nonzeros

Table 8

Distribution of Zeros by Triage Level

Registration	Triage 1–2	Triage 3	Triage 4	Triage 5
Nonzeros	39 (60.9%)	270 (90.9%)	321 (98.2%)	204 (99.0%)
Zeros	25 (39.1%)	27 (9.1%)	6 (1.8%)	2 (1.0%)
Nursing assess.	Triage 1–2	Triage 3	Triage 4	Triage 5
Nonzeros	46 (71.9%)	287 (96.6%)	311 (95.1%)	198 (96.1%)
Zeros	18 (28.1%)	10 (3.4%)	16 (4.9%)	8 (3.9%)

to ensure that each case sums to 1. We then run the same logit-normal model as before, but this time on the entire data-set.

Table 10 displays the output for this model, along with the coefficients from the earlier model for the nonzero subset of the data (in the column labeled "Subset"). The coefficient estimates for the new model are quite similar to those in the Subset column in both the location and dispersion submodels. The only mismatch in coefficient significance is the State*Intervention coefficient in the location submodel, which had $p = .019$ in the subset model and $p = .067$ in the new model. Given these results, we would conclude that the zeros are not sufficiently distinct from the nonzeros to merit a hurdle model, whereas the replacement method seems to provide a unified model of the entire data-set by treating the zeros as rounded rather than essential.

Employment Data Example: Logit-Normal Versus Beta Regression Models

In this example, we compare a log-ratio Gaussian model with a probability-ratio beta regression model. However, we will do this by transforming the log-ratio Gaussian model into a probability-ratio logit-normal model, so that we are comparing models of the dependent variable on the same scale.

Our example consists of employment data and Gross Domestic Product (GDP) per capita for 188 countries in 2015, from the International Labor Organization, ILOSTAT database (ILO, 2015; data retrieved in September 2018). The employment figures comprise the proportions of employed persons who are female versus male, and from each of three sectors: agricultural, industrial, and services. The employment composition therefore has six parts: $\{y_{af}, y_{am}, y_{if}, y_{im}, y_{sf}, y_{sm}\}$, where *f* denotes female, *m* male, *a* agriculture, *i* industry, and *s* services.

Suppose that we want to test two hypotheses about the effect of a nations' wealth (as measured by GDP per capita) on relative employment rates. The first hypothesis is that in wealthier countries female employment rates relative to male employment rates will be higher in the service sector and lower in the other two sectors. The second hypothesis is that in wealthier countries the employment rates for men and women combined in agriculture and industry will be lower relative to employment rates in the service sector. The resulting probability-ratios are:

$$\begin{aligned}
\gamma_{afm} &= y_{af}/(y_{af} + y_{am})\\
\gamma_{ifm} &= y_{if}/(y_{im} + y_{im})\\
\gamma_{sfm} &= y_{sf}/(y_{sf} + y_{sm})\\
\gamma_{as} &= (y_{af} + y_{am})/(y_{af} + y_{am} + y_{sf} + y_{sm})\\
\gamma_{is} &= (y_{if} + y_{im})/(y_{if} + y_{im} + y_{sf} + y_{sm})
\end{aligned} \tag{16}$$

The first three probability-ratios test the first hypothesis, and the latter two test the second hypothesis. We will use log(GDP/cap) as a predictor in both the location and dispersion submodels. To model the dependencies in the logit-normal and beta regression models, we will use t-copulas (see the supplemental materials for further details on the copula part of the models).

Table 11 shows the location submodel coefficients for each of the probability-ratios. All of the log(GDP/cap) coefficients are significant in both the logit-normal and beta regression models. Moreover, they are quite similar for both models. Wealthier countries have smaller proportions of women working in agriculture and industry relative to men's proportions, and this is reflected in the negative coefficients for γ_{afm} and γ_{ifm} in the logit-normal and beta models. In wealthier countries, women's employment rate in the service sector relative to men's rate is higher, as shown by the positive coefficient for γ_{sfm} in both regression models. Finally, the second hypothesis also is supported, with negative coefficients for γ_{as} and γ_{is}

Table 9

Logistic Regression Model for Zeros

Coefficient	Estimate	Std. error	*t* value	*p*
Intercept	−4.1757	0.2947	−14.169	<.001
Stage	2.7343	0.2976	9.189	<.001
Triage12	1.0790	0.3298	3.271	<.001
Triage3	1.2337	0.3421	3.606	<.001
Stage × Triage3	−1.2798	0.4715	−2.714	.007
Ambulance	1.2639	0.3038	4.160	.001
Stage × Ambulance	−1.9566	0.4626	−4.229	<.001

Table 10

Three-Part Composition Probability-Ratio Model With Zeros Replaced

Location submodel: Coefficient	Subset	Estimate	Std. error	*t* value	*p*
Intercept	−0.062	−0.204	0.080		
Stage	−0.110	−0.139	0.114	−1.212	.226
LOS	−0.140	−0.121	0.008	−15.518	<.001
Triage12	−1.607	−1.795	0.122	−14.728	<.001
Triage3	−0.527	−0.570	0.066	−8.621	<.001
Ambulance	−0.831	−0.942	0.094	−9.990	<.001
Stage × Ambulance	0.460	0.592	0.133	4.456	<.001
Intervention	−1.001	−0.988	0.097	−10.191	<.001
Stage × Intervention	0.310	0.246	0.134	1.835	.067
Dispersion Submodel:					
Coefficient	Subset	Estimate	Std. error	*t* value	*p*
Intercept	0.279	0.337	0.033		
Stage	0.034	0.064	0.034	1.897	.058
Triage3	−0.130	−0.126	0.037	−3.436	<.001
Intervention	−0.162	−0.187	0.036	−5.167	<.001

Note. LOS = length of stay.

Table 11

Location Submodel Coefficients, Logit-Normal, and β Regression

Prob.-ratio	Model	Intercept	Std. error	Log(GDPcap)	Std. error
γ_{afm}	logit-normal	1.847	0.353	−0.325	0.046
	beta	1.260	0.294	−0.237	0.036
γ_{ifm}	logit-normal	−0.115	0.333	−0.138	0.038
	beta	0.018	0.284	−0.140	0.032
γ_{sfm}	logit-normal	−1.282	0.271	0.121	0.029
	beta	−1.149	0.243	0.108	0.026
γ_{as}	logit-normal	7.697	0.396	−1.040	0.045
	beta	6.434	0.386	−0.872	0.045
γ_{is}	logit-normal	−0.096	0.213	−0.106	0.023
	beta	−0.042	0.206	−0.107	0.023

confirming that employment rates in agriculture and industry relative to the service sector decline as GDP per capita increases.

We now turn to the dispersion submodels for both regression models, whose coefficients are displayed in Table 12. There is a general tendency for dispersion in the probability-ratios to decrease as GDP per-capita increases, with the exception of the female-to-male agricultural sector ratios. We cannot directly compare the coefficients for the logit-normal model with the beta model because their dispersion models do not share the same scale. However, we can make qualitative comparisons by examining the direction and significance (indicated by the stars * in the table) of these coefficients. The models disagree on γ_{ifm} and γ_{as}, with the beta regression submodel finding significant negative effects and the logit-normal model not finding significant effects in both cases.

Goodness of fit for the regression models and their respective copulas can be compared via AIC values. In Table 13, we can see that the beta model fits best for the first three probability-ratios (the ones comparing female to male employment rates in each of the three sectors), whereas the logit-normal models were better for the remaining two probability-ratios (which compare employment rates in agriculture and industry with the rates in the service sector). The AIC values for the copulas are similar, slightly in favor of the logit-normal model.

On this basis, it would be reasonable to construct a "hybrid" model with the beta marginals for the first three probability-ratios and logit-normal models for the remaining two probability-ratios. The resulting copula has AIC = −183.0, slightly worse than the other two models' copulas. However, the gains in fit for the marginal distribution models more than compensates for that. Table 14 displays the estimated correlations from the copula model along with the sample correlations. The copula model captures the correlation pattern reasonably well.

Table 12

Dispersion Submodel Coefficients, Logit-Normal, and β Regression

Prob.-ratio	Model	Intercept	Std. error	Log(GDPcap)	Std. error
γ_{afm}	logit-normal	−2.330	0.340	0.263*	0.039
	beta	−2.079	0.285	0.114*	0.032
γ_{ifm}	logit-normal	−0.274	0.282	−0.002	0.032
	beta	−0.695	0.265	−0.074*	0.030
γ_{sfm}	logit-normal	0.923	0.317	−0.174*	0.036
	beta	−0.165	0.280	−0.141*	0.032
γ_{as}	logit-normal	−0.331	0.314	0.026	0.036
	beta	−0.066	0.307	−0.134*	0.037
γ_{is}	logit-normal	−0.114	0.291	−0.078*	0.033
	beta	−0.869	0.276	−0.088*	0.032

Note. * Significant coefficient, $p < .05$.

Table 13

AIC Values for the Marginal Regression and Copula Models

Model components	Logit-normal	beta
γ_{afm}	−143.7	−177.7
γ_{ifm}	−255.8	−286.7
γ_{sfm}	−221.7	−229.5
γ_{as}	−326.8	−319.6
γ_{is}	−364.4	−358.7
copula	−190.2	−187.6

Probability Judgment Example: Alternative Log-Ratio Models

We now present an example of two alternative log-ratio models applied to the same data-set. We illustrate the isometric log-ratio model, a type of model popular in the compositional analysis literature, and compare it with an alternative model that employs amalgamations via a "stick-breaking" procedure. We will show that although both models appear to be addressing the same questions about the data, they end up with differing results and interpretations. We have two goals. One is to alert researchers to pay sufficient attention to the details of a compositional model to ensure that the model fulfills their specific requirements or goals. The second is to introduce readers to an ongoing debate in the compositional analysis literature about the utility and legitimacy of these models.

Our example is a reanalysis of data from Budescu et al. (2009), a study of how 223 members of the public interpret Intergovernmental Panel on Climate Change (IPCC) probabilistic uncertainty phrases in the IPCC's fourth report. Here, we focus on the data from participants' numerical translations of the phrases "likely," "very likely," "unlikely," and "very unlikely." Each of these phrases was used in three sentences presented to participants. Participants provided their "best" estimate of the probability intended in each sentence, and their lowest and highest values that they believed encompasses the range of plausible estimates.

Participants in the study were assigned to one of four conditions: *control*, where they were given no instructions in how to interpret the probability expressions; *treatment*, where they were

Table 14

Hybrid Copula Model Estimates and Empirical Correlations: Employment Data

Correlations	Estimate	Std. err.	Sample
$\hat{\rho}_{afm,ifm}$	0.278	0.153	0.217
$\hat{\rho}_{afm,sfm}$	0.348	0.166	0.278
$\hat{\rho}_{afm,as}$	0.408	0.091	0.403
$\hat{\rho}_{afm,is}$	−0.000	0.248	−0.057
$\hat{\rho}_{ifm,sfm}$	0.545	0.161	0.438
$\hat{\rho}_{ifm,as}$	0.136	0.198	0.134
$\hat{\rho}_{ifm,is}$	0.213	0.447	0.268
$\hat{\rho}_{sfm,as}$	−0.061	0.298	−0.136
$\hat{\rho}_{sfm,is}$	−0.226	0.244	−0.233
$\hat{\rho}_{as,is}$	0.192	0.168	0.187

able to view the IPCC guidelines for numerically translating these expressions; *narrow*, where a small interval (lengths of 10% for "likely" and "unlikely," 5% for "very likely" and "very unlikely") of appropriate values was displayed next to each expression in the sentence text; and *wide*, where the considerably wider ranges used by the IPCC were displayed next to each expression.

Our primary interest is in assessing the effects of the conditions on the width of the interval between the lower and upper estimates, but we also are interested in identifying effects of expression valence (the presence of "likely" versus "unlikely" in each expression), and expression extremity (the presence or absence of "very" in the expressions). Whereas such an assessment would be cumbersome if we worked with the probability estimates themselves, it is simplified by converting the estimates into three interval widths, where b_1 is the width of the interval from 0 to the lower probability estimate, b_2 is the width of the interval between the lower and upper estimate, and b_3 is the width of the interval from the upper estimate to 1. These three interval widths form a three-part composition.

We compare two models, an isometric log-ratio (ILR) model and a stick-breaking (SB) model. The rationale behind both models is that we are interested primarily in the impact of the experimental variables on the width of the interval between participants' lower and upper probability estimates (i.e., the middle interval, b_2), relative to the remaining amount of the (0,1) interval. There are two ways we can examine the relative width of b_2.

First, the ILR model compares the middle interval b_2 with the geometric mean of b_1 and b_3, the first and third intervals, and the third interval b_3 with the first interval b_1. From Equation 13, the two ILR log-ratios may be written as

$$\begin{aligned}\sqrt{2/3}\log\left(b_2/\sqrt{b_1 b_3}\right) &= \lambda_1\\ \sqrt{1/2}\log(b_3/b_1) &= \lambda_2\end{aligned} \tag{17}$$

Second, the SB model compares the middle interval with the sum of the other two intervals (the first break in a "stick" consisting of $\{b_2, b_3, b_1\}$), and the third interval with the first interval (the break in the remaining part of the "stick"). Thus, the SB log-ratios are

$$\begin{aligned}\log\left(b_2/(b_1+b_3)\right) &= \lambda_1\\ \log(b_3/b_1) &= \lambda_2\end{aligned} \tag{18}$$

The primary difference between the two models is between their first log-ratios. The nub of the aforementioned debate in the compositional literature is whether an amalgamation model such as the SB is legitimate. Several authorities, for example, Pawlowsky-Glahn and Egozcue (2006), have raised the criticism that because amalgamations are sums, they are not linear in the simplex and therefore should be ruled out. They prescribe geometric means as a substitute, as seen in our ILR model, and they also show that an ILR forms an orthonormal basis of the compositional data vectors. However, other authors, for example, Greenacre (2020), have argued that amalgamations not only are permissible but also considerably easier to interpret than geometric means or ILRs generally. We shall investigate these claims through our example here. We turn now to the regression analysis. To enhance the clarity of this illustration, we make the following simplifications:

1. Each participant's three-sentence estimates are averaged, so that the data for each participant consist of a single set of interval widths for each condition-valence-extremity combination (i.e., 12 data-points per participant).

2. Both the ILR and SB models are limited to main effects from condition, valence, and extremity.

3. The distinction between the two log-ratios in each model is treated as a fixed effect, while condition is a between-subjects effect and therefore is treated as fixed. Valence and probability expression, on the other hand, are treated as random effects.

The resulting regression model is:

$$\lambda_{ij} = \delta_{0ij} + \beta_{nj}C_{nij} + \beta_{tj}C_{tij} + \beta_{wj}C_{wij} + \delta_{1ij}V_{ij} + \delta_{2ij}E_{ij} \tag{19}$$

where λ denotes a log-ratio, with $j = 1, 2$ indexing log-ratios and $i = 1, \ldots, 223$ indexing participants; the C variables are (0,1) indicator variables corresponding to the narrow, treatment, and wide conditions, respectively, V is an indicator variable denoting negative valence when $V = 1$ and positive valence when $V = 0$, E is an indicator variable denoting the absence of "very" from the probability expression when $E = 1$ and its presence when $E = 0$. The random-effects coefficients are defined as follows:

$$\begin{aligned}\delta_{0ij} &= \beta_{0j} + u_{0ij}\\ \delta_{1ij} &= \beta_{1j} + u_{1ij}\\ \delta_{2ij} &= \beta_{2j} + u_{2ij}\end{aligned} \tag{20}$$

The u_{kij} have the usual multilevel random-effect error-term properties, that is, means of 0 and the conventional variance-covariance matrix. In our analyses, we deal with the j index via another indicator variable, which we denote by D, and whose role is specific to the ILR and SB models (and therefore described in each of these models).

ILR and SB Models

A full elaboration of the ILR and SB regression results is contained in the supplemental materials. Both models fit the data reasonably well, and to similar degrees. The ILR model's correlations between fitted values and empirical data are .814 for the first variate and .831 for the second, while the SB model's correlations are .845 for the first variate and.836 for the second.

Here, we focus on the disagreement between the two regression models. Table 15 displays the fixed-effects t statistics for both variates in the ILR and SB models. The models broadly agree on the fixed-effects for the b_3/b_1 variate because this variate is identical in both models up to multiplication by a scalar. However, for their first log-ratios the models differ on effects from extremity and valence. The ILR model finds significant negative effects for both variables, whereas the SB model finds only a marginally significant negative effect for valence and, more drastically, finds a significant *positive* effect for extremity ("very" absent).

Greenacre et al. (2021) demonstrate similar disagreements between ILR and amalgamation models to those we have identified in our example. They argue that given a constant sum of two parts, say $b_1 + b_3$, their geometric means still may vary due to variability in their ratio, b_1/b_3. Thus, a geometric mean of parts cannot be regarded as equivalent to combining the parts. This

Table 15
Isometric Log-Ratio and Stick-Breaking Regression t Statistics

Fixed effects	ILR *t* statistics	*p*	SB *t* statistics	*p*
b_3 vs. b_1				
Intercept	−11.914	<.001	−11.980	< .001
Narrow	−1.771	.078	−1.771	.078
Treatment	−1.046	.297	−1.046	.297
Wide	−1.871	.063	−1.871	.063
“Very” absent	14.190	<.001	15.180	< .001
Negative valence	7.386	<.001	7.901	<.001
D	12.095	<.001	8.323	<.001
b_2 vs. b_3 and b_1				
Intercept	5.962	<.001	−5.565	<.001
Narrow	−4.080	<.001	−4.997	<.001
Treatment	−1.190	.235	−1.963	.083
Wide	−0.463	.644	−1.407	.161
“Very” absent	−3.512	.001	3.455	.001
Negative valence	−4.017	<.001	−2.011	.046

intuitive argument can be made more explicit and rigorous by comparing the inverse transformations of the ILR and SB log-ratios, that is, the functions that relate these log-ratios back to the composition parts.

The SB inverse transformation is

$$b_1 = \frac{1}{(e^{\lambda_1}+1)(e^{\lambda_2}+1)}$$
$$b_2 = \frac{e^{\lambda_1}}{e^{\lambda_1}+1} \tag{21}$$
$$b_3 = \frac{e^{\lambda_2}}{(e^{\lambda_1}+1)(e^{\lambda_2}+1)}$$

Equation 21 cleanly separates the effects on λ_1 and λ_2 when translating these into effects on the composition parts. We can see that b_2 is influenced solely by λ_1, so that any regression coefficients for the first variate apply directly to effects on b_2. Moreover, it is clear that increasing λ_1 both increases b_2 and simultaneously decreases both b_1 and b_3 (not just their sum) if λ_2 is held constant. Effects on λ_2, on the other hand, affect b_1 and b_3 but not b_2, with its increase boosting b_3 and decreasing b_1.

The ILR inverse transformation is

$$b_1 = \frac{1}{e^{\sqrt{\frac{3}{2}}\lambda_1+\frac{\lambda_2}{\sqrt{2}}}+e^{\sqrt{2}\lambda_2}+1}$$
$$b_2 = \frac{e^{\sqrt{\frac{3}{2}}\lambda_1+\frac{\lambda_2}{\sqrt{2}}}}{e^{\sqrt{\frac{3}{2}}\lambda_1+\frac{\lambda_2}{\sqrt{2}}}+e^{\sqrt{2}\lambda_2}+1} \tag{22}$$
$$b_3 = \frac{e^{\sqrt{2}c_2}}{e^{\sqrt{\frac{3}{2}}\lambda_1+\frac{\lambda_2}{\sqrt{2}}}+e^{\sqrt{2}\lambda_2}+1}$$

Equation 22 shows that the ILR model does not separate effects on λ_1 from those on λ_2 in its treatment of b_2. Instead, effects on both λ_1 and λ_2 influence all three composition parts. In particular, it is clear that in the ILR model b_2 is driven both by its ratio to the geometric mean of b_3 and b_1 and by the ratio of b_3 to b_1. In contrast, the SB model b_2 is driven solely by its ratio to the sum of b_3 and b_1.

The geometric mean of compositional parts cannot be interpreted as simply combining the parts, whereas their sum can. On the other hand, the geometric mean is the appropriate mean in the simplex. The take-home lesson is that a researcher’s choice of a model in compositional analysis must be closely guided by their goals or hypotheses. For example, if we want a model for which effects on b_2 correspond solely with effects on λ_1, then the SB model is a better alternative than the ILR model. However, if instead we are interested in how b_2 relates to the barycenter of the other two intervals, then the ILR model is the best alternative.

Discussion and Conclusions

New and exciting methods of data collection are sometimes avoided because of a lack of methods for accurately analyzing the data collected. We believe this happened in social science with regard to ipsative data collection. Moreover, sometimes inaccurate inferences are drawn from data due to dependencies that are overlooked, such as constraints on parts to sum to a constant for individual cases. For compositional data, both of these problems can be addressed through a broader understanding and application of compositional data analysis. We therefore have provided a practical introduction to compositional data analysis for the social scientist, including demonstrations of the different forms of social science data that are well suited for analysis using this method.

We have introduced several important distinctions regarding compositional data analysis that are helpful for designing data collection and choosing analysis strategies. Data may be optionally or unavoidably compositional. Optionally compositional data can be viewed through compositional and noncompositional perspectives to answer different research questions. Unavoidably compositional data requires the compositional approach to generate accurate inferences.

Compositions can be transformed out of the simplex using log-ratios or probability-ratios (which we summarize in detail next). Zeros cannot not be transformed out of the simplex. Our approach to handling zeros depends on whether they are essential or rounded. This distinction enables researchers to choose whether to separate out cases with essential zeros for analysis via a hurdle model, to use amalgamations for removing sparse parts, or to replace rounded zeros with small positive values.

Having introduced various approaches to analyzing compositional data, a summing-up is in order with an overview of the strengths and weaknesses of each approach, along with some guidance for choosing among them. We then conclude with descriptions of outstanding problems and topics of active research and development in compositional analysis.

The log-ratio approach is the most well-developed of the two approaches presented here. It has an excellent theoretical rationale, including well-developed statistical and geometric theory. It provides researchers the option of analysis within the simplex and via transforms to Cartesian coordinates. It can model associations among composition parts, and other kinds of relations among parts, notably distances between compositions. Its main drawbacks are in the domain of practice rather than theory, per se. No obvious way to model dispersion is presented in the conventional log-ratio framework; although, as we have demonstrated, it is available via heteroscedastic linear regression. Likewise, there is an overreliance on normal theory regression models. Both of these deficiencies can be remedied. Finally, some transforms (e.g., ILR) are preferred by some advocates of the log-ratio framework on theoretical grounds but are difficult to interpret, whereas others, such as amalgamations, are proscribed on theoretical grounds but possess practical utility. As indicated earlier, these issues currently are debated.

The probability-ratio approach is new and relatively untried. It does have a theoretical basis, and is congruent with almost all aspects of the log-ratio approach; any set of log-ratios can be converted into probability-ratios and vice versa. Indeed, for some practical purposes it can be treated as an alternative “point of view” within the log-ratio framework. Its main advantages is that it affords ready access to all of the distributions whose support is (0,1), such as the beta distribution, and therefore is able to connect with univariate regression methods for (0,1) variables as well. Dispersion modeling is inbuilt and flexible, because of its access to a wide variety of marginal distributions and the absence of any constraints on dispersion parameters. It is also well-suited to modeling covariance structures with copulas, thereby going beyond the restriction to multivariate normal distribution models. However, some aspects of model diagnostics for it are underdeveloped, and as of this writing it lacks dedicated software.

There are at least four open problems or areas for development pertaining to compositional analysis generally. First, the zeros-and-ones problem has not seen an entirely satisfactory resolution, especially for essential zeros. In addition to the three somewhat ad hoc solutions reviewed in this article, Greenacre (2018) suggests using correspondence analysis as a substitute for log-ratio analysis, because of their close connections and the fact that it can handle zeros in the data. He does note that correspondence analysis is not subcomposition coherent, but he argues that it is near enough to coherence to be useful. We have not seen mention of hurdle models in the composition analysis literature, but we view it as the most satisfactory way of dealing with essential zeros so far.

Second, the ILR transform and amalgamations controversies in the log-ratio community have yet to be resolved. The usual complaint against ILR transforms is their interpretative opacity and inability to directly test certain kinds of hypotheses. The standard argument against amalgamations is that they are not linear in the simplex. Our own position regarding amalgamations is similar to the view that amalgamations have practical utility and are legitimate if appropriately used (Greenacre et al., 2021).

A third topic for development has not been featured in the compositional analysis literature, but we raise it here: methods for quantile compositional analysis. Applying quantile regression to log-odds-ratios would seem to be straightforward, thereby affording the advantages of quantile regression to compositional analysis (e.g., requiring fewer assumptions about the underlying distribution). However, we are not aware of any systematic treatment of quantile regression for compositional data. In the probability-ratios framework, the cdf-quantile family (Smithson & Shou, 2017) is available. These distributions model the median with a location parameter and the remaining quantiles with a dispersion parameter in tandem with the location one. There has yet to be a proper comparison between quantile regression and cdf-quantile regression, and compositional data analysis would constitute one appropriate setting for such a comparison.

Fourth, and finally, it seems that the compositional data analysis literature has rarely discussed Bayesian methods. Major textbooks do not treat them in any depth (e.g., Pawlowsky-Glahn et al., 2015; Van den Boogaart & Tolosana-Delgado, 2013). Most of the published articles on Bayesian methods for compositional data analysis employ the Dirichlet distribution rather than the log-ratio approach (e.g., Van der Merwe, 2019), or are restricted to analyzing counts rather than continuous variables (e.g., Huston & Schwarz, 2012; Napier et al., 2015). One understandable reason for neglecting this topic is that the traditional log-ratio approach utilizes normal-theory regression and therefore is amenable to well-known hierarchical Bayesian modeling techniques. Nonetheless, the employment of alternative distributions may motivate further developments in Bayesian approaches to compositional analysis.

We have endeavored in this article not only to introduce and compare techniques for compositional data analysis, but also to persuade readers that analyzing data from a compositional standpoint is useful. Two of our three examples involve data that are optionally compositional, and these have illustrated how a compositional approach can provide novel insights that otherwise would be inaccessible. Software resources are available for conducting all of the varieties of compositional data analysis featured in this article, and guidance to those resources is provided in the supplemental materials.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, Ltd.

Bowen, C.-C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis*, *10*(3), 240–259. https://doi.org/10.1108/eb028952

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science*, *20*(3), 299–308.

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508–512. https://doi.org/10.1038/nclimate2194

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review*, *51*(5), 292–303. https://doi.org/10.1037/h0057299

Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures (Psychometric Monograph No. 14)*. Psychometric Society.

Comas Cufí, M., & Fernández de Henestrosa, S. T. I. (2011). *Codapack 2.0: a stand-alone, multi-platform compositional software* [Computer software manual]. Universitat de Girona. Department d’Informàtica i Matemàtica Aplicada.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, *39*(5), 829–844. https://doi.org/10.2307/1909582

Cunningham, W. H., Cunningham, I. C., & Green, R. T. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly*, *41*(3), 379–384. https://doi.org/10.1086/268394

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology*, *35*(6), 407–412. https://doi.org/10.1037/h0058853

Graffelman, J., Pawlowsky-Glahn, V., Egozcue, J. J., & Buccianti, A. (2018). Exploration of geochemical data with compositional canonical biplots. *Journal of Geochemical Exploration*, *194*, 120–133. https://doi.org/10.1016/j.gexplo.2018.07.014

Greenacre, M. (2018). *Compositional data analysis in practice*. CRC Press.

Greenacre, M. (2020). Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their log ratios have an inverse transformation. *Applied Computing and Geosciences*, *5*, Article 100017. https://doi.org/10.1016/j.acags.2019.100017

Greenacre, M., Grunsky, E., & Bacon-Shone, J. (2021). A comparison of isometric and amalgamation log ratio balances in compositional data analysis. *Computers & Geosciences*, *148*, Article 104621. https://doi.org/10.1016/j.cageo.2020.104621

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167–184. https://doi.org/10.1037/h0029780

Huston, C., & Schwarz, C. (2012). Hierarchical Bayesian strategy for modeling correlated compositional data with observed zero counts. *Environmental and Ecological Statistics*, *19*(3), 327–344. https://doi.org/10.1007/s10651-012-0189-0

ILO. (2015). ILOSTAT. https://ilostat.ilo.org/topics/employment/

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678–703.

Karni, E., & Vierø, M.-L. (2013). Reverse Bayesianism: A choice-based theory of growing awareness. *American Economic Review*, *103*(7), 2790–2810. https://doi.org/10.1257/aer.103.7.2790

Lewi, P. J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneimittel-Forschung*, *26*(7), 1295–1300.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, *35*(3), 253–278. https://doi.org/10.1023/A:1023866030544

Napier, G., Neocleous, T., & Nobile, A. (2015). A composite Bayesian hierarchical model of compositional data with zeros. *Journal of Chemometrics*, *29*(2), 96–108. https://doi.org/10.1002/cem.2681

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Pawlowsky-Glahn, V., & Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, *264*(1), 1–10. https://doi.org/10.1144/GSL.SP.2006.264.01.01

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Wiley.

Shou, Y., & Smithson, M. (2019). cdfquantreg: An R package for CDF-quantile regression. *Journal of Statistical Software*, *88*(1), 1–30. https://doi.org/10.18637/jss.v088.i01

Sisson, E. D. (1948). Forced choice—The new army rating 1. *Personnel Psychology*, *1*(3), 365–381. https://doi.org/10.1111/j.1744-6570.1948.tb01316.x

Smithson, M. (2019). Imprecise compositional data analysis: Alternative statistical methods. *Proceedings of Machine Learning Research*, *103*, 364–366.

Smithson, M., & Shou, Y. (2017). CDF-quantile distributions for modelling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 412–438. https://doi.org/10.1111/bmsp.12091

Smithson, M., & Shou, Y. (2019). *Generalized linear models for bounded and limited quantitative variables*. SAGE Publications.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314.

Templ, M., Hron, K., & Filzmoser, P. (2011). *robcompositions: An R-package for robust statistical analysis of compositional data* [Computer software manual].

Tsagris, M., & Athineou, G. (2016). Compositional: Compositional Data Analysis (R package version 5.2). https://CRAN.R-project.org/package=Compositional

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2008). “Compositions”: A unified R package to analyze compositional data. *Computers & Geosciences*, *34*(4), 320–338. https://doi.org/10.1016/j.cageo.2006.11.017

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R* (Vol. 122). Springer.

Van der Merwe, S. (2019). A method for Bayesian regression modelling of composition data. *South African Statistical Journal*, *53*(1), 55–64.

van Eijnatten, F. M., van der Ark, L. A., & Holloway, S. S. (2015). Ipsative measurement and the analysis of organizational values: An alternative approach for data analysis. *Quality and Quantity*, *49*(2), 559–579. https://doi.org/10.1007/s11135-014-0009-8

Verkuilen, J., & Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, *37*(1), 82–113. https://doi.org/10.3102/1076998610396895

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, *21*(4), 1–21. https://doi.org/10.18637/jss.v021.i04

Yoon, P., Steiner, I., & Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine*, *5*(3), 155–161.

Received August 10, 2021
Revision received October 13, 2021
Accepted October 18, 2021 ■