Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 45, Number 4, 2016

Special Issue on Compositional Data Analysis

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 45, Number 4, 2016

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 45, Heft 4, 2016

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Statistics Austria & Vienna University of Technology
Editorial Board:	Peter Filzmoser, Vienna University of Technology Herwig Friedl, TU Graz Bernd Genser, University of Konstanz Peter Hackl, Vienna University of Economics, Austria Wolfgang Huf, Medical University of Vienna, Center for Medical Physics and Biomedical Engineering Alexander Kowarik, Statistics Austria, Austria Johannes Ledolter, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien & Management Sciences, University of Iowa Werner Mueller, Johannes Kepler University Linz, Austria Josef Richter, University of Innsbruck Milan Stehlik, Department of Applied Statistics, Johannes Kepler University, Linz, Austria Wolfgang Trutschnig, Department for Mathematics, University of Salzburg Regina Tüchler, Austrian Federal Economic Chamber, Austria Helga Wagner, Johannes Kepler University Walter Zwirner, University of Calgary, Canada
Book Reviews:	Ernst Stadlober, Graz University of Technology
Printed by	Statistics Austria, A-1110 Vienna

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

Compositional data (CoDa) are random vectors representing parts of a whole which carry only relative information such as parts per unit, percentages or ppm along with other relative units such as molar compositions. Different fields have typical examples, for instance in Geology (geochemical elements), Economics (income/expenditure distribution), Medicine (body composition: fat, bone, lean), Genetics (genotype frequency), Chemistry (chemical composition), Ecology (abundance of different species), Paleontology (foraminifera taxa), Agriculture (nutrient balance ionomics), Environmental Sciences (soil contamination), Sociology (time-use surveys), the Food Industry (food composition: fat, sugar, etc.) or questionnaire surveys (ipsative data). The particular nature of CoDa renders most classical statistical techniques on compositions incoherent, as they were devised for unbounded real vectors. Modern CoDa analysis is founded on an own geometric structure for the simplex, i.e. an appropriate representation of the sample space of CoDa. Practitioners interested in CoDa can find a forum where information, material and ideas can be exchanged on the CoDaWeb (www.compositionaldata.com).

Specialist researchers, data analysts, postgraduate students or simply those with a general interest in CoDa or other constrained data sets, meet periodically at *CoDaWork*, the international Workshop on Compositional Data Analysis. This volume is a collection of noteworthy contributions to CoDaWork 2015 (L'Escala, Girona, Spain) rounded off with the fourth paper *The Mathematics of Compositional Analysis* where the authors provide a precise and unequivocal definition of the concepts of composition, CoDa sample space and subcomposition, on which all CoDa analysis is based. From the central, fundamental idea that a composition is an equivalence class and the sample space is the corresponding quotient space, it is shown that a logarithmic isomorphism induces a metric space structure. This structure allows for standard statistical analyses on the coordinates of compositions to be carried out.

The first contribution to this volume (*A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros*) extends the additive logistic normal distribution to handle essential zeros for continuous CoDa; where an essential zero in CoDa is a zero component which is not caused by rounding or some other difficulty in measurement, but rather, is genuinely believed to be zero.

In the second paper (*Changing the Reference Measure in the Simplex and Its Weighting Effects*), among the number of weighting techniques presented, the authors show changes that appear in the algebraic-geometric structure of the simplex, as well as some effects in elementary statistics and exploratory tools, when one applies a change of reference measure of the simplex.

The third contribution, entitled *Bayesian Estimation of the Orthogonal Decomposition of a Contingency Table*, introduces a Bayesian approach for a decomposition of a table into an independent table and an interaction table. Using a Dirichlet prior distribution for

the multinomial probabilities, a simulation of its posterior allows for the independence of the observed contingency table and cell interactions to be checked.

Finally, the last paper (*Compositional Uncertainty Should Not Be Ignored in High-Throughput Sequencing Data Analysis*) deals with the compositions generated by high throughput sequencing. The approach illustrated, one which can be extended to high-dimensional count CoDa, merges Bayesian estimation with log-ratio techniques. When examining the effect of using various approaches to estimating the value of zero, the combination of estimating technical variation and the centered log-ratio transformation is shown to provide a large increase in selectivity.

We would like to thank the authors and the referees involved in this volume, as well as the editorial team from the Austrian Journal of Statistics, for their outstanding contribution. After six editions of CoDaWork, this volume demonstrates that this hot research topic is continuously growing and evolving with the new advances in its theoretical basis and new methodological developments, all of which will have an enormous impact on applied fields. The next CoDaWork will be in Abbadia San Salvatore (Siena, Italy), 2017 5-9 June: you are kindly invited to participate!

Josep Antoni Martín-Fernández, Santiago Thió Fernández de Henestrosa
(Guest Editors)

University of Girona
Department of Computer Science, Applied Mathematics and Statistics
Campus Montilivi
17003 Girona
Spain

Girona, June 2016

Austrian Journal of Statistics
September 2016, Volume 45, 3–23.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.117

A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros

John Bear
Statistical Consulting Lab,
University of Arizona, U.S.A.

Dean Billheimer
Statistical Consulting Lab,
University of Arizona, U.S.A.

Abstract

The usual candidate distributions for modeling compositions, the Dirichlet and the logistic normal distribution, do not include zero components in their support. Methods have been developed and refined for dealing with zeros that are rounded, or due to a value being below a detection level. Methods have also been developed for zeros in compositions arising from count data. However, essential zeros, cases where a component is truly absent, in continuous compositions are still a problem.

The most promising approach is based on extending the logistic normal distribution to model essential zeros using a mixture of additive logistic normal distributions of different dimension, related by common parameters. We continue this approach, and by imposing an additional constraint, develop a likelihood, and show ways of estimating parameters for location and dispersion. The proposed likelihood, conditional on parameters for the probability of zeros, is a mixture of additive logistic normal distributions of different dimensions whose location and dispersion parameters are projections of a common location or dispersion parameter. For some simple special cases, we contrast the relative efficiency of different location estimators.

Keywords: composition, subcomposition, essential zero, logistic normal, projection.

1. Introduction

An essential zero in compositional data is a zero component which is not caused by rounding or some other difficulty in measurement, but rather, is genuinely believed to be zero. This is fundamentally a different problem than that addressed by recent work on rounded zeros, or below-detection level zeros, such as in Palarea-Albaladejo and Martín-Fernández (2015) and references therein. Although there are recent workable Bayesian approaches to zeros in compositions from count data, Martín-Fernández, Hron, Templ, Filzmoser, and Palarea-Albaladejo (2014) and references therein, essential zeros in continuous compositions still present a problem.

We develop an approach proposed by Aitchison and Kay (2003) to extend the logistic normal distribution to accommodate essential zeros. Aitchison (1986) and Aitchison and Kay (2003) note that a key feature compositional data is that ratios of the components contain all pertinent information about the composition. Essential zeros complicate this feature in that

they contain no information about the other components of the composition. In addition, an observation containing an essential zero is at the boundary of the simplex and is a composition of smaller dimension.

2. Previous work

In addition to the work mentioned above, there have been other approaches to zeros in compositions. Work by Butler and Glasbey (2008) mapped a latent Gaussian variable to a composition, but seems only to work for two and three-part compositions. An additional concern is that it does not preserve ratios of parts in subcompositions. In contrast, Leininger, Gelfand, Allen, and Silander Jr (2013) have developed a more practical treatment of compositions as coming from a latent Gaussian random variable where the compositional component is zero when the latent Gaussian component is less than or equal to zero. They develop a hierarchical power model with the transformation $X_k = \frac{(\max(0,Z_k))^\gamma}{1+\sum_{k'=1}^{d}(\max(0,z_{k'}))^\gamma}$ where Z_k is the k^{th} normal component and X_k is the corresponding compositional component. D is the number of parts in the composition, $d = D-1$, and $X_D = (1+\sum_{k'=1}^{d}(\max(0,Z'_k))^\gamma)^{-1}$. The corresponding inverse transformation is $Z_k = (X_k/X_D)^{1/\gamma}$ if $X_k > 0$, and $Z_k \leq 0$ (latent) if $X_k = 0$, for $k = 1, 2, \ldots, d$. To estimate parameters they use MCMC. One limitation of their approach is also a limitation of ours: we require one component of the composition to be strictly positive.

Work by Stewart and Field (2011) uses a multiplicative logistic normal mixture model that allows them to consider the univariate log odds for the i^{th} component to be normally distributed where the i^{th} component is not zero. It works well for their applications, in particular regression, but does not capture covariance easily.

Scealy and Welsh (2011) transform compositions into directional data on the hypersphere, and develop a regression model using the Kent distribution, Kent (1982), which tolerates zeros, though they write, "When any of the components of **u** are distributed too close to 0, boundary issues arise and in this case we need to pursue alternative approaches since the fitted Kent model (and the von Mises-Fisher model) may have significant support outside the positive orthant." A further issue with their approach is that their square root transformation does not preserve ratios of parts in subcompositions.

Our goal here is to extend the additive logistic normal distribution to handle essential zeros for continuous data.

3. Motivating example

Suppose we have compositional data on how much Bill spends on rice, lentils, and spices when he buys food. Suppose he buys in bulk, and occasionally the store is out of either the spices or lentils, but they always have plenty of rice. Table 1 shows a set of such compositions where some of the entries, for spices or lentils, are zero. Our goal is to develop a model for data like these by extending the additive logistic normal distribution.

4. Definitions (from Aitchison, 1986)

Definition: The *d-dimensional simplex embedded in D-dimensional real space* is the set of compositions, $\mathbf{x}$, defined by

$$\mathcal{S}^d = \{\mathbf{x} = (x_1, \ldots, x_d, x_D) : x_1 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = 1\},$$

where $d = D-1$. If $\mathbf{x} = (x_1, x_2, \ldots, x_d, x_D)^T$, then $\mathbf{x}_{-D} = (x_1, x_2, \ldots, x_d)^T$. The *additive logratio transformation,* alr, is defined as follows:

$$\text{alr} : \mathcal{S}^d \to \mathbb{R}^d$$

Table 1: Composition of food expense

	spices	lentils	rice
1	0.16	0.00	0.84
2	0.17	0.00	0.83
3	0.16	0.00	0.84
4	0.00	0.37	0.63
5	0.00	0.37	0.63
6	0.00	0.37	0.63
7	0.12	0.33	0.55
8	0.11	0.34	0.56
9	0.12	0.32	0.56
10	0.10	0.34	0.56
11	0.10	0.33	0.57
12	0.11	0.33	0.55

$$\mathbf{x} \mapsto \mathbf{y} = (\log(x_1/x_D), \log(x_2/x_D), \ldots, \log(x_d/x_D))^T. \tag{1}$$

We define the shorthand $\log(\mathbf{x}_{-D}/x_D) = (\log(x_1/x_D), \log(x_2/x_D), \ldots, \log(x_d/x_D))^T$. Since alr is one-to-one, its inverse exists. It is called the *logistic transformation*, alr^{-1}, defined as

$$\begin{aligned} \text{alr}^{-1} : \mathbb{R}^d &\to \mathcal{S}^d \\ \mathbf{y} \mapsto \mathbf{x} &= (x_1, x_2, \ldots, x_d, x_D)^T, \text{ where for } (i = 1, \ldots, d), \\ x_i &= \exp(y_i)/\{\exp(y_1) + \cdots + \exp(y_d) + 1\} \\ x_D &= 1/\{\exp(y_1) + \cdots + \exp(y_d) + 1\}. \end{aligned} \tag{2}$$

5. Simplifying assumption

In this section we outline our method for building a mixture distribution for dealing with compositions containing essential zeros, but leave most of the details about the weights for later. A key simplifying assumption we make throughout is that one of the parts of the composition, the D^{th} component, is never zero. We allow zeros anywhere else but not in the last component.

In a set of logistic normal data without zeros, the likelihood has been shown to be permutation invariant (Aitchison 1986). In our extension which allows zeros, if some parts are never zero, the likelihood is invariant to the choice of which one of those nonzero parts is chosen as the reference provided the same reference part is used throughout the data set.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_d, x_D)^T$ be a composition with $x_i < 1$ for all $i \in \{1, 2, \ldots, d, D\}$ and $x_D > 0$. For $i \in \{1, 2, \ldots, d\}$, consider two possibilities. Either $x_i = 0$ or $x_i > 0$. Let $W = \{i : i \in \{1, 2, \ldots, d\}, x_i > 0\}$. That is, W is the set of indices for the parts of $\mathbf{x}$ (other than x_D) which are nonzero (positive). For any given composition $\mathbf{x}$, W is the set of all the indices of the nonzero components of $\mathbf{x}$. There are $2^d - 1$ possible sets W. There are $2^d - 1$ and not 2^d because W cannot be empty. If W were empty that would require that $x_D = 1$ in order for $\mathbf{x}$ to be a composition, but we have already said we require all $x_i < 1$ including x_D. Each pattern of zeros corresponds to a different set W. We index them as W_ℓ with $\ell \in \{1, 2, \ldots, 2^d - 1\}$. They are elements of the power set, $W_\ell \in \mathcal{P}(\{1, 2, \ldots, d\})$. Sometimes we refer to these sets with incidence vectors where the i^{th} component $V_{W_\ell i} = 1 \iff x_i > 0$ and $V_{W_\ell i} = 0 \iff x_i = 0$.

Each W_ℓ has some probability of occurrence, $P(W_\ell)$. Although some pattern can be not

present $P(W_\ell) = 0$, the probabilities must sum to one,

$$\sum_{\ell=1}^{2^d-1} P(W_\ell) = 1.$$

We use the probabilities $P(W_\ell)$ as the weights in a mixture distribution. For the other distributions making up our mixture, we use logistic normal distributions $\mathcal{L}(\mathbf{x}; \boldsymbol{\mu}_{W_\ell}, \boldsymbol{\Omega}_{W_\ell})$ derived from a single parent logistic normal distribution $\mathcal{L}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Omega})$. They are in fact projections from the parent. We will call the distributions derived from the parent distribution *subdistributions* once we define them. So the mixture distribution will be denoted as follows, once we define a few more terms,

$$g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Omega}) = \sum_{\ell=1}^{2^d-1} P(W_\ell)\mathcal{L}(\mathbf{x}; \boldsymbol{\mu}_{W_\ell}, \boldsymbol{\Omega}_{W_\ell}).$$

In the parent distribution, $\mathcal{L}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Omega})$, $\boldsymbol{\mu}$ is a d-part location parameter vector, $\boldsymbol{\mu} \in \mathbb{R}^d$, and $\boldsymbol{\Omega}$ is a $d \times d$ positive definite dispersion matrix. To ease the discussion we will refer to $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ as *mean vector* and *variance-covariance matrix* respectively, although they are not moments of the distribution. For the distributions derived from the logistic normal parent distribution, the parameters $\boldsymbol{\mu}_{W_\ell}$ and $\boldsymbol{\Omega}_{W_\ell}$ are defined in terms of the parameters $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$, and the set of indices of nonzero components of $\mathbf{x}$, W_ℓ, and a selection matrix $\mathbf{B}_{W_\ell}$.

Let $W_\ell \subset \{1, 2, 3, \ldots, d\}$ be a nonempty set of indices (of the nonzero components of $\mathbf{x}$); without loss of generality we can order the indices from least to greatest

$$W_\ell = \{j_1, j_2, \ldots, j_J\} \text{ where } 0 < j_1 < j_2 < \ldots < j_J \leq d.$$

Now we define our $J \times d$ selection matrix, $\mathbf{B}_{W_\ell} = [B_{i,m}]$. For $i \in \{1, 2, \ldots, J\}$, and $m \in \{1, 2, \ldots, d\}$, with $W_\ell = \{j_1, j_2, \ldots, j_J\}$, we define the elements of $[B_{i,m}]$ to be $B_{i,j_i} = 1$ and $B_{i,m\neq j_i} = 0$. For example, let $\mathbf{x} = (.2, 0, .3, 0, .25, .25)$, a 6-part composition, with $x_6 > 0$. The set of nonzero indices is $W_\ell = \{1, 3, 5\}$, and the selection matrix is

$$\mathbf{B}_{W_\ell} = \mathbf{B}_{\{1,3,5\}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Now it is easy to define $\boldsymbol{\mu}_{W_\ell}$ and $\boldsymbol{\Omega}_{W_\ell}$. We define:

$$\boldsymbol{\mu}_{W_\ell} = (\mathbf{B}_{W_\ell})(\boldsymbol{\mu}).$$

$$\boldsymbol{\Omega}_{W_\ell} = (\mathbf{B}_{W_\ell})(\boldsymbol{\Omega})(\mathbf{B}_{W_\ell}^T).$$

With this structure, the mixture distribution can be more fully specified.

$$g(\mathbf{x}; \boldsymbol{\mu}, \Omega) = \sum_{\ell=1}^{2^d-1} P(W_\ell)\mathcal{L}^{||W_\ell||}(\mathbf{x}; \boldsymbol{\mu}_{W_\ell}, \boldsymbol{\Omega}_{W_\ell}) \text{ where}$$

- $||W_\ell||$ refers to the cardinality of the set W_ℓ.
- $\sum P(W_\ell) = 1$.
- $\boldsymbol{\mu}$ is a d-part vector in $\mathbb{R}^d$.
- $\boldsymbol{\mu}_{W_\ell}$ is a subvector of $\boldsymbol{\mu}$ corresponding to the W_ℓ pattern of zeros.
- $\boldsymbol{\Omega}_{W_\ell}$ is a submatrix of a $d \times d$ positive definite covariance matrix corresponding to the W_ℓ pattern of zeros.

5.1. Multivariate normal foundation

Now we extend the notation for the inverse of the additive logratio transformation, alr^{-1}, from Aitchison (1986). We use the new symbols, $\tilde{\text{alr}}$ and $\tilde{\text{alr}}^{-1}$. We define them in terms of W and D, the maximum index. Let $W \subset \{1, 2, \ldots, d\}$ be a pattern of zeros, i.e., a set of indices of nonzero components of $\mathbf{x}$ and denote them: $W = \{i_1, i_2, \ldots, i_r\}$, and let $j \in \{1, 2, \ldots, d, D\}$.

In our approach there is a tight correspondence between the y_i variables of a multivariate normal vector and the x_i parts of a composition, possibly one containing essential zeros.

$$y_i = \log(x_i/x_D) \text{ for } i = 1, 2, 3, \ldots, d.$$

$$\begin{aligned} \text{Composition: } \mathbf{x} &= (x_1, x_2, x_3, \ldots, x_d, x_D)^T \\ &\quad | \quad | \quad | \qquad | \\ \text{alr transformed vector: } \mathbf{y} &= \log(\mathbf{x}_{-D}/x_D) = (y_1, y_2, y_3, \ldots, y_d, \odot)^T. \end{aligned}$$

When there is an essential zero in the composition in one of the x_i parts, e.g., in x_2, we use $\odot$ as a placeholder so things line up, for example:

$$\begin{aligned} \text{Composition: } \mathbf{x} &= (x_1, 0, x_3, \ldots, x_d, x_D)^T \\ &\quad | \qquad | \qquad | \\ \tilde{\text{alr}} \text{ transformed subvector: } \mathbf{y} &= \log(\mathbf{x}_{-\{2,D\}}/x_D) = (y_1, \odot, y_3, \ldots, y_d, \odot)^T. \end{aligned} \tag{3}$$

When we have an essential zero in the i^{th} part of the composition, we have a selected subvector of the y's which does not contain an element corresponding to y_i. The requirement that $x_D > 0$ is what allows us to maintain this strict correspondence between x_i and y_i.

Now we define $\tilde{\text{alr}}^{-1}$

$$\tilde{\text{alr}}^{-1}(\mathbf{y}, W, D) = (x_1, \ldots, x_d, x_D)^T \text{ where,} \tag{4}$$

$$x_j = \begin{cases} \exp(y_j)/\{\exp(y_{i_1}) + \exp(y_{i_2}) + \ldots + \exp(y_{i_r}) + 1\} & \text{if } j \in W \\ 0 & \text{if } j \notin W \ \& \ j \in \{1, \ldots, d\} \\ 1/\{\exp(y_{i_1}) + \exp(y_{i_2}) + \ldots + \exp(y_{i_r}) + 1\} & \text{if } j = D. \end{cases}$$

Next we turn to defining an extension to the logistic normal distribution. Let $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_d, x_D)^T$ be a composition, and let $\mathbf{y} = \log(\mathbf{x}_{-D}/x_D) = (y_1, y_2, y_3, \ldots, y_d)^T$. Then $\mathbf{x}$ is defined to have a logistic normal distribution with location $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\Omega}$, written $\mathbf{x} \sim \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, if $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Note also that if $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, and $\mathbf{B}$ is a selection matrix as mentioned in Section 5, then $\mathbf{By}$ is also multivariate normal with distribution:

$$\mathbf{By} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T).$$

And finally, recall that also in Section 5 we used W_ℓ to represent the set of indices of the nonzero components of $\mathbf{x}$. $\mathbf{B}_{W_\ell}$ is the selection matrix that selects those components when we perform these matrix multiplications:

$$\boldsymbol{\mu}_{W_\ell} = (\mathbf{B}_{W_\ell})(\boldsymbol{\mu}),$$

$$\boldsymbol{\Omega}_{W_\ell} = (\mathbf{B}_{W_\ell})(\boldsymbol{\Omega})(\mathbf{B}_{W_\ell}^T).$$

With all these definitions in place we are now in a position to define a logistic normal distribution with essential zeros.

Definition:

Let $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_d, x_D)^T$ be a composition with $\mathbf{x}_D > 0$.

Let $W_\ell = \{i_1, i_2, \ldots, i_r\} \subset \{1, 2, \ldots, d\}$ be a nonempty set of indices of nonzero components of $\mathbf{x}$.

Let $\mathbf{B}_{W_\ell}$ be the corresponding selection matrix.

Let $\mathbf{y} = \log(\mathbf{B}_{W_\ell}\mathbf{x}_{-D}/x_D) = (y_{i_1}, y_{i_2}, \ldots, y_{i_r})^T = \tilde{\text{alr}}(\mathbf{x}, W_\ell, D)$ be the logratios of the nonzero components of $\mathbf{x}$.

If for every set W_ℓ of indices of nonzero components of $\mathbf{x}$, we have $\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu}_{W_\ell}, \boldsymbol{\Omega}_{W_\ell})$, then $\mathbf{x}$ has a logistic normal distribution with essential zeros, written $\mathbf{x}\sim\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, with probability density function

$$g(\mathbf{x}; \boldsymbol{\mu}, \Omega) = \sum_{\ell=1}^{2^d-1} P(W_\ell)\mathcal{L}^{||W_\ell||}(\mathbf{x}; \boldsymbol{\mu}_{W_\ell}, \boldsymbol{\Omega}_{W_\ell}),$$

where

$\sum P(W_\ell) = 1.$

$\boldsymbol{\mu}$ is a d-part vector in R^d.

$\boldsymbol{\mu}_{W_\ell}$ is a subvector of $\boldsymbol{\mu}$ corresponding to the W_ℓ pattern of zeros.

$\boldsymbol{\Omega}$ is a $d \times d$ positive definite covariance matrix.

$\boldsymbol{\Omega}_{W_\ell}$ is a square submatrix of $\boldsymbol{\Omega}$, corresponding to the W_ℓ pattern of zeros.

For the case where $W = \{1, 2, \ldots, d\}$ the composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T$ has the additive logistic normal distribution, $\mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

5.2. Common expectations and variances

The definition of $\tilde{\text{alr}}^{-1}$ enables compositions from different subdistributions to be used to estimate parameters of their shared parent distribution. Let $\mathbf{x}_1 = (x_{11}, x_{21}, \ldots, x_{D1})^T$, and let $\mathbf{x}_2 = (x_{12}, x_{22}, \ldots, x_{D2})^T$ with

$$\mathbf{x}_1\sim\mathcal{L}^{||W_1||}(\boldsymbol{\mu}_{W_1}, \boldsymbol{\Omega}_{W_1}), \text{ and} \tag{5}$$

$$\mathbf{x}_2\sim\mathcal{L}^{||W_2||}(\boldsymbol{\mu}_{W_2}, \boldsymbol{\Omega}_{W_2}). \tag{6}$$

The two sets of nonzero indices, W_1, W_2 need not have any elements in common, nor do they need to have the same number of elements, though $\mathbf{x}_1$ and $\mathbf{x}_2$ both have D elements. Suppose they have an index, m, in common: $m \in W_1 \cap W_2$. By properties of the logistic normal distribution (Aitchison 1986, p. 116), and the definition of $\tilde{\text{alr}}^{-1}$ in Equation 4 we have:

$$E\log(x_{m1}/x_{D1}) = Ey_m = \mu_m = Ey_m = E\log(x_{m2}/x_{D2}). \tag{7}$$

And similarly,

$$\text{Var}[\log(x_{m1}/x_{D1})] = \text{Var}[y_m] = \sigma_m^2 = \text{Var}[y_m] = \text{Var}[\log(x_{m2}/x_{D2})]. \tag{8}$$

Thus, compositions from different subdistributions of the same logistic normal distribution can be used to estimate the parameters of their shared parent distribution.

6. Data blocks

Now that we have a correspondence between multivariate normal variables and compositions with zeros, we could derive a density function using the standard formula for transformed variables, analogous to Aitchison (1986, chapter 6). However, for estimating parameters it is more convenient to work in the space of the transformed variables (multivariate normal projections).

Here we apply the techniques and notation of block matrices and matrix calculus to do some preparation in order to build a likelihood and attack the problem of finding estimators for the

parameters. We discuss two sets of estimators, a general maximum likelihood estimator, and a simpler pair of estimators reminiscent of method of moments estimators.

6.1. Block matrices of compositions

We write a collection of compositional data with zeros, $\mathbf{X}$, as a column of blocks of compositions where each block, $\mathbf{X}_\ell$, has a particular pattern of zeros throughout. That is, for a particular block, $\mathbf{X}_\ell$, and $i \in \{1, 2, \ldots, d\}$, the i^{th} column of $\mathbf{X}_\ell$ is either all positive, or all zero. Let

$$\underset{n\times D}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_b \end{bmatrix}. \tag{9}$$

The dimensions of the blocks are: $\underset{r_1\times D}{\mathbf{X}_1}, \underset{r_2\times D}{\mathbf{X}_2}, \ldots, \underset{r_b\times D}{\mathbf{X}_b}$ and the sum of their vertical dimensions is $r_1 + r_2 + \ldots + r_b = n$, where n is the number of data points.

We use ℓ to indicate a block, and t to indicate a composition (row) in that block. Next we define the patterns of zeros in each block. Here $i \in \{1, 2, \ldots, D\}$. For $\ell \in \{1, 2, \ldots, b\}$, let $W_\ell \subset \{1, 2, \ldots, d\}$ be the set of indices of strictly positive components of $\mathbf{X}_\ell$. For $\ell \in \{1, 2, \ldots, b\}$,

$$\mathbf{X}_\ell = [x_{ti}], \text{ where } \begin{cases} x_{ti} > 0 \text{ if } i = D, \\ x_{ti} > 0 \text{ if } i \in W_\ell, \\ x_{ti} = 0 \text{ if } i \notin W_\ell \text{ and } i \neq D. \end{cases} \tag{10}$$

6.2. Transformations - ratios and logratios

We have already defined the alr transformation for the case where there are no zeros in (1). Next we extend alr to $\tilde{\text{alr}}$ for a block matrix of compositions, $\underset{r_\ell\times D}{\mathbf{X}}$ which may contain zeros. We do this by defining a selection matrix $\mathbf{B}_{W_\ell}$ corresponding to set W_ℓ. We still have $W_\ell \subset \{1, 2, 3, \ldots, d\}$ being a nonempty set of indices of the nonzero components of $\mathbf{x}$, and without loss of generality we can order the indices from least to greatest:

$$W_\ell = \{j_1, j_2, \ldots, j_J\} \text{ where } 0 < j_1 < j_2 < \ldots < j_J < D. \tag{11}$$

Now we define our $(J + 1) \times D$ selection matrix, $\mathbf{B}_{W_\ell} = [B_{p,m}]$. We use $J + 1$ here because we construct the selection matrix so that the final, D^{th}, component of the data is always selected. This is slightly different than before. Previously we constructed B to conform to the parameters $\boldsymbol{\mu}(d \times 1)$ and $\boldsymbol{\Omega}(d \times d)$.

$$\text{For } p \in \{1, 2, \ldots, J + 1\}, \text{ and } m \in \{1, 2, \ldots, D\}, \text{ with } W_\ell = \{j_1, j_2, \ldots, j_J\}, \tag{12}$$

$$\text{we define the elements of } [B_{p,m}] \text{ to be } B_{p,j_p} = 1 \text{ and } B_{p,m\neq j_p} = 0. \tag{13}$$

$\mathbf{X}_\ell \mathbf{B}_{W_\ell}^T$ is a matrix where each row vector is a composition without zeros.

$$\mathbf{X}_\ell \mathbf{B}_{W_\ell}^T = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_{r_\ell^T} \end{bmatrix} \mathbf{B}_{W_\ell}^T = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1(J+1)} \\ x_{21} & x_{22} & \ldots & x_{2(J+1)} \\ \vdots & \vdots & & \vdots \\ x_{r_\ell 1} & x_{r_\ell 2} & \ldots & x_{r_\ell(J+1)} \end{bmatrix}. \tag{14}$$

$$\text{We define } \tilde{\text{alr}}(\mathbf{X}_\ell, W_\ell, D) = \text{alr}(\mathbf{X}_\ell \mathbf{B}_{W_\ell}^T) = \underset{r_\ell \times (J+1)}{\begin{bmatrix} \text{alr}(\mathbf{x}_1^T) \\ \text{alr}(\mathbf{x}_2^T) \\ \vdots \\ \text{alr}(\mathbf{x}_{r_\ell}^T) \end{bmatrix}} = \underset{r_\ell \times J}{\begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_{r_\ell}^T \end{bmatrix}}. \tag{15}$$

$$\text{Let } \underset{r_\ell \times J}{\mathbf{Y}_\ell} = \tilde{\text{alr}}(\mathbf{X}_\ell, W_\ell, D) \tag{16}$$

Each row vector in $\mathbf{Y}_\ell$ is a vector of reals, all potentially from the multivariate normal distribution corresponding to the ℓ^{th} pattern of zeros. Note that we cannot form a single block matrix, $\mathbf{Y}$, from the collection of $\mathbf{Y}_\ell$ because they can have different numbers of columns.

6.3. Illustration - spices, lentils, and rice

In our example about compositions of spending on spices, lentils, and rice (Table 1), there are three patterns of zeros. Tables 2-4 show the result of applying the $\tilde{\text{alr}}$ transformation.

$\mathbf{X}_1$ corresponds to rows 1-3 and its set of indices is $W_1 = \{1\}$.
$\mathbf{X}_2$ corresponds to rows 4-6 and its set of indices is $W_2 = \{2\}$.
$\mathbf{X}_3$ corresponds to rows 7-12 and its set of indices is $W_3 = \{1, 2\}$.

Table 2: $\mathbf{Y}_1 = \tilde{\text{alr}}(\mathbf{X}_1, \{1\}, 3)$

	log(spices/rice)
1	-1.66
2	-1.59
3	-1.68

Table 3: $\mathbf{Y}_2 = \tilde{\text{alr}}(\mathbf{X}_2, \{2\}, 3)$

	log(lentils/rice)
4	-0.52
5	-0.52
6	-0.51

Table 4: $\mathbf{Y}_3 = \tilde{\text{alr}}(\mathbf{X}_3, \{1, 2\}, 3)$

	log(spices/rice)	log(lentils/rice)
7	-1.56	-0.52
8	-1.64	-0.51
9	-1.52	-0.57
10	-1.72	-0.51
11	-1.77	-0.53
12	-1.58	-0.51

6.4. Means

The matrix $\mathbf{Y}_\ell$ contains rows of compositions with the same pattern of zeros. We refer to the t^{th} row vector of $\mathbf{Y}_\ell$ as $\mathbf{y}_{\ell t}^T$. We refer to the mean as the vector $\bar{\mathbf{y}}_\ell$, and define it as:

$$\underset{J \times 1}{\bar{\mathbf{y}}_\ell} = \frac{1}{r_\ell} (\mathbf{1}_{r_\ell}^T \mathbf{Y}_\ell)^T. \tag{17}$$

Here we are using $\mathbf{1}_{r_\ell}$ to represent an $r_\ell \times 1$ column vector of ones. $\bar{\mathbf{y}}_\ell$ is a column vector. We define it this way because of how we intend to use it in quadratic forms from the multivariate normal density.

7. Simple estimators

7.1. Mean

It is also possible to construct simpler estimators relying on properties of the normal distribution. For the location, if $\underset{n\times D}{\mathbf{X}} = [x_{ti}]$ is a collection of n compositional data points with zeros, and the D^{th} component always strictly positive, we can define a simple estimator of the mean, $\overset{*}{\boldsymbol{\mu}} = (\overset{*}{\mu}_1, \overset{*}{\mu}_2, \ldots, \overset{*}{\mu}_d)^T$. Let n_i be the number of elements of the i^{th} column of $\mathbf{X}$ that are nonzero. For $i \in \{1, 2, \ldots, d\}$, and $t \in \{1, 2, \ldots, n\}$, define

$$\overset{*}{\mu}_i = \frac{1}{n_i} \sum_{\{t:x_{ti}\neq 0\}} \log(x_{ti}/x_{tD}). \tag{18}$$

By the assumption of normality of the logratios, the estimator $\overset{*}{\boldsymbol{\mu}}$ is unbiased.

In the spices-lentils-rice example, $\overset{*}{\mu}$ = (log(spices/rice): -1.635 , log(lentils/rice): -0.523). For ease of interpretation, we convert the estimator back to a composition with the alr^{-1} transformation giving: (spices: 0.109 , lentils: 0.332, rice: 0.559). That is, our estimate of Bill's mean expenditure is 10.9% on spices , 33.2% on lentils, and 55.9% on rice.

7.2. Variance

Here we show how to find estimators for variances and covariances using maximum likelihood estimators for normal random variables. For a single random composition, $\mathbf{x}$, with components $x_1, x_2, \ldots, x_D$, we substitute $\log(x_i/x_D)$ into the MLE for variances of normal random variables. We use $\overset{*}{\sigma}{}^2_{ii}$ for the estimator of the variances of the logratios $\log(x_i/x_D)$, for $i \in \{1, 2, \ldots, d\}$, and $t \in \{1, 2, \ldots, n_i\}$.

$$\overset{*}{\sigma}{}^2_{ii} = \frac{1}{n_i} \sum_{\{t:x_{ti}\neq 0\}} \left(\log(x_{ti}/x_{tD}) - \overset{*}{\mu}_i\right)^2. \tag{19}$$

If we want an unbiased estimator, we can divide by $(n_i - 1)$ instead of n_i. As with means, the different $\overset{*}{\sigma}_{ii}$ are based on different numbers of observations, n_i.

7.3. Covariance

It only makes sense to talk about estimating the covariance of the variables $\log(x_i/x_D)$ and $\log(x_j/x_D)$ when both x_{ti} and x_{tj} are not 0 so we define $n_{ij} = ||\{t : x_{ti} \neq 0 \;\&\; x_{tj} \neq 0\}||$. That is, n_{ij} is the number of data points where both x_{ti} and x_{tj} are not 0. As we did with variance, we can start with the canonical maximum likelihood formula for estimating covariance among normally distributed variables, and substitute in appropriate logratios.

$$\overset{*}{\sigma}_{ij} = \frac{1}{n_{ij}} \sum_{\{t:x_{ti}\neq 0 \;\&\; x_{tj}\neq 0\}} (\log(x_{ti}/x_{tD}) - \overset{*}{\mu}_i)(\log(x_{tj}/x_{tD}) - \overset{*}{\mu}_j) \tag{20}$$

Note that $\overset{*}{\sigma}_{ij}$ is based on n_{ij} observations, while $\overset{*}{\mu}_i$ and $\overset{*}{\mu}_j$ are based on n_i and n_j observations, respectively. The formula in Equation 20 is based on the maximum likelihood estimator for covariance of normal variables. For unbiased estimators we would divide by $(n_{ij} - 1)$ instead of n_{ij}.

Our estimator for the $d \times d$ variance-covariance matrix is $\overset{*}{\boldsymbol{\Omega}} = [\overset{*}{\sigma}_{ij}]$. There are two potential problems with this approach. There could be $i, j, i \neq j$, such that whenever $x_i > 0$, $x_j = 0$. In that case we cannot estimate the covariance. Also, irrespective of that, the estimate of the variance-covariance matrix, $\overset{*}{\boldsymbol{\Omega}}$, might not be positive definite.

In the spices-lentils-rice example,

$$\overset{*}{\boldsymbol{\Omega}} = \begin{pmatrix} 0.00648 & -0.00096 \\ -0.00096 & 0.00035 \end{pmatrix}$$

which is positive definite.

8. Maximum likelihood estimators

For the case where there are no zeros, the location estimator described earlier is a maximum likelihood estimator (MLE), but in general the estimator we found earlier is not an MLE. From now on we will call that estimator the simple estimator, to contrast it with the MLE, which we derive next.

We start by finding the location MLE given $\boldsymbol{\Omega}$ for 3-part compositions, show it is unbiased, and then show the relative efficiency of the simple estimator with to the MLE. Assume we have a set of logistic normal compositional data with b different patterns of zeros as in (9).

$$\begin{aligned}
&\mathbf{x}_{11}, \ldots, \mathbf{x}_{1r_1} \overset{i.i.d.}{\sim} \mathcal{L}^{||W_1||}(\mathbf{B}_{W_1}\boldsymbol{\mu}, \mathbf{B}_{W_1}\boldsymbol{\Omega}\mathbf{B}_{W_1}^T) && \text{(rows of } \mathbf{X}_1\text{)} \\
&\mathbf{x}_{21}, \ldots, \mathbf{x}_{2r_2} \overset{i.i.d.}{\sim} \mathcal{L}^{||W_2||}(\mathbf{B}_{W_2}\boldsymbol{\mu}, \mathbf{B}_{W_2}\boldsymbol{\Omega}\mathbf{B}_{W_2}^T) && \text{(rows of } \mathbf{X}_2\text{)} \\
&\qquad\vdots \\
&\mathbf{x}_{b1}, \ldots, \mathbf{x}_{br_b} \overset{i.i.d.}{\sim} \mathcal{L}^{||W_b||}(\mathbf{B}_{W_b}\boldsymbol{\mu}, \mathbf{B}_{W_b}\boldsymbol{\Omega}\mathbf{B}_{W_b}^T). && \text{(rows of } \mathbf{X}_b\text{)}
\end{aligned} \tag{21}$$

In a block of data, as in (21), we use $\mathbf{x}_{\ell t}$ to refer to the t^{th} compositional observation with W_ℓ pattern of zeros. We define $\mathbf{y}_{\ell t} = \tilde{\text{alr}}(\mathbf{x}_{\ell t}, W_\ell, D)$, and to ease notation, we write in terms of $\mathbf{y}_{\ell t}$.

8.1. Likelihood

First we write the full likelihood and log likelihood for D-part compositions, and then restrict ourselves to 3-part compositions. The full likelihood is:

$$\begin{aligned}
&L(\boldsymbol{\mu}, \boldsymbol{\Omega} | r_1, \ldots, r_b, \mathbf{y}_{11}, \ldots, \mathbf{y}_{br_b}) = \\
&\prod_{\ell=1}^{b} \prod_{t=1}^{r_\ell} \frac{P(W_\ell)}{(2\pi)^{||W_\ell||/2} |\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T|^{1/2}} \exp\big[-\frac{1}{2}(\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})^T (\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T)^{-1}(\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})\big].
\end{aligned} \tag{22}$$

The constant

$$\prod_{\ell=1}^{b} \prod_{t=1}^{r_\ell} \frac{P(W_\ell)}{(2\pi)^{||W_\ell||/2} |\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T|^{1/2}} \tag{23}$$

is independent of $\boldsymbol{\mu}$, so for purposes of maximizing the likelihood with respect to $\boldsymbol{\mu}$, we can treat it as a single constant, C.

$$\begin{aligned}
&L(\boldsymbol{\mu}, \boldsymbol{\Omega} | r_1, \ldots, r_b, \mathbf{y}_{11}, \ldots, \mathbf{y}_{br_b}) \\
&= C \prod_{\ell=1}^{b} \prod_{t=1}^{r_\ell} \exp\big[-\frac{1}{2}(\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})^T (\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T)^{-1}(\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})\big]
\end{aligned} \tag{24}$$

$$= C \exp\big[-\frac{1}{2} \sum_{\ell=1}^{b} \sum_{t=1}^{r_\ell} (\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})^T (\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T)^{-1} (\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})\big]. \tag{25}$$

Taking the log gives:

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Omega} | r_1, \ldots, r_b, \mathbf{y}_{11}, \ldots, \mathbf{y}_{br_b}) \tag{26}$$

$$= \log C - \frac{1}{2} \sum_{\ell=1}^{b} \sum_{t=1}^{r_\ell} (\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu})^T (\mathbf{B}_{W_\ell}\boldsymbol{\Omega}\mathbf{B}_{W_\ell}^T)^{-1} (\mathbf{y}_{\ell t} - \mathbf{B}_{W_\ell}\boldsymbol{\mu}). \tag{27}$$

For the simple case of three-part compositional data with some zeros in component one, and some zeros in component two, the parent distribution for the transformed data is bivariate normal,

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}) \text{ where } \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ and } \boldsymbol{\Omega} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} \tag{28}$$

For the full bivariate normal distribution,

$$A = \boldsymbol{\Omega}^{-1} = \frac{1}{s_{11}s_{22} - s_{12}^2} \begin{bmatrix} s_{22} & -s_{12} \\ -s_{12} & s_{11} \end{bmatrix}. \tag{29}$$

For the two univariate normal distributions, the inverses of the variances are: $\frac{1}{s_{11}}$ and $\frac{1}{s_{22}}$.
In these formulas,
y_{1j1} is the j^{th} data point among the univariate data from the first component.
y_{2j2} is the j^{th} data point among the univariate data from the second component.
$\mathbf{y}_{3j}$ is a 2-part vector with data from both components,

$$\begin{bmatrix} y_{1j1} \\ \odot \end{bmatrix} \qquad \begin{bmatrix} \odot \\ y_{2j2} \end{bmatrix} \qquad \begin{bmatrix} y_{3j1} \\ y_{3j2} \end{bmatrix}. \tag{30}$$

In the example, the y_{1j1} correspond to elements of Table 2, log(spices/rice). The y_{2j2} correspond to elements of Table 3, log(lentils/rice), and y_{3j1} correspond to elements of Table 4, both log(spices/rice) and log(lentils/rice).

We define the means of these matrices in the usual way.

$$\frac{1}{r_1}\sum_{j=1}^{r_1} y_{1j1} = \bar{y}_{11} \qquad \frac{1}{r_2}\sum_{j=1}^{r_2} y_{2j2} = \bar{y}_{22} \qquad \frac{1}{r_3}\sum_{j=1}^{r_3} \mathbf{y}_{3j} = \begin{bmatrix} \bar{y}_{31} \\ \bar{y}_{32} \end{bmatrix} \tag{31}$$

Partial derivatives

$$\frac{\partial \log L(\boldsymbol{\mu}|\mathbf{Y}, \boldsymbol{\Omega}, r_1, r_2, r_3)}{\partial \mu_1} =$$
$$\frac{1}{s_{11}} r_1(\bar{y}_{11} - \mu_1) + \frac{s_{22}}{s_{11}s_{22} - s_{12}^2} r_3(\bar{y}_{31} - \mu_1) + \frac{-s_{12}}{s_{11}s_{22} - s_{12}^2} r_3(\bar{y}_{32} - \mu_2) \tag{32}$$

$$\frac{\partial \log L(\boldsymbol{\mu}|\mathbf{Y}, \boldsymbol{\Omega}, r_1, r_2, r_3)}{\partial \mu_2} =$$
$$\frac{1}{s_{22}} r_2(\bar{y}_{22} - \mu_2) + \frac{s_{11}}{s_{11}s_{22} - s_{12}^2} r_3(\bar{y}_{32} - \mu_2) + \frac{-s_{12}}{s_{11}s_{22} - s_{12}^2} r_3(\bar{y}_{31} - \mu_1) \tag{33}$$

MLE for location, given $\mathbf{\Omega}$

We set the partial derivatives equal to zero, replace μ with $\hat{\mu}$, and solve. The result is:

$$\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3 = \frac{(r_1\bar{y}_{11} + r_3\bar{y}_{31})(r_2 + r_3)s_{11}s_{12} - r_1\bar{y}_{11}r_2s_{12}^2 + (\bar{y}_{22} - \bar{y}_{32})r_2r_3s_{11}s_{12}}{(r_1 + r_3)(r_2 + r_3)s_{11}s_{22} - r_1r_2s_{12}^2} \tag{34}$$

$$\hat{\mu}_2|\mathbf{\Omega}, r_1, r_2, r_3 = \frac{(r_2\bar{y}_{22} + r_3\bar{y}_{32})(r_1 + r_3)s_{11}s_{12} - r_2\bar{y}_{22}r_1s_{12}^2 + (\bar{y}_{11} - \bar{y}_{31})r_1r_3s_{12}s_{22}}{(r_1 + r_3)(r_2 + r_3)s_{11}s_{22} - r_1r_2s_{12}^2} \tag{35}$$

In the case where there are no univariate data from the second component, i.e, $r_2 = 0$, we have:

$$(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3)\Big|_{r_2=0} = \frac{r_1\bar{y}_{11} + r_3\bar{y}_{31}}{(r_3 + r_1)} = \frac{1}{(r_3 + r_1)}\left[\sum_{j=1}^{r_3} y_{3j1} + \sum_{j=1}^{r_1} y_{1j1}\right]. \tag{36}$$

That shows that when we have $r_2 = 0$, the MLE $(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3)$ is equal to our simple estimator for μ_1. Similarly, when $r_1 = 0$, $(\hat{\mu}_2|\mathbf{\Omega}, r_1, r_2, r_3)$ is equal to our simple estimator for μ_2. It also turns out that $(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3)\Big|_{r_3=0} = \bar{y}_{11}$, and when $r_3 = 0$, the simple estimator is also $\bar{y}_{11}$, so they are equal in that case as well.

8.2. Unbiasedness of conditional MLE for 3-part composition

To show that $\hat{\boldsymbol{\mu}}|\mathbf{\Omega}, r_1, r_2, r_3$ is unbiased, we start by pointing out the expectations of the various means:

$$E[\bar{y}_{11}] = E\left[\frac{1}{r_1}\sum_{j=1}^{r_1} y_{1j1}\right] = \frac{1}{r_1}\sum_{j=1}^{r_1} E\left[y_{1j1}\right] = \mu_1 \tag{37}$$

$$E[\bar{y}_{22}] = E\left[\frac{1}{r_2}\sum_{j=1}^{r_2} y_{2j2}\right] = \frac{1}{r_2}\sum_{j=1}^{r_2} E\left[y_{2j2}\right] = \mu_2 \tag{38}$$

$$E\left[\begin{bmatrix} \bar{y}_{31} \\ \bar{y}_{32} \end{bmatrix}\right] = E\left[\frac{1}{r_3}\sum_{j=1}^{r_3} \mathbf{y}_{3j}\right] = \frac{1}{r_3}\sum_{j=1}^{r_3} E\left[\mathbf{y}_{3j}\right] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{39}$$

When we take the expectation in expression(34), the term with $(\bar{y}_{22} - \bar{y}_{32})$ vanishes because $E[\bar{y}_{22}] = E[\bar{y}_{32}]$. That leaves only terms with $E[\bar{y}_{11}] = \mu_1$ and $E[\bar{y}_{31}] = \mu_1$, which we can factor:

$$E\left[\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_2\right] = \frac{\mu_1\left[(r_1 + r_3)(r_2 + r_3)s_{11}s_{12} - r_1r_2s_{12}^2\right]}{(r_1 + r_3)(r_2 + r_3)s_{11}s_{22} - r_1r_2s_{12}^2} = \mu_1 \tag{40}$$

This shows that $\hat{\mu}_1$ is unbiased. By symmetry we get that $\hat{\mu}_2$ is unbiased.

8.3. General maximum likelihood estimators

For the general case of MLE for higher dimensions than shown here, the log likelihood can be differentiated, and the score functions can be solved with a computer algebra system. In addition, the Hessian can be checked to verify the solution is a maximum. We have done this for the case of 3-part compositions and do not anticipate any obstacles to extending the program to handle the general case of D-dimensions.

9. Variances of location estimators

Next we find variances of the two location estimators, the MLE, and the simple estimator. Both are unbiased. A question we need to answer is, what is the efficiency of the simple estimator relative to the MLE. We have been using $\hat{\boldsymbol{\mu}}$ for the MLE. We continue to use $\overset{*}{\boldsymbol{\mu}}$ for the simple estimator (of the location). In our discussion,

$$\text{efficiency}(\overset{*}{\mu}_1, \hat{\mu}_1) = \frac{\text{Var}(\hat{\mu}_1)}{\text{Var}(\overset{*}{\mu}_1)}. \tag{41}$$

9.1. Variances of location estimators

The variances of the MLE and the simple location estimator are derived in the Appendix. They are:

$$\text{Var}(\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3) = \frac{r_3\left((r_3+r_2)^2 s_{11}^3 s_{22}^2 + r_2^2 s_{11}^2 s_{12}^2 s_{22} - 2r_2(r_3+r_2)s_{11}^2 s_{12}^2 s_{22}\right) + r_2 r_3^2 s_{11}^2 s_{12}^2 s_{22} + r_1 s_{11}((r_3+r_2)s_{11}s_{22} - r_2 s_{12}^2)^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)^2} \tag{42}$$

$$\text{Var}(\hat{\mu}_2|\boldsymbol{\Omega}, r_1, r_2, r_3) = \frac{r_3\left((r_3+r_1)^2 s_{22}^3 s_{11}^2 + r_1^2 s_{22}^2 s_{12}^2 s_{11} - 2r_1(r_3+r_1)s_{22}^2 s_{12}^2 s_{11}\right) + r_1 r_3^2 s_{22}^2 s_{12}^2 s_{11} + r_2 s_{22}((r_3+r_1)s_{22}s_{11} - r_1 s_{12}^2)^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)^2} \tag{43}$$

$$\text{Var}(\overset{*}{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3) = \frac{s_{11}}{r_1 + r_3}. \tag{44}$$

$$\text{Var}(\overset{*}{\mu}_2|\boldsymbol{\Omega}, r_1, r_2, r_3) = \frac{s_{22}}{r_2 + r_3}. \tag{45}$$

9.2. Relative efficiency of location estimators

The first thing we show is that when the covariance element of $\boldsymbol{\Omega}$ is zero, i.e, $s_{12} = 0$, then $\text{Var}(\hat{\mu}) = \text{Var}(\overset{*}{\mu})$.

$$\text{Var}(\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3) = \frac{r_3\left((r_3+r_2)^2 s_{11}^3 s_{22}^2 + r_2^2 s_{11}^2 s_{12}^2 s_{22} - 2r_2(r_3+r_2)s_{11}^2 s_{12}^2 s_{22}\right) + r_2 r_3^2 s_{11}^2 s_{12}^2 s_{22} + r_1 s_{11}((r_3+r_2)s_{11}s_{22} - r_2 s_{12}^2)^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)^2} \tag{46}$$

Evaluate at $s_{12} = 0$.

$$\text{Var}(\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3)\Big|_{s_{12}=0} = \frac{r_3\left((r_3+r_2)^2 s_{11}^3 s_{22}^2\right) + r_1 s_{11}((r_3+r_2)s_{11}s_{22})^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22})^2} \tag{47}$$

Factor numerator and denominator.

$$\text{Var}(\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3)\Big|_{s_{12}=0} = \frac{(r_3+r_2)^2 s_{11}^3 s_{22}^2 (r_3+r1)}{(r_3+r_1)^2(r_3+r_2)^2 s_{11}^2 s_{22}^2} = \frac{s_{11}}{r_3+r_1} = \text{Var}(\overset{*}{\mu}_1). \tag{48}$$

Similarly,

$$\text{Var}(\hat{\mu}_2|\boldsymbol{\Omega}, r_1, r_2, r_3)\Big|_{s_{12}=0} = \frac{s_{22}}{r_3+r_2} = \text{Var}(\overset{*}{\mu}_2). \tag{49}$$

We have already shown in Section 8.1.2 that when $r_2 = 0, \hat{\mu}_1 = \overset{*}{\mu}_1$, and when $r_1 = 0, \hat{\mu}_2 = \overset{*}{\mu}_2$; and when $r_3 = 0$, $\hat{\mu}_1 = \overset{*}{\mu}_1$, and $\hat{\mu}_2 = \overset{*}{\mu}_2$. Next we need to compare the variance of $\overset{*}{\mu}$ with the variance of $\hat{\mu}$ in cases where the estimators are not obviously the same.

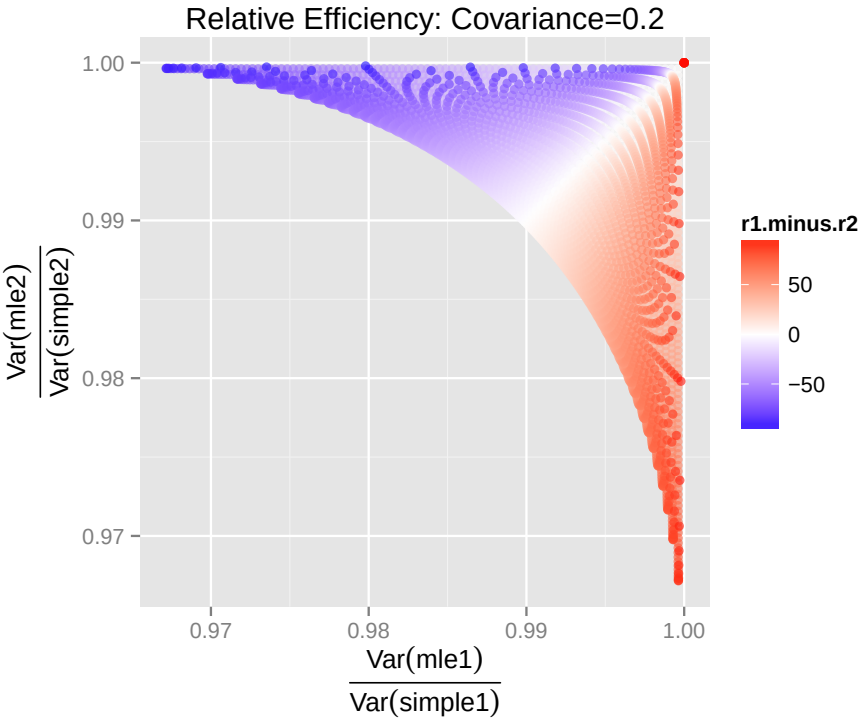

Figure 1: Efficiency of $\overset{*}{\mu}$ relative to $\hat{\mu}$ with low covariance (0.2)

We consider a sample of 100 compositions from a logistic normal distribution with the number of zeros in part 1 ranging from 0 to 100, and similarly for part 2. We calculate the relative efficiency. These are not simulations; they are calculations based on the expressions for the variances of the estimators. We consider all possible combinations of r_1, r_2, r_3 such that $r_1 + r_2 + r_3 = 100$. A larger sample would give roughly the same picture, just with finer granularity. In addition, while we want to understand the effect of the covariance term s_{12} for every possible value between -1 and 1, we get a feel for the space by choosing three values, $s_{12} \in \{0, 0.2, 0.8\}$. For simplicity we choose $s_{11} = s_{22} = 1$.

In all three figures, we plot $\text{Var}(\hat{\mu}_2)/\text{Var}(\overset{*}{\mu}_2)$ versus $\text{Var}(\hat{\mu}_1)/\text{Var}(\overset{*}{\mu}_1)$. In Figure 1 we use a small covariance term, $s_{12} = 0.2$. In Figure 2 we use a large covariance, $s_{12} = 0.8$. In both figures, we shade by the size of r_1 relative to r_2. We already showed in (48) and (49) that when $s_{12} = 0$, the relative efficiency of $\overset{*}{\boldsymbol{\mu}}$ with respect to $\hat{\boldsymbol{\mu}}$ is 1, so there is no plot for $s_{12} = 0$.

Figure 1 shows the relationship between efficiency of $\overset{*}{\mu}_1$ and $\overset{*}{\mu}_2$ and the relative sizes of r_1 and r_2. In the worst case, when $r_1 >> r_2$, the efficiency of $\overset{*}{\mu}_1$ approaches 1, and the efficiency of $\overset{*}{\mu}_2$ falls off toward 0.97. A point to note here is that for a relatively small covariance, 0.2, the simple estimator, $\overset{*}{\mu}$ has a variance almost as small as that of $\hat{\mu}$. We will save discussion of the bands or striations for Figure 3.

Figure 2, which shows efficiency based on a covariance of 0.8, has the same pattern as Figure 1, but with larger variances for $\overset{*}{\mu}$, smaller efficiency. Here the worst cases can have an efficiency of less than 0.5 for either component of $\overset{*}{\mu}$, though when the efficiency of $\overset{*}{\mu}_1$ is that small, the efficiency of $\overset{*}{\mu}_2$ is very near 1.

Figure 3 shows the same points, for a covariance of 0.8, but shaded by the value of r_3. To

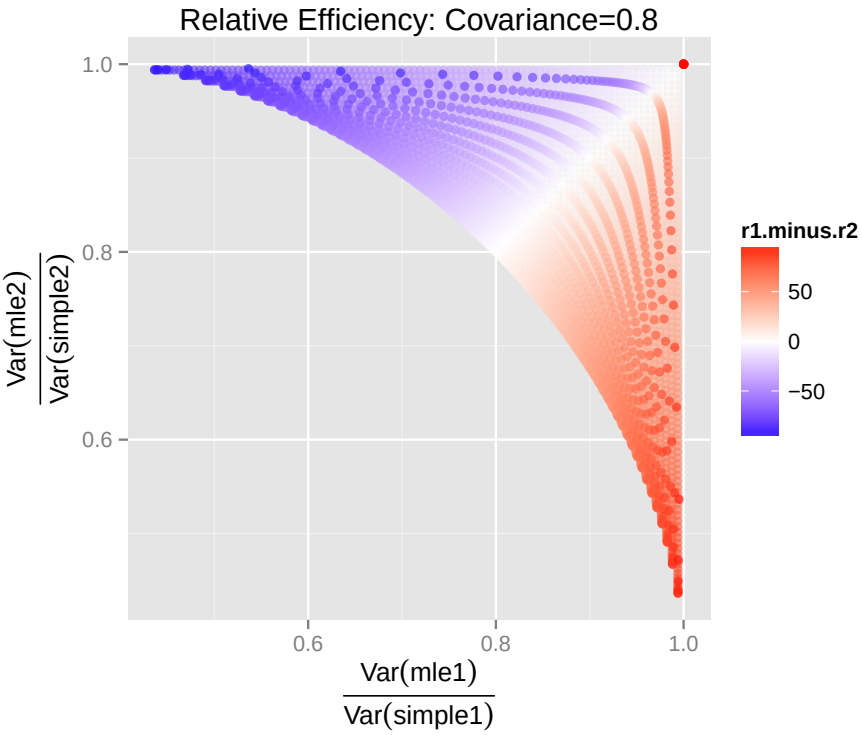

Figure 2: Efficiency of $\overset{*}{\mu}$ relative to $\hat{\mu}$ with high covariance (0.8)

help decipher it, we show a subset of the points in Figure 4.

Figure 4 shows a subset of the points, only the points where $r_3 \in \{1, 2, 3, 4, 61, 62, 63, 64\}$. When r_3 is very small there is a wide range of possibilities for r_1 and r_2. The four leftmost points in the upper left of Figure 4 are points where r_1 is 1 or 2; r_2 is somewhere between 94 and 97, and r_3 is 2, 3, or 4. In these cases, the sample for estimating μ_1 is very small, from 3 to 6 points, some from univariate data and some from the bivariate data. In that case, the MLE has a much smaller variance than the simple estimator. In that same case, there is a much larger sample from univariate data for estimating μ_2, upwards of 90 points, plus a handful of points from the bivariate data. In that case, the difference between the variance of $\overset{*}{\mu}_2$ and $\hat{\mu}_2$ is very small.

Graphs with negative covariances, -0.2, and -0.8 look the same as with positive covariances, and are omitted for the sake of brevity.

9.3. Summary of relative efficiency

Both the simple estimator for the location, $\overset{*}{\boldsymbol{\mu}}$, and the maximum likelihood estimator, $\hat{\boldsymbol{\mu}}$, are unbiased given $\boldsymbol{\Omega}$. The simple estimator's efficiency relative to the MLE tends to decrease as the covariance component of $\boldsymbol{\Omega}$ increases. We say "tends" because even with a covariance of 0.8, there are cases where the efficiency of both components of $\overset{*}{\boldsymbol{\mu}}$ relative to $\hat{\boldsymbol{\mu}}$ is very close to one.

When there are relatively few zeros, and they are balanced, $\overset{*}{\boldsymbol{\mu}}$ has a variance almost as small as $\hat{\boldsymbol{\mu}}$. The more zeros there are, or the more unbalanced their distribution is, the larger the variance of one or more components of the simple estimator.

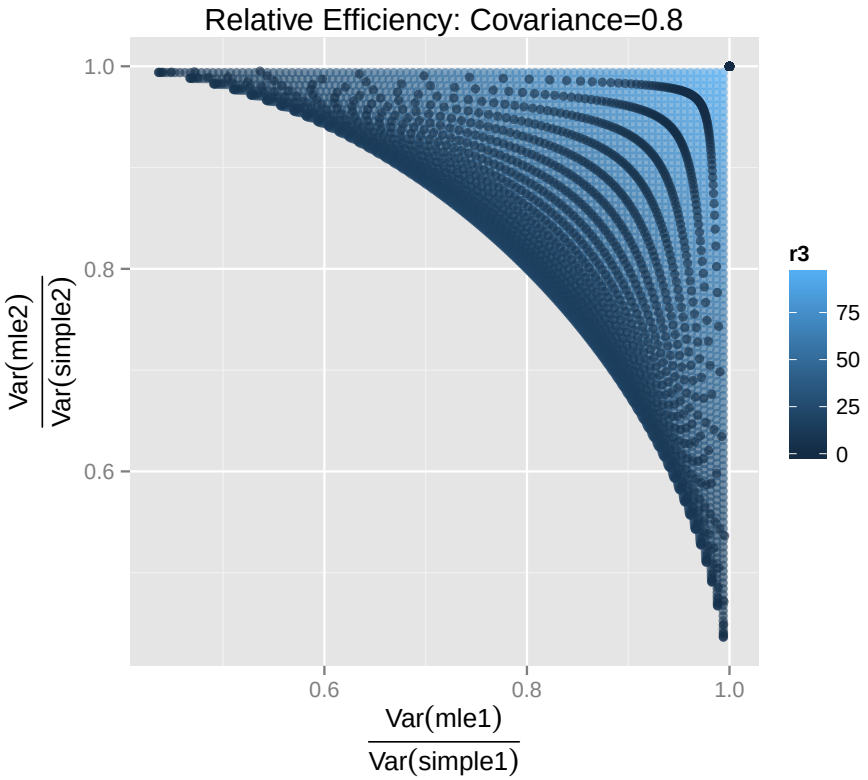

Figure 3: Efficiency of $\overset{*}{\mu}$ relative to $\hat{\mu}$ with high covariance (0.8) and relative to r_3

10. Subcompositional coherence

One of the reasons for using the logistic normal approach is that, in the base case without zeros, it preserves subcompositional coherence, described by Aitchison and Egozcue (2005) p. 831, as follows, "Subcompositional coherence demands that two scientists, one using full compositions and the other using subcompositions of these full compositions, should make the same inference about relations within the common parts." This implies that the subcomposition of a location estimator equals the location estimator for the subcomposition.

In the presence of zeros, do we maintain this property? It depends on which estimators are used. We have shown that in general when there are zeros, the MLE for the mean is not the same as the simple estimator for the mean. The MLE does not preserve subcompositional coherence when we have zeros. The simple estimators, by construction, do preserve subcompositional coherence provided the same D^{th} component is in both. Thus for inference, there is a choice to be made between maintaining subcompositional coherence and maximizing likelihood.

The issue of the relationship between compositions containing zeros, and subcompositional coherence, has been addressed from other points of view as well. Greenacre (2011) introduced a measure of *subcompositional incoherence* and suggested ways of getting it small enough for practical purposes in the paradigm of correspondence analysis. Scealy and Welsh (2014) argue more generally that although logratio methods for analyzing compositions have their uses, some of the principles that have been used to motivate them, such as subcompositional coherence, should not be taken to be as important as has been argued, e.g., by Aitchison (1994).

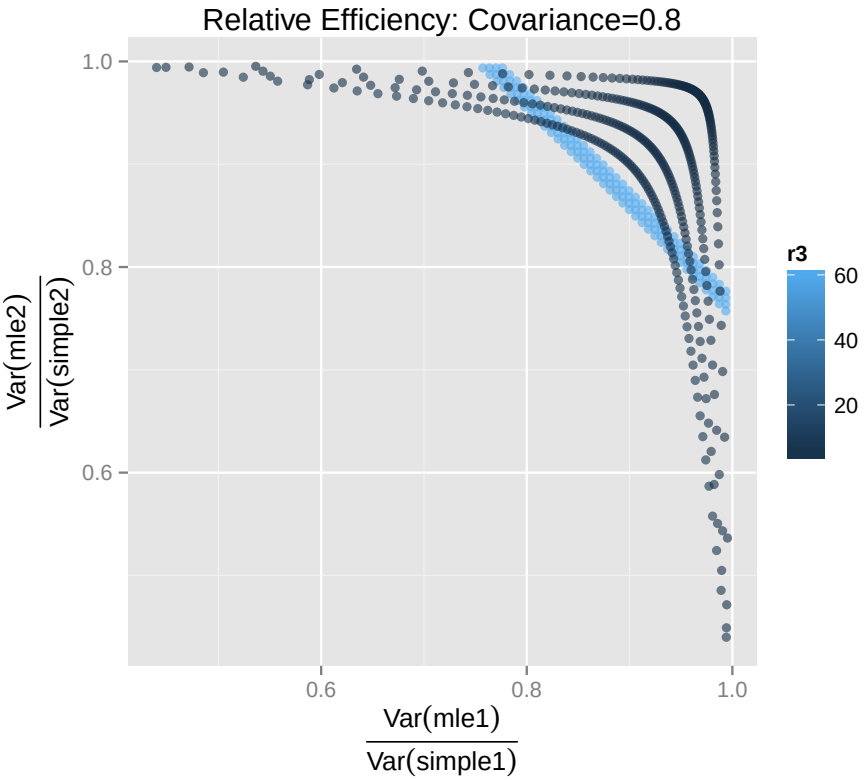

Figure 4: Efficiency of $\overset{*}{\hat{\mu}}$ relative to $\hat{\mu}$ with high covariance (0.8). $r_3 \in \{1:4, 61:64\}$.

11. Discussion

The goal has been to extend the additive logistic normal distribution to cope with essential zeros. We have done that by requiring that the final component of each composition be nonzero, and by projecting compositions with zeros onto smaller dimensional subspaces, thereby addressing the issues of division by zero, and the log of zero. We arrive at a mixture of logistic normals where each distribution has a mean and a covariance parameter which are projections from a common mean and covariance.

We construct two sets of estimators, simple estimators, $\overset{*}{\boldsymbol{\mu}}, \overset{*}{\boldsymbol{\Omega}}$, and maximum likelihood estimators, $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}}$. These are estimated using all of the compositions in the data, regardless of where the zeros occur, assuming that the D^{th} component is always nonzero. The simple estimators preserve subcompositional coherence, while the maximum likelihood estimators do not.

There are some limitations to this approach. In addition to the assumption that the D^{th} part is always nonzero, we assume that each composition has at least one more nonzero part, i.e., the vertices of the simplex are not in the support of the distribution. We assume a common mean and variance. Obviously, for a data set where different zero patterns have different means or variances or both, this model would not be appropriate. It is possible for the simple estimator of the covariance to produce a nonpositive definite matrix. If that happens, one possible approach is to estimate the covariance matrix using only the compositions that do not contain zeros. Another possible approach, once more work is done, would be to use the MLE. Currently, though, we do not have a general software solution for finding the MLE. One last concern is that a data set might have two parts which are never positive at the same time, in which case, the simple estimator for the covariance cannot be found.

In compositional data zeros are a common occurrence. We developed this logistic normal mixture model with the intention of making analysis of such data easier. For future work, we plan to extend existing compositional data methods for inference, graphing, clustering, etc., to work with this distribution.

12. Acknowledgements

We would like to thank two anonymous reviewers for their detailed and insightful comments. In addition, we would like to thank Shripad Sinari for helpful discussions. This research was supported by the University of Arizona/BIO5 Institute TRIF funds.

13. Appendix

13.1. Variance of location MLE, $\hat{\boldsymbol{\mu}}$

Next we derive the variance of the location MLE, $\hat{\boldsymbol{\mu}}|\boldsymbol{\Omega}, r_1, r_2, r_3$. First we rewrite the expression (34) so that each of the $\bar{y}$ terms stands alone.

$$\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3 = \frac{(r_3^2 + r_2 r_3)s_{11}s_{22}\bar{y}_{31} - r_2 r_3 s_{11}s_{12}\bar{y}_{32} + r_2 r_3 s_{11}s_{12}\bar{y}_{22} + ((r_1 r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)\bar{y}_{11}}{(r_3^2 + (r_2 + r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2} \tag{50}$$

To find the variance of $\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3$, we need to replace $r_1\bar{y}_{11}$ with $\sum_{j=1}^{r_1} y_{1j1}$; $r_2\bar{y}_{22}$ with $\sum_{j=1}^{r_2} y_{2j2}$; $r_3\bar{y}_{31}$ with $\sum_{j=1}^{r_3} y_{3j1}$; and $r_3\bar{y}_{32}$ with $\sum_{j=1}^{r_3} y_{3j2}$. We also make some other substitutions to simplify the algebra.

$$\text{Let } k_{31} = (r_3 + r_2)s_{11}s_{22}. \tag{51}$$
$$\text{Let } k_{32} = r_2 s_{11}s_{12}. \tag{52}$$
$$\text{Let } k_{22} = r_3 s_{11}s_{12}. \tag{53}$$
$$\text{Let } k_{11} = (r_3 + r_2)s_{11}s_{22} - r_2 s_{12}^2. \tag{54}$$
$$\text{Let } k_{denom} = (r_3^2 + (r_2 + r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2. \tag{55}$$

With these in place, we get

$$\begin{aligned}\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3 =& \frac{1}{k_{denom}}\left(k_{31}\sum_{j=1}^{r_3} y_{3j1} - k_{32}\sum_{j=1}^{r_3} y_{3j2} + k_{22}\sum_{j=1}^{r_2} y_{2j2} + k_{11}\sum_{j=1}^{r_1} y_{1j1}\right)\\ =& \frac{1}{k_{denom}}\left(\sum_{j=1}^{r_3}(k_{31}y_{3j1} - k_{32}y_{3j2}) + k_{22}\sum_{j=1}^{r_2} y_{2j2} + k_{11}\sum_{j=1}^{r_1} y_{1j1}\right).\end{aligned} \tag{56}$$

The y_{2j2} are i.i.d. univariate normal; the y_{1j1} are i.i.d. univariate normal; and the $\mathbf{y}_{3j}$ are i.i.d bivariate normal, so the variance of the estimator is:

$$\text{Var}(\hat{\mu}_1|\boldsymbol{\Omega}, r_1, r_2, r_3) = \left(\frac{1}{k_{denom}}\right)^2\left[\text{Var}\left(\sum_{j=1}^{r_3}(k_{31}y_{3j1} - k_{32}y_{3j2})\right) + \text{Var}\left(k_{22}\sum_{j=1}^{r_2} y_{2j2}\right) + \text{Var}\left(k_{11}\sum_{j=1}^{r_1} y_{1j1}\right)\right]. \tag{57}$$

$\text{Var}(y_{2j2}) = s_{22}$ and $\text{Var}(y_{1j1}) = s_{11}$, so

$$\text{Var}(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3) = \left(\frac{1}{k_{denom}}\right)^2 \left[\text{Var}\left(\sum_{j=1}^{r_3}(k_{31}y_{3j1} - k_{32}y_{3j2})\right) + k_{22}^2 r_2 s_{22} + k_{11}^2 r_1 s_{11}\right]. \tag{58}$$

To find the variance of the remaining sum requires the facts that $\mathbf{y}_{3j}$ are i.i.d., and that $\text{Cov}(y_{3j1}, y_{3j2}) = s_{12}$.

$$\begin{aligned}
\text{Var}\left(\sum_{j=1}^{r_3}(k_{31}y_{3j1} - k_{32}y_{3j2})\right) &= \sum_{j=1}^{r_3}\text{Var}(k_{31}y_{3j1} - k_{32}y_{3j2}) \\
&= \sum_{j=1}^{r_3}\left[\text{Var}(k_{31}y_{3j1}) + \text{Var}(k_{32}y_{3j2}) - 2k_{31}k_{32}\text{Cov}(y_{3j1}, y_{3j2})\right] \\
&= \sum_{j=1}^{r_3}\left[k_{31}^2 s_{11} + k_{32}^2 s_{22} - 2k_{31}k_{32}s_{12}\right] \\
&= r_3\left[k_{31}^2 s_{11} + k_{32}^2 s_{22} - 2k_{31}k_{32}s_{12}\right].
\end{aligned} \tag{59}$$

With that we can write the variance of the MLE, $\hat{\mu}_1$.

$$\begin{aligned}
&\text{Var}(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3) = \\
&\left(\frac{1}{k_{denom}}\right)^2 \left[r_3\left(k_{31}^2 s_{11} + k_{32}^2 s_{22} - 2k_{31}k_{32}s_{12}\right) + k_{22}^2 r_2 s_{22} + k_{11}^2 r_1 s_{11}\right].
\end{aligned} \tag{60}$$

Substituting the values for the k's back in gives:

$$\begin{aligned}
&\text{Var}(\hat{\mu}_1|\mathbf{\Omega}, r_1, r_2, r_3) = \\
&\frac{r_3\left((r_3+r_2)^2 s_{11}^3 s_{22}^2 + r_2^2 s_{11}^2 s_{12}^2 s_{22} - 2r_2(r_3+r_2)s_{11}^2 s_{12}^2 s_{22}\right) + r_2 r_3^2 s_{11}^2 s_{12}^2 s_{22} + r_1 s_{11}((r_3+r_2)s_{11}s_{22} - r_2 s_{12}^2)^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)^2}
\end{aligned} \tag{61}$$

Symmetry also gives the variance of $\hat{\mu}_2$ given $\mathbf{\Omega}$.

$$\begin{aligned}
&\text{Var}(\hat{\mu}_2|\mathbf{\Omega}, r_1, r_2, r_3) = \\
&\frac{r_3\left((r_3+r_1)^2 s_{22}^3 s_{11}^2 + r_1^2 s_{22}^2 s_{12}^2 s_{11} - 2r_1(r_3+r_1)s_{22}^2 s_{12}^2 s_{11}\right) + r_1 r_3^2 s_{22}^2 s_{12}^2 s_{11} + r_2 s_{22}((r_3+r_1)s_{22}s_{11} - r_1 s_{12}^2)^2}{((r_3^2 + (r_2+r_1)r_3 + r_1 r_2)s_{11}s_{22} - r_1 r_2 s_{12}^2)^2}
\end{aligned} \tag{62}$$

13.2. Variance of simple location estimator, $\overset{*}{\boldsymbol{\mu}}$

Our simple estimator for the location is $\overset{*}{\boldsymbol{\mu}} = \begin{bmatrix} \overset{*}{\mu}_1 \\ \overset{*}{\mu}_2 \end{bmatrix}$. Here we concern ourselves with $\text{Var}(\overset{*}{\mu}_1)$ and then rely on symmetry to arrive at the variance of $\overset{*}{\mu}_2$.

$$\overset{*}{\mu}_1 = \frac{1}{r_1 + r_3}\left[\sum_{j=1}^{r_1} y_{1j1} + \sum_{j=1}^{r_3} y_{3j1}\right] \tag{63}$$

$$\begin{aligned}
\text{Var}(\overset{*}{\mu}_1) &= \text{Var}\left(\frac{1}{r_1 + r_3}\left[\sum_{j=1}^{r_1} y_{1j1} + \sum_{j=1}^{r_3} y_{3j1}\right]\right) \\
&= \frac{1}{(r_1 + r_3)^2}\text{Var}\left(\sum_{j=1}^{r_1} y_{1j1} + \sum_{j=1}^{r_3} y_{3j1}\right)
\end{aligned}$$

$$= \frac{1}{(r_1 + r_3)^2} (r_1 s_{11} + r_3 s_{11})$$
$$= \frac{s_{11}}{r_1 + r_3}. \tag{64}$$

$$\text{By symmetry, } \mathrm{Var}(\overset{*}{\mu}_2) = \frac{s_{22}}{r_2 + r_3}. \tag{65}$$

References

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

Aitchison J (1994). *Principles of Compositional Data Analysis*, pp. 73–81. In Anderson, Olkin, and Fang (1994).

Aitchison J, Egozcue JJ (2005). "Compositional Data Analysis: Where Are We and Where Should We Be Heading?" *Mathematical Geology*, **37**(7), 829–850.

Aitchison J, Kay JW (2003). "Possible Solution of Some Essential Zero Problems in Compositional Data Analysis." In *Proceedings of CoDaWork 2003, The First Compositional Data Analysis Workshop*. Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada. http://ima.udg.edu/Activitats/CoDaWork03/paper_Aitchison_and_Kay.pdf.

Anderson TW, Olkin I, Fang K (eds.) (1994). *Multivariate Analysis and Its Applications*. Institute of Mathematical Statistics, Hayward, CA.

Bacon-Shone J (2008). "Discrete and Continuous Compositions." In *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, May 27–30, University of Girona, Girona (Spain), CD-ROM*. Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada.

Billheimer D, Guttorp P, Fagan WF (2001). "Statistical Interpretation of Species Composition." *Journal of the American Statistical Association*, **96**(456), 1205–1214. http://dx.doi.org/10.1198/016214501753381850.

Butler A, Glasbey C (2008). "A Latent Gaussian Model for Compositional Data with Zeros." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**(5), 505–520.

Daunis-i-Estadella J, Martín-Fernández JA (eds.) (2008). *Proceedings of CODAWORK 2008, The 3rd Compositional Data Analysis Workshop*. University of Girona. CD-ROM (ISBN: 978-84-8458-272-4).

Fry JM, Fry TR, Mclaren KR (2000). "Compositional Data Analysis and Zeros in Micro Data." *Applied Economics*, **32**(8), 953–959. http://www.tandfonline.com/doi/pdf/10.1080/000368400322002.

Greenacre M (2011). "Measuring Subcompositional Incoherence." *Mathematical Geosciences*, **43**, 681–693. doi:10.1007/s11004-011-9338-5. URL http://link.springer.com/content/pdf/10.1007/s11004-011-9338-5.pdf.

Kent JT (1982). "The Fisher–Bingham Distribution on the Sphere." *Journal of the Royal Statistical Society, B*, **44**, 71–80.

Leininger TJ, Gelfand AE, Allen JM, Silander Jr JA (2013). "Spatial Regression Modeling for Compositional Data with Many Zeros." *Journal of Agricultural, Biological, and Environmental Statistics*, **18**(3), 314–334.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2014). "Bayesian-Multiplicative Treatment of Count Zeros in Compositional Data Sets." *Statistical Modelling*. doi:10.1177/1471082X14535524.

Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011). *Dealing With Zeros*, pp. 43–58. In Pawlowsky-Glahn and Buccianti (2011).

Palarea-Albaladejo J, Martín-Fernández J (2008). "A Modified EM Alr-Algorithm for Replacing Rounded Zeros in Compositional Data Sets." *Computers & Geosciences*, **34**(8), 902–917.

Palarea-Albaladejo J, Martín-Fernández JA (2015). "zCompositions - R Package for Multivariate Imputation of Left-Censored Data under a Compositional Approach." *Chemometrics and Intelligent Laboratory Systems*, **143**, 85–96. doi:10.1016/j.chemolab.2015.02.019.

Palarea-Albaladejo J, Martín-Fernández JA, Olea RA (2014). "A Bootstrap Estimation Scheme for Chemical Compositional Data with Nondetects." *Journal of Chemometrics*, **28**(7), 585–599.

Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. Wiley.

Scealy J, Welsh A (2011). "Regression for Compositional Data by Using Distributions Defined on the Hypersphere." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3), 351–375.

Scealy J, Welsh A (2014). "Colours and Cocktails: Compositional Data Analysis 2013 Lancaster Lecture." *Australian & New Zealand Journal of Statistics*, **56**(2), 145–169.

Stewart C, Field C (2011). "Managing the Essential Zeros in Quantitative Fatty Acid Signature Analysis." *Journal of Agricultural, Biological, and Environmental Statistics*, **16**(1), 45–69.

Affiliation:

John Bear
Statistical Consulting Lab
University of Arizona
888 N Euclid Ave
85719-4824 Tucson, AZ, USA
E-mail: jbear@email.arizona.edu

Dean Billheimer
Statistical Consulting Lab
University of Arizona
888 N Euclid Ave
85719-4824 Tucson, AZ, USA
E-mail: dean.billheimer@arizona.edu

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 45 *Submitted:* 2015-10-16
September 2016 *Accepted:* 2016-05-03

Austrian Journal of Statistics
September 2016, Volume 45, 25–44.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.126

Changing the Reference Measure in the Simplex and Its Weighting Effects

Juan José Egozcue
Universitat Politècnica de Catalunya, Spain

Vera Pawlowsky-Glahn
Universitat de Girona, Spain

Abstract

Under the assumption that the Aitchison geometry holds in the simplex, standard analysis of compositional data assumes a uniform distribution as reference measure of the space. Changing the reference measure induces a weighting of parts. The changes that appear in the algebraic-geometric structure of the simplex are analysed, as a step towards understanding the implications for elementary statistics of random compositions. Some of the standard tools in exploratory analysis of compositional data, such as center, variation matrix and biplots are studied in some detail, although further research is still needed. The main result is that through a progressive down-weighting of some parts, the geometry of the space approaches that of the corresponding subcomposition. In this way, the coherence between standard and down-weighted analyses is preserved.

Keywords: simplex, sigma-additive measures, subcomposition, weighting, Bayes space, biplot, center, variability.

1. Introduction

When analysing a composition, some parts may heavily influence the results. A typical example are inaccuracies in the measurements in some not fully relevant parts. They can dominate the analysis, producing a large contribution to variability or to distances. Also, relevance of some parts in a given problem can call for weighting techniques to adapt the simplex geometry accordingly. There are a number of weighting techniques that can be useful in this sense (e.g. Filzmoser and Hron 2015). Among them, the change of reference measure of the simplex has several implications that need to be fully understood for a consistent analysis. This contribution is aimed at showing changes that appear in the algebraic-geometric structure of the simplex, as well as some effects in elementary statistics and exploratory tools.

One of the most fruitful concepts in compositional analysis is that of subcomposition (Aitchison 1986). In Aitchison (1992), some reasonable principles for a coherent analysis of subcompositions were established. Beyond the idea that compositional analyses should be scale invariant, those principles included the assumption that distances between compositions should be greater than or equal to those observed in a subcomposition. This principle, called subcompositional dominance (Aitchison 1992; Aitchison, Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2000; Egozcue 2009), highlights a change of the geometry of subcomposi-

tions (for instance, a change in inter-distances between two data-points in the subcomposition) with respect to the original geometry of the full composition. Taking a subcomposition can be considered as an extreme case of down-weighting, since the influence of some parts of the composition is removed from the analysis. However, there are cases in which the complete removal of the influence of some parts of the original composition is not desirable. This motivates the idea of weighting compositions as a continuous transition from the full composition, endowed with the corresponding Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001), to a subcomposition, endowed with the induced Aitchison geometry, which differs in dimension and metrics (distances, inner product, norm).

Apparently, there are many ways of weighting compositions so that a transition from a full composition to a subcomposition is performed. However, fulfilling all coherence requirements is quite challenging. One option deserving attention is the one proposed for Bayes spaces (Boogaart, Egozcue, and Pawlowsky-Glahn 2010; Egozcue, Pawlowsky-Glahn, Tolosana-Delgado, Ortego, and Boogaart 2013b) and, more specifically, for Bayes Hilbert spaces (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2006; Boogaart, Egozcue, and Pawlowsky-Glahn 2014). Bayes Hilbert spaces are spaces of measures and densities, and their algebraic-geometric structure is an extension of the Aitchison geometry of the simplex. In fact, in (Boogaart *et al.* 2014), it is shown that the simplex, endowed with the Aitchison geometry, is a particular case of a Bayes Hilbert space. In the development of Bayes Hilbert spaces, a reference probability measure is introduced as a parameter regulating the geometry of the measures and densities in the space. This kind of approach provides a way of coherently introducing weighting strategies, both in the simplex and in the analysis of compositional data. The present aim is to start studying the change of reference measure in the simplex, being conscious that there is a long way from the general theory of Bayes Hilbert spaces to applications in compositional data analysis. Special attention is paid to the transition from the geometry of the simplex $\mathcal{S}^D$ for compositions to the geometry of $\mathcal{S}^d$, $d < D$, where subcompositions are defined. The main difficulties are interpretative, as usual in compositional data analysis.

The structure of the paper is as follows: Section 2 translates the milestones of Bayes Hilbert spaces into the case of compositions, with special emphasis on the role of the reference measure. Section 3 introduces the centered log-ratio transformation (clr) with respect to an arbitrary reference measure in the simplex, following the definition in Boogaart *et al.* (2014) for general Bayes Hilbert spaces. Section 4 gets into details of metric concepts under a change of the reference measure, such as orthogonality, bases, and balances. A proposition on dominance of distances is there stated (see proof in Appendix A). Section 5 gives an introduction to distributions of random compositions, their variability and centre under a weighted geometry of the simplex. Section 6 shows how variation matrix and biplots work under weighting using an example of electoral results.

2. Change of reference measure for compositions

Consider D categories $c_1, c_2, \ldots, c_D$; they represent a partition of a measurable space Ω. A D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ in the D-part simplex $\mathcal{S}^D$ assigns a proportion x_i to the category c_i. Assuming that the composition $\mathbf{x}$ is closed to 1, the proportion assigned to the whole space Ω is just 1. For any subset of categories, the proportion assigned is the sum of the corresponding proportions. For instance, the proportion assigned to the subset $\{c_1\}$ is x_1, and the proportion assigned to the subset $\{c_1, c_2\}$ is $x_1 + x_2$. From this point of view, the composition $\mathbf{x}$ defines a finite additive measure on Ω, which is denoted $\mu_{\mathbf{x}}\{\cdot\}$. The argument of this measure is any subset of Ω. Examples are $\mu_{\mathbf{x}}\{\Omega\} = 1$, $\mu_{\mathbf{x}}\{\emptyset\} = 0$, $\mu_{\mathbf{x}}\{c_1\} = x_1$, $\mu_{\mathbf{x}}\{c_1, c_2\} = x_1 + x_2$.

Measures can be represented by densities. The idea is that sums (integrals) on a subset of Ω give the measure of this subset. In the case of the simplex $\mathcal{S}^D$, the density is identified with

the composition $\mathbf{x}$, as for any subset $A \subseteq \Omega$ it satisfies

$$\mu_{\mathbf{x}}\{A\} = \sum_{c_i \in A} x_i P_0\{c_i\} \quad , \quad P_0\{c_i\} = 1 \ , \ i = 1, 2, \ldots, D \ ,$$

where the uniform measure $P_0\{\cdot\}$ on Ω has been made explicit as reference measure. Note that $P_0\{\Omega\} = D$ and addends of sums (integrals) along the composition are equally weighted with $1 = P_0\{c_i\}$. The reference measure specified as $\mathbf{p}_0 = (P_0\{c_1\}, P_0\{c_2\}, \ldots, P_0\{c_D\})$ is a non-closed uniform measure. Therefore, it is compositionally equivalent to the neutral element of the simplex $\mathbf{n} = (1/D, 1/D, \ldots, 1/D)$. The conclusion is that a composition $\mathbf{x} \in \mathcal{S}^D$ defines a measure $\mu_{\mathbf{x}}$ on Ω specifying the measure of each elementary subset $\{c_i\}$ and, at the same time, $\mathbf{x}$ is the density of $\mu_{\mathbf{x}}$ with respect to the uniform reference measure P_0, which density is $\mathbf{p}_0$. In mathematical terms, the density (composition) $\mathbf{x}$ is the Radon-Nikodym derivative of $\mu_{\mathbf{x}}$ with respect to the reference measure P_0 which can be written as

$$\mathbf{x} = \frac{d\mu_{\mathbf{x}}}{dP_0} \quad , \quad \mu_{\mathbf{x}}\{A\} = \int_A \frac{d\mu_{\mathbf{x}}}{dP_0} \, dP_0 = \sum_{c_i \in A} x_i \ P_0\{c_i\} \ ,$$

for any $A \subseteq \Omega$. When P_0 is the unitary and uniform reference measure, there is no need to distinguish between $\mathbf{x}$ as a composition, as a measure or as a density. These facts change when weights are introduced through the reference measure.

To analyse the effects of a change of reference measure as a means to introduce weights, consider an arbitrary array of positive weights, $\mathbf{p} = (p_1, p_2, \ldots, p_D)$. The corresponding measure P is then characterised by $P\{c_i\} = p_i$, for $i = 1, 2, \ldots, D$, and by the measure of the whole space, $P\{\Omega\} = \sum_{i=1}^{D} p_i$. Note that $\mathbf{p}$ is the density of P with respect to the uniform measure P_0. A question is now to look for the density of the measure $\mu_{\mathbf{x}}$ with respect to the new reference measure P. This density is $\mathbf{y} = \mathbf{x}/\mathbf{p} = (x_1/p_1, x_2/p_2, \ldots, x_D/p_D)$. In fact, for $A \subseteq \Omega$,

$$\mu_{\mathbf{x}}\{A\} = \sum_{c_i \in A} x_i = \sum_{c_i \in A} y_i \ p_i = \sum_{c_i \in A} \frac{x_i}{p_i} \ p_i \, . \tag{1}$$

The measure $\mu_{\mathbf{x}}$ is thus retrieved from two different densities, $\mathbf{x}$ when considering the uniform reference P_0, and $\mathbf{y}$ for a reference P. Note that $\mathbf{y}$ is a vector which components do not add to one, i.e. it is not closed. However, it is compositionally equivalent to $\mathcal{C}\mathbf{y} = \mathbf{x} \ominus \mathbf{p}$, as its components are proportional (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

If the reference measure P is represented by the vector of weights $\mathbf{p}$, the composition $\mathcal{C}\mathbf{y}$ is just a perturbation of $\mathbf{x}$, a shift in the simplex, recalling that the perturbation-difference $\ominus$ includes the closure, $\mathcal{C}$, and, consequently, $\mathcal{C}\mathbf{y} = \mathbf{x} \ominus \mathbf{p} = \mathbf{x} \ominus \mathcal{C}\mathbf{p}$. From now on, the non-closed version of $\mathbf{y}$ is denoted $\mathbf{y}^{(\mathbf{p})}$ when the reference measure needs to be specified. Following Boogaart *et al.* (2010) and Boogaart *et al.* (2014), a weighted perturbation and powering can be defined for densities like $\mathbf{y}^{(\mathbf{p})}$ such that they operate linearly in the weighted simplex. However, their use is not recommended in this context as standard perturbation ($\oplus$) and powering ($\odot$) are easily interpreted and computed in the applications. This avoids linear operations with the shifted densities $\mathcal{C}\mathbf{y}^{(\mathbf{p})} = \mathbf{x} \ominus \mathbf{p}$. In practice, weighted compositions will be used only in the computation of distances and inner products, as explained below.

3. Centred log-ratio with respect to a reference measure

In Boogaart *et al.* (2014), the clr-transformation of a density f with respect to a given reference measure P, is defined as

$$\mathrm{clr}_P(f)(x) = \log f(x) - \frac{1}{P\{\Omega\}} \int_\Omega \log f(\xi) \ dP\{\xi\} \ , \quad x \in \Omega \ , \tag{2}$$

where Ω is the measurable set where the density f is defined. In the present case, Ω is the set of the D parts or categories of $\mathcal{S}^D$, namely c_i, $i = 1, 2, \ldots, D$. Therefore, the values of x

in such an expression correspond to the c_i's. Since f is a density of a measure with respect to the reference measure P, it can be identified with the density $\mathbf{y} = \mathbf{x}/\mathbf{p}$, as introduced in Section 2. With these identifications, the $\mathrm{clr}_\mathbf{p}$-transformation of the simplex with respect to the measure P, represented by $\mathbf{p} = (p_1, p_2, \ldots, p_D)$, is

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = \left(\log \frac{y_1}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \ , \ \log \frac{y_2}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \ , \ldots , \log \frac{y_D}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \right) , \quad \mathrm{g}_\mathbf{p}(\mathbf{y}) = \exp\left(\frac{1}{s_\mathbf{p}} \sum_{i=1}^{D} p_i \log y_i \right) \ , \quad (3)$$

where $s_\mathbf{p} = \sum_{i=1}^{D} p_i$, and $\mathrm{g}_\mathbf{p}(\cdot)$ denotes a weighted geometric mean of the parts y_i. It is remarkable that $\mathbf{p}$, the reference measure of the categories c_i, is not closed to D, and that $P\{\Omega\} = s_\mathbf{p}$, while for P_0 the uniform reference measure $s_{\mathbf{p}_0} = D$. Note also that $\mathbf{y}$ can be closed or not, as Equation 3 is scale invariant.

An important characteristic of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ is that the weighted sum of its D components is zero, that is

$$\sum_{i=1}^{D} p_i \log \frac{y_i}{\mathrm{g}_\mathbf{p}(\mathbf{y})} = 0 \ , \quad (4)$$

generalising the ordinary clr in $\mathcal{S}^D$, for which the sum of its components (weights equal to 1) is zero. This has a geometric interpretation in the space $\mathbb{R}^D$, where a point has coordinates $\log(\mathbf{y}) = (\log y_1, \log y_2, \ldots, \log y_D)$. As illustrated in Figure 1, which shows a scheme for $D = 2$, to obtain the ordinary clr of a generic point $\log(\mathbf{y})$, the point is orthogonally projected onto a hyperplane through the origin whose orthogonal vector is $(1, 1, \ldots, 1)$ (Aitchison 1986; Pawlowsky-Glahn *et al.* 2015). When using a non-uniform $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ the procedure to get $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ is to orthogonally project the point $\log(\mathbf{y})$ onto a hyperplane whose orthogonal vector is $\mathbf{p}$, as shown by the inner product in $\mathbb{R}^D$ implicit in Equation 4. Summarising, $\mathrm{clr}_\mathbf{p}$ is a projection of $\log(\mathbf{y})$ on a hyperplane whose normal vector is $\mathbf{p}$.

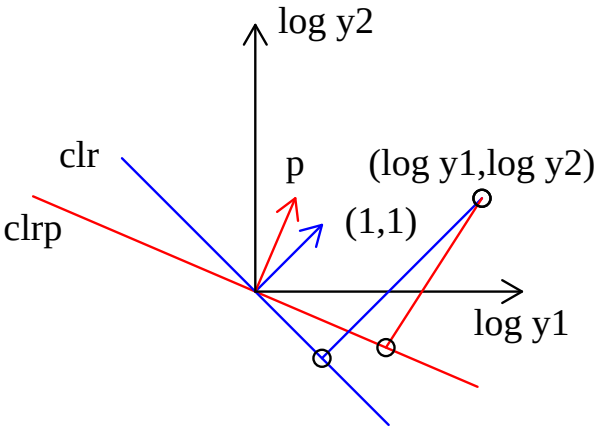

Figure 1: Generic 2-part composition (y_1, y_2), log-transformed into $(\log y_1, \log y_2)$. Two reference measures with densities $(1, 1)$ (uniform, blue arrow) and $\mathbf{p}$ (red arrow) are considered. The point $(\log y_1, \log y_2)$ is projected, parallel to the reference arrow, on the clr-plane (blue) and on the $\mathrm{clr}_\mathbf{p}$-plane (red), thus obtaining the respective transformations.

A particular case of interest is that of

$$p_i = 1 \ , \ i = 1, 2, \ldots, D-1 \ , \ p_D = \epsilon \ , \quad (5)$$

for which $P\{\Omega\} = (D-1) + \epsilon$. When $\epsilon \to 0$, the D-th part is down-weighted from 1 to $\epsilon \ll 1$. For small enough ϵ, the weighted geometric mean $\mathrm{g}_\mathbf{p}$ in Equation 3 approaches the ordinary geometric mean of the first $D-1$ parts of $\mathbf{y}$. A consequence is that the first $D-1$ components of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ approach the ordinary clr of the subcomposition formed by $(y_1, y_2, \ldots, y_{D-1})$. This

suggests that this kind of reference measures may approach the induced Aitchison geometry on the subcomposition.

4. Metrics under change of reference

The clr transformation can be used to define the inner product in $\mathcal{S}^D$, as was done in Bayes Hilbert spaces (Boogaart *et al.* 2014, Def. 2). There, the proposed definition was

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{B^2} = \frac{1}{P\{\Omega\}} \langle \mathrm{clr}_P(\mathbf{y}_2), \mathrm{clr}_P(\mathbf{y}_1) \rangle \ ,$$

where $\langle \cdot, \cdot \rangle$ is the ordinary inner product in $\mathbb{R}^D$. This definition leads to an inner product in $\mathcal{S}^D$ which, for a uniform reference measure P_0, with weights $\mathbf{p}_0 = (1, 1, \ldots, 1)$, is

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_a = \frac{1}{D} \langle \mathrm{clr}(\mathbf{y}_2), \mathrm{clr}(\mathbf{y}_1) \rangle \ , \tag{6}$$

which is not the standard in compositional data analysis due to the factor $1/D$. This inner product is not suitable for compositional data analysis, as it does not fulfill the principle of subcompositional dominance of distances. For instance, consider the 3-part compositions $\mathbf{u} = (0.1, 0.7, 0.2)$ and $\mathbf{v} = (1/3, 1/3, 1/3)$. Their distance, in the geometry induced by the inner product (6) in $\mathcal{S}^3$, is $\mathrm{d}_3(\mathbf{u}, \mathbf{v}) = 0.805$. Taking the subcomposition formed by the first and second part and computing the distance in $\mathcal{S}^2$ according to (6), the result is $\mathrm{d}_2(\mathbf{u}, \mathbf{v}) = 0.973$. Since $\mathrm{d}_3(\mathbf{u}, \mathbf{v}) < \mathrm{d}_2(\mathbf{u}, \mathbf{v})$, the principle of subcompositional dominance is violated.

The discussion about the role of the constant $1/D$ in the inner product is related to the fact that in Boogaart *et al.* (2014) the reference measure was assumed to satisfy $P_0\{\Omega\} = 1$. If $0 < P_0\{\Omega\} < +\infty$, the value $P_0\{\Omega\}$ is irrelevant when one does not try to compare results of an analysis using different reference measures, as was the case in that contribution. On the contrary, in Egozcue *et al.* (2006) the reference is implicitly assumed to be proportional to the length of the interval supporting the densities of the Hilbert space, that is $P_0\{\Omega\}$ is adapted for each support Ω. Here this second strategy has been adopted so that analytical results using different references become comparable, fulfilling the subcompositional coherence requirements. This strategy of normalizing the reference measures has a consequence which might be uncomfortable for some readers, namely that $\mathbf{p}_0$, or in general $\mathbf{p}$, are not only non-closed compositions, but convey also information about the size of Ω, $P\{\Omega\} = \sum_{i=1}^{D} P\{c_i\}$. In the following development, $\mathbf{p}$ or $\mathbf{p}_0$ appear to be closed when represented as elements of the simplex, but retain their absolute values when the components are used as weights in sums (integrals) along compositions or clr images.

To match the present definition to the standard practice in compositional data analysis (Aitchison 1986; Aitchison and Egozcue 2005; Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras 2011; Pawlowsky-Glahn *et al.* 2015) and to the subcompositional dominance of distances, the factor $1/D$ in (6) is suppressed. Remember that multiplication by a real scalar in an inner product does not change its character.

In the case of using a reference measure represented by the weights $\mathbf{p}$, the appropriate definition of the weighted Aitchison inner product is

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{\mathbf{p}} = \sum_{i=1}^{D} p_i \log \frac{y_{1i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_1)} \log \frac{y_{2i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_2)} \ , \tag{7}$$

where $\mathbf{y}_k = \mathbf{y}_k^{(\mathbf{p})}$, $k = 1, 2$ are in $\mathcal{S}^D$. The expression in the right hand side of Equation (7) is an inner product of the $\mathrm{clr}_{\mathbf{p}}$ as real vectors with respect to the measure P.

The weighted Aitchison norm, derived from the inner product, is $\|\mathbf{y}\|^2_{\mathbf{p}} = \langle \mathbf{y}, \mathbf{y} \rangle_{\mathbf{p}}$, and an explicit expression of the distance is

$$\mathrm{d}^2_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2) = \langle \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2), \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2) \rangle_{\mathbf{p}} = \sum_{i=1}^{D} p_i \left(\log \frac{y_{1i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_1)} - \log \frac{y_{2i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_2)} \right)^2 .$$

This expression of weighted distance can be written in matrix notation

$$\mathrm{d}^2_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2) = (\mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2))\ \mathrm{diag}(\mathbf{p})\ (\mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2))^\top ,$$

where the $\mathrm{clr}_{\mathbf{p}}$ are row vectors and $\mathrm{diag}(\mathbf{p})$ is a diagonal (D, D)-matrix containing the weights $\mathbf{p}$. These definitions coincide with those of the ordinary Aitchison geometry of $\mathcal{S}^D$ whenever $\mathbf{p} = \mathbf{p}_0 = (1, 1, \ldots, 1)$. When $\mathbf{p} \neq \mathbf{p}_0$, the inner product differs from the ordinary Aitchison inner product and, consequently, also norm and distance are different.

To get a further intuition of what is changing with $\mathbf{p}$, it is instructive to build orthonormal basis of the simplex according to the change of reference. It allows to show how these bases appear under a change of $\mathbf{p}$ in particular cases.

A straightforward technique for obtaining orthonormal basis of the simplex and their respective coordinates (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003) is that of sequential binary partitions (SBP) (Egozcue and Pawlowsky-Glahn 2005, 2006). Like in the standard case (reference measure P_0), when using a reference measure with the weights $\mathbf{p}$, the procedure is based on a partition coded as in Table 1, but the formulae to obtain the contrast matrix are modified. Table 1 shows a generic sign code for an SBP, adding weights $\mathbf{p}$ as column labels (second row) for further comment on the generalised technique.

Table 1: A generic table of an SBP for a five-part composition. Weights from the reference measure are placed in the second row, under the part label. Rows are labelled as balances b_i for further reference.

parts	y_1	y_2	y_3	y_4	y_5
weights	p_1	p_2	p_3	p_4	p_5
b_1	+1	−1	−1	−1	+1
b_2	+1	0	0	0	−1
b_3	0	+1	−1	−1	0
b_4	0	0	+1	−1	0

Denote the entries of the matrix code as θ_{ij}, $i = 1, 2, \ldots, D-1$, $j = 1, 2, \ldots, D$. For the case in Table 1, $D = 5$ and, for instance, $\theta_{32} = +1$. When using the standard reference measure $\mathbf{p}_0 = (1, 1, \ldots, 1)$, the clr coefficients of an element of the basis, that is of a balancing element, are given by

$$\psi_{ij} = \begin{cases} +\frac{1}{n_i^+} \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = +1 \\ -\frac{1}{n_i^-} \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = -1 \\ 0 & \text{if } \theta_{ij} = 0 \ , \end{cases} \tag{8}$$

where n_i^+ denotes the number of $+1$, respectively n_i^- of -1, in the i-th row of the code table. When using the reference measure which weights p_j are not unity, these formulas for the $\mathrm{clr}_{\mathbf{p}}$ of balancing elements are the same except that n_i^+, n_i^- are

$$n_i^+ = \sum_{\theta_{ij}=+1} p_j \quad , \quad n_i^- = \sum_{\theta_{ij}=-1} p_j \ .$$

The contrast matrix Ψ, with entries ψ_{ij}, $i = 1, 2, \ldots, D$, $j = 1, 2, \ldots, D-1$, fulfills the conditions

$$\Psi\ \mathrm{diag}(\mathbf{p})\ \Psi^\top = I_{D-1} \quad , \quad \mathrm{diag}(\mathbf{p})\ \Psi^\top \Psi = I_D - \frac{1}{D} \mathbf{p}^\top \mathbf{1} \ , \tag{9}$$

where I_m is the (m,m)-identity matrix; $\mathbf{p}$ and $\mathbf{1} = (1,1,\ldots,1)$ are taken as row D-vectors, and $\mathrm{diag}(\mathbf{p})$ is a (D,D) diagonal matrix with entries equal to the components of $\mathbf{p}$. The first condition is equivalent to saying that balancing elements are unitary compositions mutually orthogonal. In fact, their $\mathrm{clr}_{\mathbf{p}}$ are unitary and orthogonal in the weighted Euclidean geometry.

Coordinates of a density $\mathbf{y} \in \mathcal{S}^D$ with respect to an orthonormal basis are found carrying out the inner product of a balancing element in the basis with the density $\mathbf{y} = \mathbf{x}/\mathbf{p}$. In general, these coordinates are termed weighted isometric log-ratio coordinates and denoted by $\mathrm{ilr}_{\mathbf{p}}$. In the particular case in which they are obtained using an SBP, they are called weighted balances. For simplicity, these weighted balances are denoted b_i, $i = 1,2,\ldots,D-1$, with no reference to the weights associated with the change of measure (as shown in Table 1). The $\mathrm{ilr}_{\mathbf{p}}$ coordinates can be obtained using the matrix expression

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathbf{b} = \mathrm{clr}_{\mathbf{p}}(\mathbf{y})\ \mathrm{diag}(\mathbf{p})\ \Psi^{\top}\ , \tag{10}$$

where compositions and their $\mathrm{clr}_{\mathbf{p}}$ and $\mathrm{ilr}_{\mathbf{p}}$ transforms are considered row-vectors. Note that each component of $\mathbf{b} = (b_1, b_2, \ldots, b_{D-1})$ is a weighted inner product of $\mathrm{clr}_{\mathbf{p}}(\mathbf{y})$ with the corresponding $\mathrm{clr}_{\mathbf{p}}$ of a balancing element. The inverse $\mathrm{ilr}_{\mathbf{p}}$ transformation is readily obtained using the properties (9) of Ψ

$$\mathcal{C}\mathbf{y} = \mathcal{C}\exp(\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})\Psi) \quad , \quad \mathrm{clr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{ilr}_{\mathbf{p}}(\mathbf{y})\Psi\ ,$$

being the first of these relations formally identical to the standard inverse ilr with reference measure P_0. The relationship of $\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})$ and $\mathrm{ilr}(\mathbf{x})$ is developed in Appendix B.

Although Equation 10 is useful from a computational point of view, an explicit expression of balances gives a deeper insight into the meaning of weighted balances. Consider a sign code of a step in an SBP, for which n_i^+, n_i^- are given. The corresponding weighted balance is

$$b_i = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}}\ \log\left(\frac{\prod_{(\theta_{ij}=+1)}\ y_j^{p_j/n_i^+}}{\prod_{(\theta_{ij}=-1)}\ y_j^{p_j/n_i^-}}\right)\ , \tag{11}$$

where the products span over the parts corresponding to the sign code θ_{ij}. When the weights $p_j = 1$, the balance reduces to the standard balances, as n_i^+, n_i^- are then the number of $+1$ and -1 in the i-th row of the sign code, respectively. The main feature, when the reference is not $\mathbf{p}_0$, is that the ratios within the logarithm are ratios of a kind of weighted geometric means. Note that, in general, n_i^+, n_i^- are not integers and each part is powered to the weight corresponding to that part. When some p_j is small, relative to other weights, it plays a minor role in these weighted geometric means. Furthermore, the weighted balances are scale invariant log-contrasts, that is, if the composition $\mathbf{y}$ is multiplied by a positive constant, the weighted balance remains unaltered.

Expressing inner products, norms, and distances as functions of weighted coordinates $\mathrm{ilr}_{\mathbf{p}}$ can be useful, because they are exactly those of the standard Euclidean geometry. For the inner product and square-distance they are

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{\mathbf{p}} = \langle \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_1), \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_2) \rangle \quad , \quad \mathrm{d}^2_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2) = \mathrm{d}^2(\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_1), \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_2))\ , \tag{12}$$

where $\langle \cdot, \cdot \rangle$, $\mathrm{d}(\cdot,\cdot)$, are the ordinary Euclidean inner product and distance.

Whenever there is a change in the geometry of compositions, the subcompositional dominance of distances is a critical point. In the standard approach, the distance between any two compositions $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{S}^D$ is $\mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2)$. After taking a given subcomposition in $\mathcal{S}^d$, $d < D$, the distance between the respective subcompositions, $\mathbf{x}_1^{(d)}, \mathbf{x}_2^{(d)}$, satisfies $\mathrm{d}_a(\mathbf{x}_1^{(d)}, \mathbf{x}_2^{(d)}) \leq \mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2)$. In this case, both spaces have integer reference measures with $P\{\Omega_D\} = D$ and $P\{\Omega_d\} = d$ and, for $D = 3$, $d = 2$ the corresponding weights are $(1,1,1)$ and $(1,1,0)$, respectively. When changing the reference measure by down weighting some of the weights, a dominance of distances is expected, as it occurs when taking subcompositions. The dominance of distances

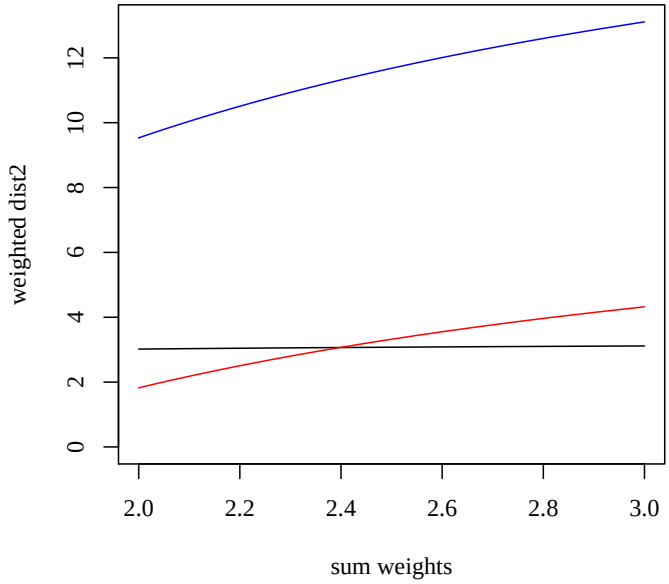

Figure 2: Evolution of weighted square-distances between three measures represented by the compositions $\mathbf{x}_1 = (0.1, 0.7, 0.2), \mathbf{x}_2 = (0.5, 0.3, 0.2), \mathbf{x}_3 = (0.9, 0.08, 0.02) \in \mathcal{S}^3$ with respect to the reference measure P_0 with weights $\mathbf{p}_0 = (1, 1, 1)$. With $\mathbf{y}_i = \mathbf{x}_i/\mathbf{p}$, square-distance curves are $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2)$ (black), $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_3)$ (blue), $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_2, \mathbf{y}_3)$ (red). Reference measure is $\mathbf{p} = (1, 1, \epsilon)$ and x-axis is scaled as $P\{\Omega\} = 1 + 1 + \epsilon$. The three square-distances monotonically increase from $P\{\Omega\} = 2$ to $P\{\Omega\} = 3$. The end points of the curves at $P\{\Omega\} = 2$ and $P\{\Omega\} = 3$ are equal to standard Aitchison square-distances in $\mathcal{S}^2$ and $\mathcal{S}^3$ respectively.

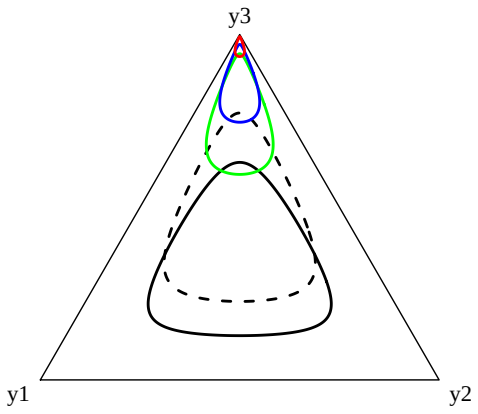

Figure 3: The unit circle (black, full line) in the uniform reference. After change of origin to $(1, 1, \epsilon)$, $\epsilon = 0.5$ (black, dashed), 0.1 (green), 0.05 (blue), and 0.01 (red), the circle is shifted towards the vertex y_3.

can be stated as follows.

PROPOSITION (dominance of distances) *Let* $\mathbf{x}_1$, $\mathbf{x}_2$ *be two compositions in* $\mathcal{S}^D$, *endowed with the reference measure* P_0, *with weights* $\mathbf{p}_0 = (1, 1, \ldots, 1)$. *Consider two reference measures,* P_1 *and* P_2, *represented by their respective weights* $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$ *and* $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$, *such that all their components are* $0 < p_{ki} \leq 1$, *for* $k = 1, 2$, $i = 1, 2, \ldots, D$ *and* $P_k\{\Omega\} = \sum_{i=1}^{D} p_{ki}$. *Define* $\mathbf{y}_j^{(\mathbf{p}_k)} = \mathbf{x}_j / \mathbf{p}_k$ *for* $k = 1, 2$ *and* $j = 1, 2$. *Then,*

$$p_{1i} \leq p_{2i},\ i = 1, 2, \ldots, D \quad \Rightarrow \quad \mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)}) \leq \mathrm{d}_{\mathbf{p}_2}(\mathbf{y}_1^{(\mathbf{p}_2)}, \mathbf{y}_2^{(\mathbf{p}_2)}) \ .$$

It is worth to remark that the notation of distances like $\mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)})$ could be changed to $\mathrm{d}_{\mathbf{p}_1}(\mathbf{x}_1, \mathbf{x}_2)$, as distances assigned to shifted $\mathbf{y}$'s are equal to those of the original compositions $\mathbf{x}$'s. This is due to the fact that $\mathbf{x}$ and $\mathbf{y}$ are densities of the same measure, namely $\mu_{\mathbf{x}}$, with respect to different reference measures.

Figure 2 shows the evolution of square-distances between three compositions $\mathbf{x}_1 = (0.1, 0.7, 0.2)$, $\mathbf{x}_2 = (0.5, 0.3, 0.2)$, $\mathbf{x}_3 = (0.9, 0.08, 0.02)$ with respect to the uniform reference in $\mathcal{S}^3$ when the reference measure changes progressively. The reference measure is $(1, 1, \epsilon)$, with ϵ going from 0 to 1. The plot is scaled according the $P\{\Omega\} = 1+1+\epsilon$. The square-distances increase monotonically, from distances corresponding to the subcomposition (y_1, y_2) to square-distances with the standard reference $\mathbf{p}_0 = (1, 1, 1)$. This result is expected after the previous proposition.

An experiment has been conducted to show how the changes of reference modify distances and shapes. Five different reference measures $\mathbf{p} = (1, 1, \epsilon)$ have been considered with ϵ equal to 1, 0.5, 0.1, 0.05, 0.01, so that they approach progressively the geometry of the subcomposition of the two first parts. The unit circle centered at the neutral element was shifted by the five reference measures. Figure 3, shows this unit circle (black) and the sequence of perturbations as a consequence of the change of origin. Note that the transformed circle is shifted to the vertex which weight is reduced, as expected after dividing each part by the corresponding weight.

After the change of origin, each point on the circles was ilr-transformed using the corresponding weights according to the SBP sign code

y_1	y_2	y_3
+1	−1	−1
0	+1	−1

,

which has been selected to avoid a balance representing the subcomposition (y_1, y_2). Figure 4 (left panel) shows the coordinates of the circles, to show the changes of the distances between points on the same circle. Note that the centers of the ellipses do not coincide, as they correspond to the closure of the reference measure $(1, 1, \epsilon)$. The main feature is the progressive stretch of the original circle. For very small ϵ the ellipse tends to degenerate into a segment following the direction of the subcomposition (y_1, y_2). Similarly, Figure 4 (right panel) shows the deformation of a grid originally at $-1, 0, 1$ in both axes (black). The new references are $\epsilon = 0.1$ (blue), and 0.01 (red). The grid is progressively tilted and distances between nodes decrease as ϵ decreases. Although straight-lines are preserved, their angles change, thus showing the change of geometry when changing the reference.

5. Elementary statistics

The change of reference measure and its associated weighting have consequences in the definitions of elementary concepts of compositional statistics. Variability and center are the two main concepts examined below. Both concepts are redefined following previous developments in the statistical analysis of compositional data, just looking for the influence of the weighting. These new definitions are intended to match the standard concepts whenever the weights are unity over the categories defining the composition.

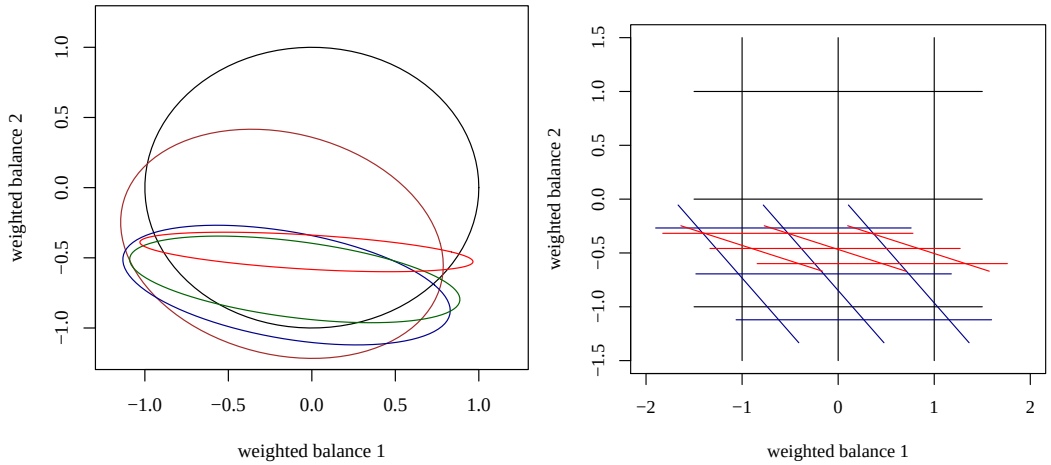

Figure 4: Left panel: the five circles in Figure 3 after weighted ilr-transformation. Reference measures are $(1, 1, \epsilon)$, $\epsilon = 1$ (black), $\epsilon = 0.5$ (brown), 0.1 (blue), 0.05 (green), and 0.01 (red). Right panel: a regular grid at points -1, 0, -1 in both axes after change of origin and weighted ilr transformation. Weights are $(1, 1, \epsilon)$, $\epsilon = 1$ (black), 0.1 (blue), and 0.01 (red).

Let $\mathbf{X}$ be a random composition (density) in $\mathcal{S}^D$ (Pawlowsky-Glahn *et al.* 2015, ch. 6) which, for some selected ilr coordinates denoted $\mathbf{X}^*$ in $\mathbb{R}^{D-1}$, is absolutely continuous with joint probability density (pdf) $f^*_{\mathbf{X}}$. Therefore, $f^*_{\mathbf{X}}(\mathbf{x})$ is a function defined on $\mathbb{R}^{D-1}$, the space of the ilr coordinates, with the standard definitions from probability theory. Assume also that a new reference measure is chosen and it is represented by a set of positive weights $\mathbf{p}$. Accordingly, the random composition $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$ corresponds to the change of reference and its distribution only differs from that of $\mathbf{X}$ in a shift of the center. The $\text{ilr}_{\mathbf{p}}$ coordinates of $\mathbf{Y}$, denoted $\mathbf{Y}^*$, are also random, but their distribution on $\mathbb{R}^{D-1}$ is a transformation of the previous pdf $f^*_{\mathbf{X}}$, here denoted as f^*, where the subscript is dropped when it corresponds to the composition $\mathbf{Y}$. In Appendix B it is shown that the transformation from $\mathbf{X}^* = \text{ilr}(\mathbf{X})$ to $\mathbf{Y}^* = \text{ilr}_{\mathbf{p}}(\mathbf{Y})$ is a linear (affine) transformation. For instance, this means that, if $\mathbf{X}$ has a normal distribution on the simplex (Mateu-Figueras, Pawlowsky-Glahn, and Egozcue 2013; Pawlowsky-Glahn *et al.* 2015) and, thus, $\mathbf{X}^*$ is multivariate normal on $\mathbb{R}^{D-1}$, the distribution of $\mathbf{Y}^*$ is also a multivariate normal on $\mathbb{R}^{D-1}$. As a conclusion, the normality of $\text{ilr}_{\mathbf{p}}$ coordinates is maintained when the weights $\mathbf{p}$ of the reference measure change.

Following the general formalism developed by Fréchet (1948) for metric spaces, the first milestone to be defined is the (total) variability of $\mathbf{Y}$ with respect to an arbitrary point $\boldsymbol{\eta} \in \mathcal{S}^D$. It is defined as

$$\text{totVar}_{\mathbf{p}}[\mathbf{Y}; \boldsymbol{\eta}] = \text{E}[\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta})] \ ,$$

provided that the expectation exists. The distance $\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta})$ is a function of the coordinates $\mathbf{Y}^* = \text{ilr}_{\mathbf{p}}(\mathbf{Y})$ and the expectation $\text{E}[\cdot]$ is taken with respect to their pdf f^*. Since $\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta}) = \sum (Y^*_i - \eta^*_i)^2$ (Equation 12), the minimum of $\text{Var}_{\mathbf{p}}[\mathbf{Y}; \boldsymbol{\eta}]$ is attained for $\boldsymbol{\eta}^* = \text{E}[\mathbf{Y}^*]$, a standard result in real multivariate statistics. Based on this result, the weighted center and total variance are

$$\text{Cen}_{\mathbf{p}}[\mathbf{Y}] = \text{ilr}^{-1}_{\mathbf{p}}(\text{E}[\mathbf{Y}^*]) = \text{clr}^{-1}_{\mathbf{p}}(\text{E}[\mathbf{Y}^*]) \quad , \quad \text{totVar}_{\mathbf{p}}[\mathbf{Y}] = \text{E}[\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \text{Cen}_{\mathbf{p}}[\mathbf{Y}])] \ . \tag{13}$$

Note that this kind of approach has been used in Pawlowsky-Glahn and Egozcue (2001) and in Boogaart and Tolosana-Delgado (2013), but total variance is there called metric variance.

Despite the previous expression of $\text{Cen}_{\mathbf{p}}[\mathbf{Y}]$ in Equation 13, the weighted center of a random composition only depends on the weights in $\mathbf{p}$ through the shift applied, that is

$$\text{Cen}_{\mathbf{p}}[\mathbf{Y}] = \text{Cen}[\mathbf{X}] \ominus \mathbf{p} \ , \quad \text{or, equivalently,} \quad \text{Cen}[\mathbf{X}] = \text{Cen}_{\mathbf{p}}[\mathbf{Y}] \oplus \mathbf{p} \ ,$$

where Cen and $\oplus$ are the ordinary center and perturbation of random compositions, respectively, and $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$, thus enhancing the linearity of expectations.

Decompositions of total variance underlays many standard statistical methods, thus remarking its upmost importance. Equation 13 leads to decompositions of the total variance when the reference measure $\mathbf{p}$ is not $\mathbf{p}_0$. Similarly to those described in Egozcue *et al.* (2011), we obtain

$$\begin{aligned} \text{totVar}_{\mathbf{p}}[\mathbf{Y}] &= \sum_{i=1}^{D-1} \text{Var}[\ \text{ilr}_{\mathbf{p},i}(\mathbf{Y})\ ] \\ &= \sum_{i=1}^{D} p_i \text{Var}[\ \text{clr}_{\mathbf{p},i}(\mathbf{Y})\ ] \\ &= \frac{1}{2s_{\mathbf{p}}} \sum_{i=1}^{D}\sum_{j=1}^{D} p_i p_j \text{Var}\left[\ln \frac{Y_i}{Y_j}\right] \ , \end{aligned} \tag{14}$$

where $s_{\mathbf{p}} = \sum_{i=1}^{D} p_i$, $\text{ilr}_{\mathbf{p},i}(\mathbf{Y}) = y_i^*$ and $\text{clr}_{\mathbf{p},i}(\mathbf{Y})$ is the i-th component of $\text{clr}_{\mathbf{p}}(\mathbf{Y})$. Note that the decomposition of $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ into $\text{ilr}_{\mathbf{p}}$ variance components points out that $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ is the trace of the covariance matrix of $\text{ilr}_{\mathbf{p}}(\mathbf{Y})$, and that $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ is not the sum of $\text{clr}_{\mathbf{p}}$ variances, but a weighted sum of them.

The decompositions of the total variance are closely related to the relationships between the covariance matrices of the $\text{ilr}_{\mathbf{p}}$ coordinates and the $\text{clr}_{\mathbf{p}}$ coefficients. These relationships can be summarized as

$$\Sigma_{\mathbf{p}} = \Psi\ \text{diag}(\mathbf{p})\ \Sigma_{\mathbf{p}}^{c}\ \text{diag}(\mathbf{p})\ \Psi^{\top} \quad , \quad \Sigma_{\mathbf{p}}^{c} = \Psi\ \Sigma_{\mathbf{p}} \Psi^{\top} \ ,$$

where Ψ is the $(D-1, D)$-contrast matrix of the $\text{ilr}_{\mathbf{p}}$, $\Sigma_{\mathbf{p}}$ is the covariance matrix of $\mathbf{Y}^*$ and $\Sigma_{\mathbf{p}}^{c}$ is the covariance matrix of $\text{clr}_{\mathbf{p}}(\mathbf{Y})$.

Also, the variation matrix (Aitchison 1986) plays an important role in the statistics of compositional data. Its entries are variances of simple log-ratios, $\ln(X_i/X_j)$. At least, it has two important uses: (a) it constitutes a simple and interpretable representation of the variability (second order moments) of the random composition, identifying the binary sources of variability relative to the total variance; and (b) each entry of the variation matrix is a measure of the compositional dissociation, as opposite of association, between the two parts involved. Point (a) is reflected in the fact that the covariance matrices of ilr-coordinates and clr coefficients can be retrieved from the variation matrix (Pawlowsky-Glahn *et al.* 2015, Appendix A). Concerning point (b), large entries, relative to other entries, point out most dissociated pairs of parts. The measurement of compositional association of two parts, understood as proportionality between them, is motivated by the fact that $\text{Var}[\ln(X_i/X_j)] = 0$ implies that X_i and X_j are strictly proportional (Egozcue, Lovell, and Pawlowsky-Glahn 2013a; Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler 2015).

Inspired by the third decomposition of weighted total variance in Equation 14, a weighted variation matrix can be defined as a (D, D)-matrix $T_{\mathbf{p}}$ with entries

$$t_{\mathbf{p},i,j} = p_i p_j \text{Var}\left[\ln \frac{Y_i}{Y_j}\right] \quad , \quad i, j = 1, 2, \ldots, D \ .$$

The relationship of $T_{\mathbf{p}}$ with the covariance matrix of $\text{ilr}_{\mathbf{p}}$ coordinates is

$$\Sigma_{\mathbf{p}} = -\frac{1}{2} \Psi T_{\mathbf{p}} \Psi^{\top} \ .$$

The decomposition of weighted total variance and the relationships between covariance matrices reduce to the standard ones whenever the reference measure is $P = P_0$, that is, whenever $\mathbf{p} = (1, 1, \ldots, 1)$.

Table 2: Weights, $\mathbf{p}$ (second row) and $p_i^{(sub)}$ (third row), for each part used in the analysis of the Cat10 data set. Weights $p_i^{(sub)}$ are only used in an example of biplot. Center of the composition, expressed in percent, for the original composition (forth row), and for the shifted composition $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$ (fifth row).

party	abs	nota	null	C's	CiU	ERC	ICV	PSC	PP	other
p_i	0.1	0.3	0.3	1	1	1	1	1	1	0.5
$p_i^{(sub)}$	0.001	0.001	0.001	1	1	1	1	1	1	0.001
Cen[$\mathbf{X}$] (%)	38.9	1.9	0.5	0.9	27.6	5.5	3.0	9.6	5.2	6.8
Cen$_{\mathbf{p}}$[$\mathbf{Y}$] (%)	84.1	1.4	0.4	0.2	6.0	1.2	0.6	2.1	1.1	2.9

6. Exploratory tools

In compositional data analysis, the main specific exploratory tools are the variation matrix (Aitchison 1986), principal component analysis of the clr transformed compositional sample (Aitchison 1983) and its corresponding biplots (Aitchison and Greenacre 2002), and the compositional dendrogram (Pawlowsky-Glahn and Egozcue 2011). These three tools take slightly different forms when taking a reference measure different from P_0. In order to show how to use and interpret the weighted versions in an exploratory analysis, the data from the Catalan parliament (Spain) elections in November 2010 (Cat10) have been selected. This data set was previously analysed in Egozcue and Pawlowsky-Glahn (2011) (see also Pawlowsky-Glahn *et al.* 2015).

The data set Cat10 contains the number of votes obtained by several parties, including abstention (abs), null (null) and none of the above or blank votes (nota) in $n = 41$ electoral districts. The major parties contesting the elections were *Convergència i Unió* (CiU), *Partit dels Socialistes de Catalunya* (PSC), *Ciutadans-Partido de la Ciudadanía* (C's), *Esquerra Republicana de Catalunya* (ERC), *Iniciativa per Catalunya Verds-Esquerra Unida i Alternativa* (ICV) and *Partit Popular* (PP). Other minor parties are amalgamated in *other*. The present analysis focusses on the whole composition of votes, that is, the $D = 10$ parts of the composition: abs, nota, null, CiU, C's, ERC, ICV, PP, PSC, other.

A first step in exploratory analysis is to choose suitable weights for the 10 parts involved. The situation in most political elections is that votes to parties show a homogeneous preference to a given party, meanwhile "abs", "nota", "null" and "other" mix non-homogeneous support to democratic elections or other situations, thus suggesting to weight them differently. Well defined parties were weighted by 1. The abstention is the more heterogeneous group of electors and the choice for its weight was 0.1. The electors that choose blank vote (nota) and null vote (null) can be considered less heterogeneous than abstention, as they express something similar to "I want to vote, but none of the contesting parties convinced me"; these two categories have been weighted by 0.3. Votes to parties included in "other" are well defined, but directed to different parties with different programmes; there is a well defined intention in the vote, but the amalgamation of different parties makes the group heterogeneous; the category "other" is weighted by 0.5. The vector of weights $\mathbf{p}$ chosen is shown in the second row of Table 2. These weights have been chosen to show the effects of weighting, and not to carry out a sound analysis of the data set. Methods to establish suitable weights should be object of further research. The third row of Table 2 shows an alternative set of weights $p_i^{(sub)}$ that will be used only for illustrating how these weights make the analysis to be close to that of a subcomposition of the well defined parties. The forth row of Table 2 shows the center of the composition, expressed in percent. The fifth row is the center Cen$_{\mathbf{p}}$[$\mathbf{Y}$] (also in percent), which is not useful for interpretation, but for comparison with Cen[$\mathbf{X}$]. Note how the percent of "abs", with weight 0.1, increased when dividing by the weight. The same fact may occur for all parts with weights less than one, but closure hides this fact. Note that the center is

Table 3: Weighted variation matrix for Cat10 data. Last column: weighted $\text{clr}_{\mathbf{p}}$ variances, $p_i \text{Var}(\text{clr}_{\mathbf{p},i}[\mathbf{Y}])$, adding to weighted total variance. Upper triangle: elements of the weighted variation matrix (values greater than or equal to 0.30 are highlighted in boldface). Lower triangle: product of weights $p_i p_j$. Two last rows: weighted total variance and total variance (uniform reference).

	Abs	Nota	Null	C's	CiU	ERC	ICV	PSC	PP	other	$\text{clr}_{\mathbf{p}}$ var.
Abs		0.002	0.005	0.029	0.007	0.020	0.008	0.006	0.011	0.007	0.001
Nota	0.03		0.008	0.164	0.009	0.027	0.040	0.031	0.073	0.017	0.014
Null	0.03	0.09		0.238	0.022	0.015	0.078	0.064	0.111	0.020	0.039
C's	0.10	0.09	0.30		**0.563**	**0.870**	0.270	**0.320**	0.160	**0.303**	**0.308**
CiU	0.10	0.30	0.30	1.00		0.077	0.157	0.142	0.268	0.034	0.054
ERC	0.10	0.30	0.30	1.00	1.00		0.249	0.262	**0.459**	0.052	0.153
ICV	0.10	0.30	0.30	1.00	1.00	1.00		0.101	0.189	0.085	0.051
PSC	0.10	0.30	0.30	1.00	1.00	1.00	1.00		0.122	0.133	0.051
PP	0.10	0.30	0.30	1.00	1.00	1.00	1.00	1.00		0.193	0.113
other	0.05	0.15	0.15	0.50	0.50	0.50	0.50	0.50	0.50		0.054
$\text{totVar}_{\mathbf{p}}$											0.836
totVar											1.020

a composition of a "mean electoral district", and that variability around this center may be large. This can be checked, for instance, on C's, which minimum percentage is 0.3% and its maximum is 2.8% across the sample of electoral districts, what in turns may represent a number of electors from 3046 up to 1,572,425 for the surroundings of Barcelona. Therefore, reporting mean values or centers needs to be complemented with the analysis of variability.

The weighted variation matrix is shown in the upper triangle of Table 3. In the lower triangle of Table 3, the cross products of weights $p_i p_j$ are specified. When the entries of the weighted variation matrix are divided by the corresponding $p_i p_j$ they result in the corresponding entry of the traditional variation matrix with reference P_0. Terms in the weighted variation matrix larger than or equal to 0.30 are highlighted in boldface. They constitute the larger sources of variability in the data set. Most of them correspond to C's, whose votes are irregularly distributed over electoral districts. This fact is confirmed by the weighted $\text{clr}_{\mathbf{p}}$ variances, as the largest value corresponds to C's as well. Small values in the weighted variation matrix suggest association between parts, i.e. approximate proportionality, although this needs further analysis to be confirmed (Egozcue *et al.* 2013a; Lovell *et al.* 2015). The strongest associations appear between abs, nota, null, with traditionally nationalist parties in Catalonia, i.e. CiU, ERC, and even with PSC. Compared to the variation matrix published in Egozcue and Pawlowsky-Glahn (2011), the possible associations appear stronger in Table 3. This is due to the fact that the 2011 analysis was performed without any weighting in the reference. Differences in the variances of simple log-ratios of not down-weighted parts are the consequence of dividing entries in Table 3 by $n-1=40$, while in 2011 the divisor was $n=41$. The weighted total variance is 0.836, smaller than that obtained with unit weights (1.020), using in both cases the same divisor ($n-1=40$).

In compositional data analysis, principal component analysis (PCA) is commonly performed using the singular value decomposition (SVD) of the clr-transformed data set (Aitchison 1983). The scores, multiplied by the singular values, are proportional to ilr-coordinates, such that their variances are proportional to the square singular values. The loadings matrix contains the clr representation of the principal directions. The last singular value is zero, as the clr data sum to zero for each data point. Similar features are expected for a PCA performed on a weighted composition using its weighted $\text{clr}_{\mathbf{p}}$ transformed values. However, when the $\text{clr}_{\mathbf{p}}$-transformed data set is SVD-decomposed, the square singular values are no longer proportional to $\text{ilr}_{\mathbf{p}}$ variances and they do not provide a decomposition of the weighted total variance. The way proposed here consists of dividing the $\text{clr}_{\mathbf{p}}$ data previous to SVD, so

that resulting square singular values add to the total variance.

Let X be a compositional data set in $\mathcal{S}^D$; therefore, X is a (n, D)-matrix and n is the size of the sample. After selecting some positive weights, $\mathbf{p}$, each row of the data matrix is accordingly shifted and is written as $Y = X \ominus \mathbf{p}$. Applying the $\mathrm{clr}_{\mathbf{p}}$ transformation to each row yields $\mathrm{clr}_{\mathbf{p}}(Y)$. This $\mathrm{clr}_{\mathbf{p}}$-transformed data set is centered and weighted with the square-root of the weights in $\mathbf{p}$, that is

$$A = [\mathrm{clr}_{\mathbf{p}}(Y) - (\overline{\mathrm{clr}_{\mathbf{p}}(Y)})]\ \mathrm{diag}(\sqrt{\mathbf{p}})\ ,$$

where $\overline{\mathrm{clr}_{\mathbf{p}}(Y)}$ denotes the average by columns of $\mathrm{clr}_{\mathbf{p}}(Y)$. The SVD of A,

$$A = U\Lambda V^{\top}\ ,$$

has the standard properties of an SVD. Among these properties, some of them are reinterpreted in the compositional framework. The singular values contained in the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{D-1}, 0)$ are positive and in decreasing order of magnitude; the last one is zero due to the property of the $\mathrm{clr}_{\mathbf{p}}(Y)$ that the weighted sum of its components adds to zero (Equation 4). The non-standardized scores $U\Lambda$ are $\mathrm{ilr}_{\mathbf{p}}$-coordinates whose sample variances are $\lambda_i^2/(n-1)$. The sample total variance is $\mathrm{totVar}_{\mathbf{p}}(Y) = \sum_{i=1}^{D-1} \lambda_i^2/(n-1)$. The $D-1$ first columns of $V^{\top}$ contain the contrast matrix corresponding to the $\mathrm{ilr}_{\mathbf{p}}$. The loadings are given by the columns of $\mathrm{diag}(1/\sqrt{\mathbf{p}})V\Lambda$, where $\mathrm{diag}(1/\sqrt{\mathbf{p}})$ appears to compensate the previous weighting in A.

A covariance biplot (Aitchison and Greenacre 2002) is a simultaneous projection of U (scores) and $V\Lambda$ (loadings) onto two principal directions, usually the first two. The percent of weighted total variance explained in such a projection is given by

$$100\ \frac{\lambda_1^2 + \lambda_2^2}{\sum_{i=1}^{D-1} \lambda_i^2}\ .$$

This kind of biplots have been obtained for the data set Cat10. Figure 5 shows four different cases: top-left panel shows the biplot when the reference is $\mathbf{p}_0 = (1, 1, \ldots, 1)$; top-right panel adopts the weights p_i shown in Table 2; bottom-left panel shows the biplot when using $p_i^{(sub)}$ also shown in Table 2 (third row). Finally, the bottom-right panel shows the biplot obtained using the subcomposition of individual parties, excluding "abs", "nota", "null", and "other", and using the reference $\mathbf{p}_0$ for the subcomposition.

The first impression is that the two biplots in the upper part of Figure 5 appear to be quite similar, as the main features are preserved. In fact, the clr-variables corresponding to well defined parties are projected very similarly. For instance, the first principal axis is dominated by the clr-variables corresponding to C's on one side, and CiU and ERC on the opposite side, which can be identified with a balance of non-nationalist *versus* nationalist Catalan parties; this fact was previously observed in the weighted variation matrix. The second principal axis is mainly influenced by the links between PP-ICV and PSC-ICV, leading to identify the second principal axis with a balance of right *versus* left wing parties. In fact, the three parties involved are perceived by electors as right wing (PP), very moderate social-democratic (PSC) and left wing (ICV). However, when looking at the clr-variables corresponding to down-weighted parts (abs, null, nota, other), the shortening of the corresponding parts proportional to $\sqrt{p_i}$ is apparent. For example, the role of clr-other in the projection has been reduced in an appreciable way.

In the bottom-left panel of Figure 5, the weights $p_i^{(sub)}$ (Table 2) have been used in order to approach a subcompositional analysis of the parties C's, CiU, ERC, ICV, PP, PSC. As the rest of the parts are severely down-weighted, they appear as very short rays from the origin (labels are overlapping). Compared with the subcompositional analysis (bottom-right panel, Figure 5), it is clear that, exception made of these short rays, the rest is almost identical in the two bottom biplots. See, for instance, that the total variance of the two cases are, respectively, 0.7020 (weights $p_i^{(sub)}$) and 0.7016 (unit weights in the subcomposition) and

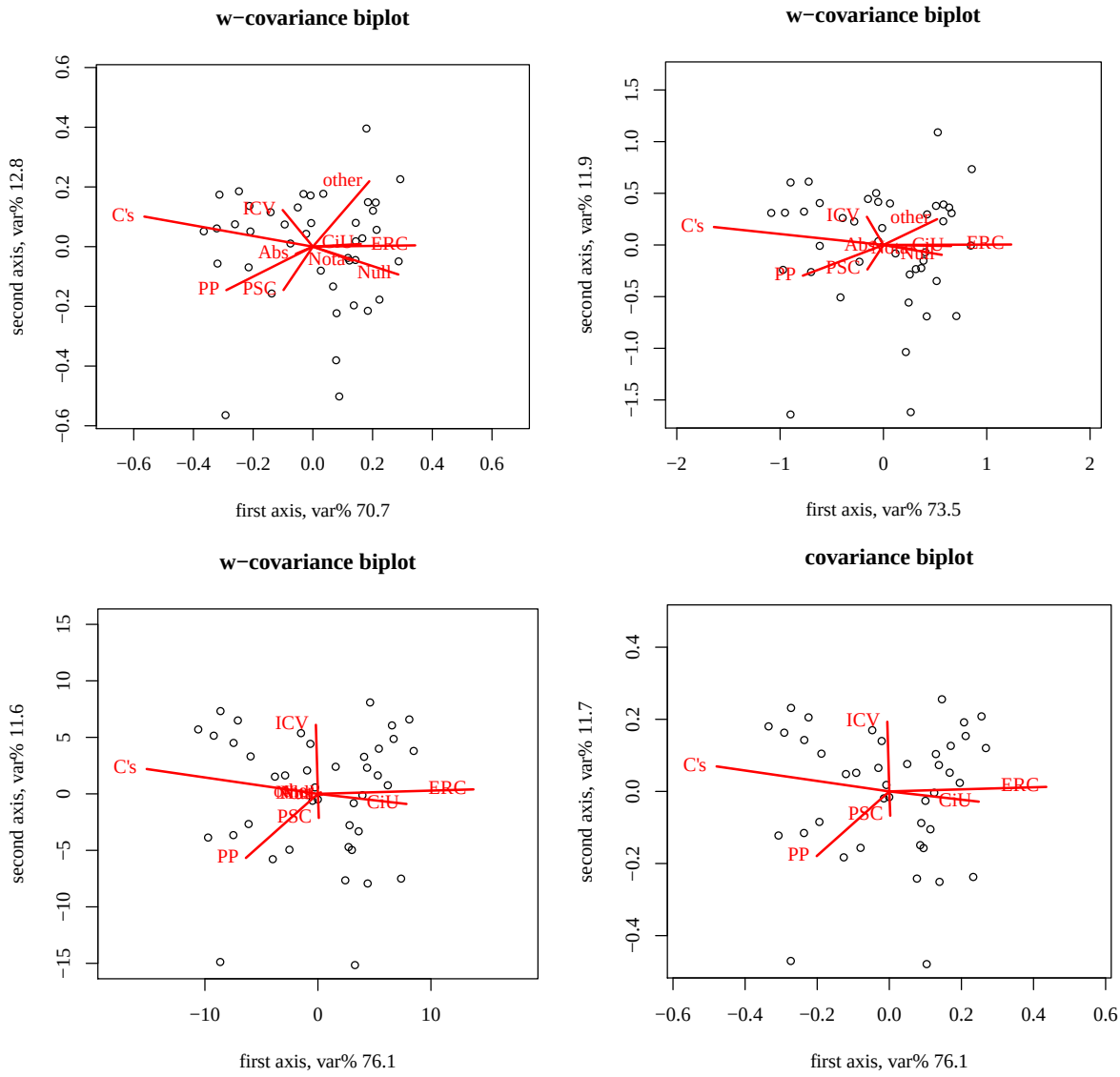

Figure 5: Covariance biplots of Cat10 dataset. Top-left panel: uniform reference $\mathbf{p}_0 = (1, 1, \ldots, 1)$, total variance 1.020. Top-right panel: weights given in Table 2, weighted total variance 0.836. Bottom-left panel: extreme weighting, given in Table 2, weighted total variance 0.7020. Bottom-right panel: subcomposition of parties, total variance 0.7016.

the corresponding proportions of explained total variance in the two dimensional projections are very close. This illustrates the fact, that down-weighting some parts is a path towards subcompositional analysis.

The fact that the projection changes only slightly from top to bottom of Figure 5 indicates that most of the variance introduced by "abs" is small (see Table 3) and that of "nota" and "null" is not well represented in the first and second principal axes. A feature that is clear in the weighted biplot (top-right panel) is that the link "null-other" is almost parallel to the second axis and to the link PSC-ICV: the variance of this two log-ratios are mainly included in the second principal component. The "nota" and "null" votes are quite associated one to each other across electoral districts as the rays appear almost parallel (see also Table 3). When they are down weighted (top-right panel) the main effect is that the corresponding rays are equally shortened as the weights were equal for these two parts.

The so called balance-dendrogram is not discussed here in detail, as the changes to be incorporated when using weights are quite obvious. Firstly, a balance-dendrogram presents

a hierarchical structure describing an SBP, which in the weighted case is identical to the standard case. The decomposition of the total variance changes quantitatively with weighting, as indicated in Equation 14 (second member). Finally, the position of mean balances is substituted by the new mean weighted balances. However, the qualitative structure of the dendrogram remains the same.

The present study of different exploratory tools for compositional data analysis is only preliminary. Details on interpretation and methods to assess weights require further study.

7. Conclusions and further research

A weighting strategy for the analysis of compositions is proposed. It is based on the theory of Bayes Hilbert spaces. However, some modifications have been introduced to fulfill the principle of dominance of distances when down-weighting some parts of the composition. When the weights considered are unitary in each part, that is, when there is no down or up-weighting, the approach is reduced to the standard compositional data analysis. If some parts are down-weighted approaching zero, the weighted geometry of the simplex tends to the ordinary Aitchison geometry of the corresponding subcomposition.

In order to use the proposed weighting approach, it is advisable to deal with compositional data as usual for linear operations, using the standard perturbation and powering. When distances or inner products are involved in the analysis, they are computed in two steps: first, shifting the compositional data by $\ominus\mathbf{p}$, that is, dividing each part by the corresponding weight; and second, computing $\mathrm{clr}_{\mathbf{p}}$ (Equations 3 or 16) or $\mathrm{ilr}_{\mathbf{p}}$ (Equation 10) to find the required distances or inner products in a straightforward way.

Statistical consequences of weighting compositions need to be studied in the future. Standard tools of exploratory analysis, as variation matrix, biplots or balance-dendrogram, clustering and others, will be influenced by weighting. The reason is that distances between compositions and computation of variances-covariances are influenced as well. Thus, the proposed weighting approach is only a first step towards developing effective weighting techniques applicable to compositional data analysis.

Acknowledgements

This research has been supported by the *Spanish Ministry of Education and Science* under projects 'METRICS' (Ref. MTM2012-33236) and 'CODA-RETOS' (Ref. MTM2015-65016-C2-1-R); and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under the project Ref. 2009SGR424. We thank the deep and detailed revision by K. Hron and an anonymous reviewer, which lead to improvements of the original contribution.

References

Aitchison J (1983). "Principal component analysis of compositional data." *Biometrika*, **70**(1), 57–65.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press), London (UK). ISBN 0-412-28060-4. 416 p.

Aitchison J (1992). "On Criteria for Measures of Compositional Difference." *Mathematical Geology*, **24**(4), 365–379.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000). "Logratio

analysis and compositional distance." *Mathematical Geology*, **32**(3), 271–275. ISSN 0882-8121.

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

Aitchison J, Greenacre M (2002). "Biplots for compositional data." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **51**(4), 375–392.

Boogaart KGvd, Egozcue JJ, Pawlowsky-Glahn V (2010). "Bayes linear spaces." *SORT - Statistics and Operations Research Transactions*, **34**(2), 201–222. ISSN 1696-2281.

Boogaart KGvd, Egozcue JJ, Pawlowsky-Glahn V (2014). "Bayes Hilbert Spaces." *Australian and New Zealand Journal of Statistics*, **56**(2), 171–194. doi:10.1111/anzs.12074.

Boogaart KGvd, Tolosana-Delgado R (2013). *Analysing compositional data with R*. Springer, Heidelberg. 280 pp.

Egozcue JJ (2009). "Reply to "On the Harker variation diagrams;..." by J. A. Cortés." *Mathematical Geosciences*, **41**(7), 829–834.

Egozcue JJ, Barceló-Vidal C, Martín-Fernández JA, Jarauta-Bragulat E, Díaz-Barrero JL, Mateu-Figueras G (2011). "Elements of simplicial linear algebra and geometry." In Pawlowsky-Glahn and Buccianti (2011), pp. 141–157. 378 p.

Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2006). "Hilbert space of probability density functions based on Aitchison geometry." volume 22, pp. 1175–1182. DOI: 10.1007/s10114-005-0678-2.

Egozcue JJ, Lovell D, Pawlowsky-Glahn V (2013a). *Testing compositional association*. In: Proceedings of the 5th Workshop on compositional data analysis, CoDaWork 2013, ISBN: 978-3-200-03103-6. Pp 28–36.

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

Egozcue JJ, Pawlowsky-Glahn V (2006). "Simplicial geometry for compositional data." In *Compositional Data Analysis in the Geosciences: From Theory to Practice*, pp. 145–159.

Egozcue JJ, Pawlowsky-Glahn V (2011). "Basic concepts and procedures." In Pawlowsky-Glahn and Buccianti (2011), pp. 12–28. 378 p.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric logratio transformations for compositional data analysis." *Mathematical Geology*, **35**(3), 279–300. ISSN 0882-8121.

Egozcue JJ, Pawlowsky-Glahn V, Tolosana-Delgado R, Ortego MI, Boogaart KGvd (2013b). "Bayes spaces: use of improper distributions and exponential families." *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas (RACSAM)*, **107**, 475–486. DOI 10.1007/s13398-012-0082-6.

Filzmoser P, Hron K (2015). "Robust coordinates for compositional data using weighted balances." *In Nordhausen, K. and Taskinen, S., (eds.), Modern Nonparametric, Robust and Multivariate Methods. Springer, Heidelberg.*

Fréchet M (1948). "Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié." *Annales de l'Institut Henri Poincaré*, **10**(4), 215–308.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015). "Proportionality: A Valid Alternative to Correlation for Relative Data." *PLoS Comput Biol*, **11**(3), e1004075. doi:10.1371/journal.pcbi.1004075. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1004075.

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2013). "The normal distribution in some constrained sample spaces." *SORT - Statistics and Operations Research Transactions*, **37**(1), 29–56. ISSN 1696-2281.

Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons. ISBN 978–0–470–71135–4. 378 p.

Pawlowsky-Glahn V, Egozcue JJ (2001). "Geometric approach to statistical analysis on the simplex." *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15**(5), 384–398.

Pawlowsky-Glahn V, Egozcue JJ (2011). "Exploring Compositional Data with the Coda-Dendrogram." *Austrian Journal of Statistics*, **40**(1 & 2), 103–113. ISSN 1026-597X.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and analysis of compositional data*. Statistics in practice. John Wiley & Sons, Chichester UK. ISBN 9781118443064. 272 pp.

A. Dominance of distances under change of reference

In Section 4 the following proposition was stated:

PROPOSITION (dominance of distances). *Let* $\mathbf{x}_1$, $\mathbf{x}_2$ *be two compositions in* $\mathcal{S}^D$, *endowed with the reference measure* P_0, *which weights are* $\mathbf{p}_0 = (1, 1, \ldots, 1)$. *Consider two reference measures,* P_1 *and* P_2, *represented by their respective weights* $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$ *and* $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$, *such that all their components are* $0 < p_{ki} \leq 1$, *for* $k = 1, 2$, $i = 1, 2, \ldots, D$ *and* $P_k(\Omega) = \sum_{i=1}^{D} p_{ki}$. *Define* $\mathbf{y}_j^{(\mathbf{p}_k)} = \mathbf{x}_j / \mathbf{p}_k$ *for* $k = 1, 2$ *and* $j = 1, 2$. *Then,*

$$p_{1i} \leq p_{2i},\ i = 1, 2, \ldots, D \quad \Rightarrow \quad \mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)}) \leq \mathrm{d}_{\mathbf{p}_2}(\mathbf{y}_1^{(\mathbf{p}_2)}, \mathbf{y}_2^{(\mathbf{p}_2)}) \ .$$

Proof: The change of reference from $\mathbf{p}_2$ to $\mathbf{p}_1$ with $p_{1i} \leq p_{2i}$, $i = 1, 2, \ldots, D$, can be conceived as a sequence of intermediate changes of reference for which only one weight p_{2i} is changed to p_{1i} at each step. These steps can be ordered, for instance, with the index $i = 1, 2, \ldots, D$. The sequence of weights can be the following.

step	initial reference		final reference
1-st	$\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$	to	$\mathbf{q}_1 = (p_{11}, p_{22}, \ldots, p_{2D})$
2-nd	$\mathbf{q}_1 = (p_{11}, p_{22}, \ldots, p_{2D})$	to	$\mathbf{q}_2 = (p_{11}, p_{12}, \ldots, p_{2D})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
i-th	$\mathbf{q}_{i-1} = (p_{11}, p_{12}, \ldots, p_{1,i-1}, p_{2i}, \ldots, p_{2D})$	to	$\mathbf{q}_i = (p_{11}, p_{12}, \ldots, p_{1i}, p_{2,i+1}, \ldots, p_{2D})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
D-th	$\mathbf{q}_{D-1} = (p_{11}, p_{12}, \ldots, p_{1,D-1}, p_{2D})$	to	$\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$

As one weight decreases at each step, the statement is proven if the distance $\mathrm{d}_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)})$ is less than or equal to $\mathrm{d}_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})})$, for $i = 1, 2, \ldots, D$, where $\mathbf{q}_D = \mathbf{p}_1$. The i-th step consists of changing the weight p_{2i} into p_{1i}, while all other weights remain equal. Consider that $\mathrm{ilr}_{\mathbf{q}_i}$ corresponds to a partition (SBP) that separates the i-th part of the composition

from the other $D-1$ parts. For both sets of weights $\mathbf{q}_i$ and $\mathbf{q}_{i-1}$ all weighted balances are equal except the first one, denoted $b_i^{(\mathbf{q}_k)}$, $k = i-1, i$. Equation 12 implies that

$$\mathrm{d}^2_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})}) - \mathrm{d}^2_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)})$$

$$= \left[b_i^{(\mathbf{q}_{i-1})}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}) - b_i^{(\mathbf{q}_{i-1})}(\mathbf{y}_2^{(\mathbf{q}_{i-1})})\right]^2 - \left[b_i^{(\mathbf{q}_i)}(\mathbf{y}_1^{(\mathbf{q}_i)}) - b_i^{(\mathbf{q}_i)}(\mathbf{y}_2^{(\mathbf{q}_i)})\right]^2 . \tag{15}$$

Using the expression of balances (11), it holds

$$b_i^{(\mathbf{q}_k)}(\mathbf{y}_\ell^{(\mathbf{q}_k)}) = \sqrt{\frac{q_{ki} n_i^-}{q_{ki} + n_i^-}} \log \frac{y_{\ell i}}{\prod_{j \neq i} y_{\ell j}^{q_{ki}/n_i^-}} , \quad k = i-1, i, \quad \ell = 1, 2,$$

where $q_{ki} = p_{2i}$ if $k = i-1$, and $q_{ki} = p_{1i}$ if $k = i$. Moreover, the values of the parts of the compositions are $y_{\ell j} = x_j / p_{2j}$ if $j < i$ and $y_{\ell j} = x_j / p_{1j}$ if $j > i$. As a result, the differences of balances in Equation 15 simplify to

$$b_i^{(\mathbf{q}_k)}(\mathbf{y}_1^{(\mathbf{q}_k)}) - b_i^{(\mathbf{q}_k)}(\mathbf{y}_2^{(\mathbf{q}_k)}) = \sqrt{\frac{q_{ki} n_i^-}{q_{ki} + n_i^-}} \log \left(\frac{x_{1i} x_{2i}}{\prod_{j \neq i} (x_j / q_{kj})^{q_{kj}/n_i^-}} \right) ,$$

where the closure constants associated with the change $\mathbf{x}_\ell = \mathbf{y}_\ell \oplus \mathbf{q}_k$ cancel within the balance, as it is scale invariant. Remarkably, the logarithmic term does not depend on k, as the weights q_{kj}, $j \neq i$, are equal for $\mathbf{q}_k$, $k = i-1, i$. Substituting these differences of balances in Equation 15 it yields

$$\mathrm{d}^2_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})}) - \mathrm{d}^2_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)}) = \frac{p_{2i} n_i^-}{p_{2i} + n_i^-} - \frac{p_{1i} n_i^-}{p_{1i} + n_i^-} = \frac{n_i^- (p_{2i} - p_{1i})}{(p_{2i} + n_i^-)(p_{1i} + n_i^-)} \geq 0 ,$$

since it was assumed that $p_{2i} \geq p_{1i}$. This proves the statement. □

B. Relationship between ordinary and weighted clr and ilr

In this appendix the relationship between ordinary and weighted clr and ilr is studied. The main goal is to prove that this relationship is linear up to additive terms. The expressions obtained are not central in the developed theory but they help to understand how the probability distributions of random compositions change under change of reference.

Let $\mathbf{x}$ be a composition in $\mathcal{S}^D$, taken as a density of a measure μ with respect to the uniform reference measure P_0, given by $\mathbf{p}_0 = (1, 1, \ldots, 1)$. An alternative reference measure represented by the weights $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ is considered, and the corresponding density of μ is then $\mathbf{y} = \mathbf{x}/\mathbf{p}$. The weighted centered log-ratio of $\mathbf{y}$ (Equation 3) is

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = \log \mathbf{y} - \log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1} ,$$

where $\mathrm{g}_\mathbf{p}(\cdot)$ denotes the weighted geometric mean of the arguments (Equation 3), $\mathbf{1}$ is a row vector of D ones and $\log \mathbf{x}$, $\log \mathbf{y}$ and $\log \mathbf{p}$ are taken as row vectors. Then, $\log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1}$ is a row vector with all components equal to $\mathrm{g}_\mathbf{p}(\mathbf{y})$. Moreover, $\log(\mathbf{y}) = \log(\mathbf{x}) - \log(\mathbf{p})$ and $\log(\mathrm{g}_\mathbf{p}(\mathbf{y})) = (1/\mathbf{s}_\mathbf{p}) \sum p_i (\log x_i - \log p_i)$, with $s_\mathbf{p} = \sum p_i$. Using matrix notation this leads to

$$\log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1} = \frac{1}{s_\mathbf{p}} (\log \mathbf{x} - \log \mathbf{p}) \, \mathbf{p}^\top \mathbf{1} .$$

Substitution into the definition of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ yields

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = (\log \mathbf{x} - \log \mathbf{p}) \left[I_D - \frac{1}{s_\mathbf{p}} \mathbf{p}^\top \mathbf{1} \right] , \tag{16}$$

where I_D is the (D, D)-identity matrix. Equation 16 shows that $\mathrm{clr}_{\mathbf{p}}(\mathbf{y})$ is a linear transformation of $\log \mathbf{x}$ up to additive terms depending on $\mathbf{p}$. The ordinary $\mathrm{clr}(\mathbf{x})$ can be written

$$\mathrm{clr}(\mathbf{x}) = \log \mathbf{x} \left[I_D - \frac{1}{D} \mathbf{1}^\top \mathbf{1} \right] ,$$

which can be substituted into Equation 16. The resulting expression is

$$\mathrm{clr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{clr}(\mathbf{x}) - \log(\mathbf{p}) - \log(\mathrm{g}_{\mathbf{p}}(\mathbf{x}))\mathbf{1} + \log(\mathrm{g}_{\mathbf{p}}(\mathbf{p}))\mathbf{1} + \log(\mathrm{g}(\mathbf{x}))\mathbf{1} . \tag{17}$$

In order to relate ordinary and weighted ilr, assume that ilr-coordinates of $\mathbf{x}$ are

$$\mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x})\ \Psi_0^\top \quad , \text{ with } \quad \Psi_0 \Psi_0^\top = I_{D-1} \ , \ \Psi_0^\top \Psi_0 = I_D - \frac{1}{D} \mathbf{1}^\top \mathbf{1} \ ,$$

that is, Ψ_0 is an ordinary contrast matrix (Egozcue *et al.* 2011). The weighted $\mathrm{ilr}_{\mathbf{p}}$-coordinates are computed as in Equation 10 using the weighted contrast matrix Ψ,

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{clr}_{\mathbf{p}}(\mathbf{y})\ \mathrm{diag}(\mathbf{p})\ \Psi^\top .$$

Substituting Equation 17 and taking into account that $\mathbf{1}\ \mathrm{diag}(\mathbf{p})\ \Psi^\top = \mathbf{0}$, it yields

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = (\mathrm{clr}(\mathbf{x}) - \log \mathbf{p})\ \mathrm{diag}(\mathbf{p})\ \Psi^\top .$$

Inserting $I_D = \Psi_0^\top \Psi_0 + (1/D)\mathbf{1}^\top \mathbf{1}$ after $\mathrm{clr}(\mathbf{x})$, the desired relationship is

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{ilr}(\mathbf{x})\ \Psi_0\ \mathrm{diag}(\mathbf{p})\ \Psi^\top - \log \mathbf{p}\ \mathrm{diag}(\mathbf{p})\ \Psi^\top , \tag{18}$$

which shows that $\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})$ is a linear transformation of $\mathrm{ilr}(\mathbf{x})$, up to additive terms depending only on $\mathbf{p}$ and the selected weighted contrast matrix Ψ.

Affiliation:

Juan José Egozcue
Universitat Politecnica de Catalunya
Jordi Girona 1-3, C2-UPC
E-08034 Barcelona, Spain
E-mail: juan.jose.egozcue@upc.edu

Vera Pawlowsky-Glahn
Universitat de Girona
Campus Montilivi, P4
E-17071 Girona, Spain
E-mail: vera.pawlowsky@udg.edu

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2015-09-23
Accepted: 2016-04-20

Austrian Journal of Statistics
September 2016, Volume 45, 45–56.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.136

Bayesian Estimation of the Orthogonal Decomposition of a Contingency Table

Maria Isabel Ortego
Universitat Politècnica de Catalunya, Spain

Juan José Egozcue
Universitat Politècnica de Catalunya, Spain

Abstract

In a multinomial sampling, contingency tables can be parametrized by probabilities of each cell. These probabilities constitute the joint probability function of two or more discrete random variables. These probability tables have been previously studied from a compositional point of view. The compositional analysis of probability tables ensures coherence when analysing sub-tables. The main results are: (1) given a probability table, the closest independent probability table is the product of their geometric marginals; (2) the probability table can be orthogonally decomposed into an independent table and an interaction table; (3) the departure of independence can be measured using simplicial deviance, which is the Aitchison square norm of the interaction table.

In previous works, the analysis has been performed from a frequentist point of view. This contribution is aimed at providing a Bayesian assessment of the decomposition. The resulting model is a log-linear one, which parameters are the centered log-ratio transformations of the geometric marginals and the interaction table. Using a Dirichlet prior distribution of multinomial probabilities, the posterior distribution of multinomial probabilities is again a Dirichlet distribution. Simulation of this posterior allows to study the distribution of marginal and interaction parameters, checking the independence of the observed contingency table and cell interactions.

The results corresponding to a two-way contingency table example are presented.

Keywords: interaction, independence, simplicial deviance, multinomial sampling, Aitchison geometry of the simplex, orthogonal decomposition, R.

1. Introduction

Contingency tables have been studied for a long time. There are many examples, dating from the beginning of the XX-th century which afforded elementary, but relevant, questions about such kind of data (e.g. Yule 1912). Along the XX-th century many advances have been achieved. The introduction of log-linear models (Nelder 1974) and generalized linear models (McCullagh and Nelder 1983; Nelder and Wedderburn 1972) were important milestones in the study of contingency tables. From the seventies up to now many extensions, improvement of methods and generalisations have been presented, for instance, see Everitt (1977), Darroch, Lauritzen, and Speed (1980), Chambers and Welsh (1993), or Goodman (1996). However, challenges are still pendent for a straightforward solution, specially for the study of n-way

contingency tables.

From the compositional point of view, the contingency tables have been studied only recently, with the early precedent of Kenett (1983). In the workshop CoDaWork 2008 (Girona, Spain) Egozcue, Díaz-Barrero, and Pawlowsky-Glahn (2008) introduced a perturbation-decomposition model for tables of multinomial parameters, thus opening new possibilities of analysis. This contribution was followed by other compositional attempts and applications (Gallo 2015; Fačevicová and Hron 2013). The approach proposed in Egozcue, Pawlowsky-Glahn, Templ, and Hron (2015) is a kind of log-linear model but it has some differences with the standard ones. The main differences are the way in which marginals are found and the definition of interactions.

Here, the model based on the orthogonal decomposition of multinomial contingency tables is used to carry out a Bayesian estimation of both close independent multinomial parameters and the subsequent interactions. The present goal is to show that orthogonal decomposition of multinomial contingency tables can be addressed using Bayesian estimation techniques. The zero problem, typical in compositional data analysis, is here overcome by estimating the model parameters (probabilities) underlying the contingency table, which are considered compositional parameters. Zeros in the observations do no produce any problem as their likelihood is well defined. This is a traditional way of dealing with zeroes in generalized linear models as multinomial logistic regression models (Nelder 1974) or Bayesian estimation of multinomial probabilities (e.g. Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

In Section 2, the main features of the model based on orthogonal decomposition of contingency tables are recalled. Some definitions of the Bayesian framework are introduced in Section 3. Examples are presented in Section 4.

2. Orthogonal decomposition model

Two-way contingency tables coming from a multinomial sampling are considered. They are generated by a row-classification into I classes, and a column-classification made of J classes. The total number, N, of classified individuals is then distributed on the $I \times J$ cells of the contingency table (CT) according to the classification. The number of individuals pertaining to the ij-cell is denoted n_{ij} for $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J$. The whole contingency table containing these counts is denoted $\mathbf{N}$. The table $\mathbf{N}$, as an array of counting random variables, is assumed to be multinomial distributed and its corresponding probability parameters denoted by p_{ij} $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J$. When arranged in a table, these probability parameters are called probability table (PT). The sample space of a CT, like $\mathbf{N}$, is $I \times J$ times the non-negative integers restricted to add to N. They are not conceived as compositional data, even when the frequencies $\mathbf{N}/N$ are computed. In fact, they can contain zero-counts and only can correspond to fractions with N as denominator. Alternatively, the probability parameters $\mathbf{P}$ are considered compositional. This can be summarized as (a) $\mathbf{P}$ is in $\mathcal{S}^D$, $D = I \cdot J$; (b) perturbation and powering, denoted $\oplus$, $\odot$ respectively, are vector space operations, and the dimension of $\mathcal{S}^D$ is $D - 1$; (c) the centered log-ratio (clr) transformation is defined and inner product, norm and distances in $\mathcal{S}^D$ are the ordinary Euclidean inner product, norm and distance of the clr transformed PT's. As well-known for $\mathcal{S}^D$ (Pawlowsky-Glahn and Egozcue 2001), the simplex endowed with $\oplus$, $\odot$, and the Aitchison inner product is a $D-1$-dimensional Euclidean space. More explicitly, consider two PT's, $\mathbf{P}$ and $\mathbf{Q}$ and a real number α. The perturbation $\mathbf{W} = \mathbf{P} \oplus \mathbf{Q}$ is a PT with entries $w_{ij} = p_{ij}q_{ij} / \sum_{km} p_{km}q_{km}$, $k = 1, \ldots, I$ and $m = 1, \ldots, J$. The α-powering $\mathbf{W} = \alpha \odot \mathbf{P}$ is a PT with entries $w_{ij} = p_{ij}^{\alpha} / \sum_{km} p_{km}^{\alpha}$. The clr of a PT is an $(I \times J)$-array, $\mathbf{V} = \mathrm{clr}(\mathbf{P})$ which entries are

$$v_{ij} = \log(p_{ij}) - \frac{1}{D} \sum_{k=1}^{I} \sum_{m=1}^{J} \log p_{km} \ .$$

The inverse clr-transformation is $\mathbf{P} = \mathcal{C} \exp(\mathbf{V})$, where exp operates componentwise and $\mathcal{C}$ is

the closure operator. For any vector of D strictly positive real components,

$$\mathbf{z} = (z_1, z_2, \ldots, z_D) \in \mathbb{R}_+^D, \quad z_i > 0 \quad \text{for all } i = 1, 2, \ldots, D \ ,$$

the closure of $\mathbf{z}$ to $\kappa > 0$ is defined as

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^D z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^D z_i}, \ldots, \frac{\kappa \cdot z_D}{\sum_{i=1}^D z_i} \right] \ .$$

Denoting $\mathrm{clr}(\mathbf{P}) = \mathbf{V}$ and $\mathrm{clr}(\mathbf{Q}) = \mathbf{W}$, the Aitchison inner product, $\langle \cdot, \cdot \rangle_a$, and distance, $\mathrm{d}_a(\cdot, \cdot)$, of PT's is

$$\langle \mathbf{P}, \mathbf{Q} \rangle_a = \langle \mathbf{V}, \mathbf{W} \rangle = \sum_{i=1}^I \sum_{j=1}^J v_{ij} w_{ij} \quad , \quad \mathrm{d}_a^2(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^I \sum_{j=1}^J (v_{ij} - w_{ij})^2 \ ,$$

where $\langle \cdot, \cdot \rangle$ denotes the ordinary Euclidean inner product of arrays.

In these definitions, commonly used in compositional data analysis, there are, at least, two key points. The first one is the interpretability of the perturbation. Perturbation of PT's correspond to apply the Bayes formula to a PT, containing prior probabilities, using a likelihood arranged as a PT, up to the closure operation. The second point is that the subcompositional coherence (Pawlowsky-Glahn *et al.* 2015; Egozcue 2009), is guaranteed. In the case of PT's, subcompositional coherence assures that distance between two sub-tables have Aitchison distance smaller than or equal to the distance between the parent PT's.

The main result in Egozcue *et al.* (2008, 2015) is that, independent PTs constitute an $(I-1)(J-1)$-dimensional linear subspace of $\mathcal{S}^D$. This means that any PT can be projected orthogonally on this subspace. The consequence is that $\mathbf{P}$ is decomposed in a unique way as

$$\mathbf{P} = \mathbf{P}_{ind} \oplus \mathbf{P}_{int} \quad , \quad \mathbf{P}_{ind} \perp \mathbf{P}_{int} \ , \tag{1}$$

where $\mathbf{P}_{ind}$ is the projection of $\mathbf{P}$ on the independent subspace, and $\mathbf{P}_{int}$ is in the orthogonal complement. The PT $\mathbf{P}_{int}$ is called interaction PT. The independent PT is on its turn decomposed into two new PT's, called marginal PT's, which have equal rows and equal columns respectively. The independent PT is then decomposed as

$$\mathbf{P}_{ind} = (\mathbf{1}_I \mathbf{r}^\top) \oplus (\mathbf{c} \mathbf{1}_J^\top) \ , \tag{2}$$

where $\mathbf{r}$, $\mathbf{c}$ are compositions in $\mathcal{S}^J$ and $\mathcal{S}^I$ respectively, and they are treated as column vectors for matrix notation. The symbols $\mathbf{1}_I$ and $\mathbf{1}_J$ are column-vectors, with I and J components respectively, all of them equal to 1.

Equations 1 and 2 can be transformed by taking clr, which yields

$$\mathrm{clr}(\mathbf{P}) = \mathrm{clr}(\mathbf{P}_{ind}) + \mathrm{clr}(\mathbf{P}_{int}) = \mathbf{1}_I (\mathrm{clr}(\mathbf{r}))^\top + \mathrm{clr}(\mathbf{c}) \mathbf{1}_J^\top + \mathrm{clr}(\mathbf{P}_{int}) \ . \tag{3}$$

It should be remarked that $\mathrm{clr}(\mathbf{r})$ and $\mathrm{clr}(\mathbf{c})$ are clr transformations of compositions in $\mathcal{S}^J$ and $\mathcal{S}^I$ respectively and they are not PT's.

The marginal row and column, $\mathbf{r}$ and $\mathbf{c}$ respectively, are obtained from $\mathbf{P}$ as the closed geometric means by columns and rows of $\mathbf{P}$ respectively. This feature indicates that the nearest independent PT, in the sense of Aitchison geometry in $\mathcal{S}^D$, is not obtained from the traditional (arithmetic) marginals. This is an important difference from common analysis of contingency tables. As a consequence, $\mathrm{clr}(\mathbf{P}_{ind})$ has the property that its arithmetic and geometric marginals are equal up to a closure; and the geometric marginals of $\mathbf{P}_{int}$ are neutral in the simplex (i.e. all their elements are equal).

The decomposition in Equation 3 implicitly defines a log-linear model which is revealed after taking clr^{-1} in Equation 3. The log-linear model is then

$$\mathbf{P} = \mathcal{C} \exp[\mathrm{clr}(\mathbf{P}_{ind}) + \mathrm{clr}(\mathbf{P}_{int})] = \mathcal{C} \exp[\mathbf{1}_I (\mathrm{clr}(\mathbf{r}))^\top + \mathrm{clr}(\mathbf{c}) \mathbf{1}_J^\top + \mathrm{clr}(\mathbf{P}_{int})] \ , \tag{4}$$

where the parameters are the J-coefficients in $\text{clr}(\mathbf{r})$, the I coefficients in $\text{clr}(\mathbf{c})$ and the $D = I \cdot J$ coefficients in $\text{clr}(\mathbf{P}_{int})$. However, coefficients of any clr add to zero, and the number of free parameters is $(J-1)+(I-1)+(IJ-1) = IJ+I+J-3$. The number of clr-parameters in Equation 4 can be reduced to $IJ+I+J-3$ using ilr-coordinates, but this strategy is not used here as the clr-parameters can be interpreted directly.

In order to interpret the results when the log-linear model is fitted to a CT, some derived parameters may be useful. When the norm $\|\mathbf{P}_{int}\|_a$ is null, $\mathbf{P}_{int}$ is the neutral element in $\mathcal{S}^D$ and $\mathbf{P}$ is an independent PT. Therefore, $\|\mathbf{P}_{int}\|_a^2$ is an overall measure of dependence which was named simplicial deviance. When considered relative to the Aitchison square norm of $\mathbf{P}$, it can be called relative simplicial deviance. The corresponding definitions are

$$\Delta^2(\mathbf{P}) = \|\mathbf{P}_{int}\|_a^2 \quad , \quad R^2_\Delta(\mathbf{P}) = \frac{\|\mathbf{P}_{int}\|_a^2}{\|\mathbf{P}_{ind}\|_a^2 + \|\mathbf{P}_{int}\|_a^2} \ , \tag{5}$$

where $\|\mathbf{P}_{ind}\|_a^2 + \|\mathbf{P}_{int}\|_a^2 = \|\mathbf{P}\|_a^2$ due to the orthogonal decomposition (Equation 1). Remarkably, $\Delta^2(\mathbf{P})$ does not depend on the marginals of $\mathbf{P}$; such a property is not shared by $R^2_\Delta(\mathbf{P})$. However $R^2_\Delta(\mathbf{P})$ has clear interpretation based on the facts of $0 \le R^2_\Delta(\mathbf{P}) \le 1$, $R^2_\Delta(\mathbf{P}) = 0$ implies independence of $\mathbf{P}$, whereas $R^2_\Delta(\mathbf{P}) = 1$ indicates that the nearest independent PT to $\mathbf{P}$ is the neutral (uniform) PT, and it can be considered as a pure interaction PT.

In order to interpret the coefficients of $\mathbf{V} = \text{clr}(\mathbf{P}_{int})$ it should be taken into account that the simplicial deviance is decomposed

$$\Delta^2(\mathbf{P}) = \|\mathbf{P}_{int}\|_a^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} v_{ij}^2 \ , \tag{6}$$

so that each cell contributes to the simplicial deviance with v_{ij}^2 thus deserving the name of cell interaction. A way of presenting these cell interactions is computing their relative value to the simplicial deviance or expressing them as percent of contribution. However, the signs of v_{ij} are important as they indicate whether the probability in the cell p_{ij} is smaller than the predicted probability using $\mathbf{P}_{ind}$ (negative v_{ij}) or it is larger than this predicted probability (positive v_{ij}). It has been proposed to use an interaction array reporting in each cell the value $\text{sign}(v_{ij})(v_{ij}^2/\Delta^2(\mathbf{P}))$. Unfortunately, the values of v_{ij} cannot be interpreted separately as they add to zero. The analyst should look for large absolute values in the interaction array coupled by positive-negative interactions. Cells interactions are then interpreted jointly as the sources of interaction are frequently coupled.

3. Bayesian analysis

Assume that an $I \times J$ contingency table $\mathbf{N}$ has been observed as the result of a multinomial sampling. After adopting the log-linear model (Equation 4), the multinomial probabilities p_{ij} can be expressed as functions of the clr's of the geometric marginals $\text{clr}(\mathbf{r}) = \mathbf{z}^{(r)} = (z_1^{(r)}, z_2^{(r)}, \ldots, z_J^{(r)})$, $\text{clr}(\mathbf{c}) = \mathbf{z}^{(c)} = (z_1^{(c)}, z_2^{(c)}, \ldots, z_I^{(c)})$, and the entries of $\mathbf{V} = \text{clr}(\mathbf{P}_{int})$ denoted v_{ij}. Hence, the likelihood of these parameters, given the observation has the form

$$L(\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V} \mid \mathbf{N}) = K \cdot \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{n_{ij}} \ ,$$

where all p_{ij} are functions of $\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V}$ and K the normalizing constant corresponding to the multinomial density. In order to simplify the estimation procedure, a Dirichlet distribution (e.g. Aitchison 1986) can be chosen as initial joint distribution of the p_{ij}. If the chosen parameters of the Dirichlet distribution are $a_{ij} > 0$, the final or posterior distribution of the parameters is again a Dirichlet distribution with parameters $p_{ij} + a_{ij}$ and, therefore, the

posterior distribution is

$$f(\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V} \mid \mathbf{N}) = \frac{\Gamma\left(\sum_k \sum_m a_{km}\right)}{\prod_k \prod_m \Gamma(a_{km})} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{n_{ij}+a_{ij}-1} \quad , \quad \sum_k \sum_m p_{ij} = 1 \ , \tag{7}$$

The goals of the Bayesian procedure are, at least, three: (a) estimation of posterior distribution of parameters $\mathbf{z}^{(r)}$, $\mathbf{z}^{(c)}$, $\mathbf{V}$ and their marginal distributions; (b) checking the hypothesis of independence of the observed CT; (c) study the distribution of the cell interactions v_{ij} and checking whether they can be considered null or not. These three tasks are hardly carried out using the explicit distribution (Equation 7). A way out consists of drawing independent realisations from Equation 7, and then, studying the simulated sample of parameters thus accomplishing goal (a).

Checking independence of the observed CT is performed through a predictive p-value (Bayarri and Berger 2000; Meng 1994) as proposed in goal (a). Assume that for each possible set of posterior parameters, $\mathbf{z}_0^{(r)}$, $\mathbf{z}_0^{(c)}$, $\mathbf{V}_0$, a likelihood ratio test is carried out on the hypothesis

$$H_0 : \ \mathbf{z}^{(r)} = \mathbf{z}_0^{(r)} \ , \ \mathbf{z}^{(c)} = \mathbf{z}_0^{(c)} \ , \ \mathbf{V} = \mathbf{0} \ , \tag{8}$$

using the statistic

$$\Lambda = -2 \log \left(\frac{L(\mathbf{z}_0^{(r)}, \mathbf{z}_0^{(c)}, \mathbf{V} = \mathbf{0} \mid \mathbf{N})}{L(\hat{\mathbf{z}}^{(r)}, \hat{\mathbf{z}}^{(c)}, \hat{\mathbf{V}} \mid \mathbf{N})} \right) \ , \tag{9}$$

where $\hat{\mathbf{z}}^{(r)}$, $\hat{\mathbf{z}}^{(c)}$, $\hat{\mathbf{V}}$ denote the maximum likelihood estimators based on the sample CT. Asymptotically with N, the statistic Λ has distribution χ^2 with degrees of freedom $IJ + I + J - 3$. This corresponds to the number of estimated parameters, compared with no free parameter in H_0. For each set of values $\mathbf{z}_0^{(r)}$, $\mathbf{z}_0^{(c)}$, $\mathbf{V}_0$, one p-value α_{p0} is obtained. The p-value α_{p0}, as a function of the observed CT, has uniform distribution under asymptotic conditions (Robins, van der Vaart, and Ventura 2000). A predictive p-value, α, with asymptotic uniform distribution, is obtained using

$$\alpha = \Phi \left(\frac{1}{m} \sum_{k=1}^{m} \Phi^{-1} \left(\alpha_{p0}^{(k)} \right) \right) \ , \tag{10}$$

where the sum goes through the set of p-values corresponding to the m-simulated sample of parameters $\mathbf{z}_0^{(r)}$, $\mathbf{z}_0^{(c)}$; and Φ denotes the standard normal distribution function (Ortego 2015). Small values of α suggest rejection of the independence H_0.

The assessment of the hypothesis that a single cell interaction v_{ij} is null is performed using Bayesian discrepancy p-values (Gelman, Meng, and Stern 1996), that is, computing the posterior probability of $v_{ij} \leq 0$ across the simulated sample. When this p-value is small (near to zero), or large (near to 1), rejection of v_{ij} is suggested. This accomplishes goal (c).

4. A simple example

4.1. Marks in a subject

The marks obtained by $N = 104$ students in a exam of a college-level statistics subject are considered. Theoretical and practical (mostly problems) questions in the exams are marked separately. In this context, we want to know if the performance in theoretical questions can be considered independent from the performance in practical questions.

The results of the exam may be classified into four groups: A, B, C, D, corresponding to the numeric interval of Spanish marks over 10 points. The results corresponding to this group of students have been organized in a two-way table T (Table 1). We assume that these marks

Table 1: Two-way contingency table containing the marks of the May examination of 104 students. Mark of the theoretical part of the exam (rows) vs. mark of the practical part (columns). The equivalence between marks A, B, C, D and traditional Spanish scores is indicated in the first column.

mark (theory)\ mark (prob)	A [8.5,10]	B [7,8.5)	C [5,7)	D [0,5)
A [8.5,10]	1	0	4	4
B [7,8.5)	2	4	6	13
C [5,7)	0	3	11	25
D [0,5)	1	1	5	24

have been observed as a result of a multinomial sampling with probabilities p_{ij}. A Bayesian framewok is chosen for the estimation of the table parameters p_{ij}. For simplicity, a joint Dirichlet distribution has been assumed for these probabilities.

A Dirichlet prior has been set for the multinomial probabilities. Then, the posterior distribution of these parameters corresponds again to a Dirichlet distribution (Equation 7). A large sample of the posterior distribution has been drawn. This sample is used to describe the uncertainty of parameter estimates and other quantities of interest derived from them. For this data set, a sample of the posterior of length 10,000 has been obtained (e.g. Table 2).

Table 2: Example of a sample PT drawn from the posterior Dirichlet distribution

t\p	A	B	C	D
A	0.01	0.00	0.06	0.08
B	0.01	0.03	0.04	0.17
C	0.00	0.04	0.10	0.24
D	0.02	0.02	0.02	0.16

The tables sampled PT's from the posterior Dirichlet distribution should be properly treated, due to their compositional character. The clr coordinates of the cells for each table have been computed (e.g. Table 3). The row and column geometric marginals of the clr coordinates have also been obtained for each of the tables of the posterior sample. Also, each of these tables has been decomposed in its independent (e.g. Table 4) and interaction table (e.g. Table 5) following Equation 1. That is, a sample of independent and interaction tables has been obtained from the sample of posterior tables. This allows to describe the uncertainty of quantities of interest derived from them, such as deviance, $\Delta^2(\mathbf{P})$, relative deviance, $R^2_\Delta(\mathbf{P})$, among others.

Table 3: Example of clr-coordinates of a sample PT drawn from the posterior Dirichlet distribution. Row and column geometric marginals.

t\p	A	B	C	D	rmarg
A	-1.36	-5.39	1.02	1.30	-1.105
B	-0.57	0.35	0.68	2.06	0.631
C	-4.19	0.52	1.48	2.38	0.048
D	-0.26	0.05	-0.08	2.00	0.426
cmarg	-1.594	-1.117	0.776	1.936	

The departure from independence for the two sets of marks of interest may be measured observing the simplicial deviance (squared Aitchison norm) of the interaction component of drawn posterior tables (Equation 5). Figure 1 shows the histogram of simplicial deviances corresponding to the obtained sample of interaction tables. As the deviance is a measure of dependence, $0 \leq \Delta^2(\mathbf{P}) < +\infty$, a visual comparison with the zero value (red line) is included. For the marks in the example, although the median value (blue line) is low, the amount of variability in the deviance values points to lack of independence between the theoretical and

Table 4: Example of clr-coordinates of the independent component of a sample PT drawn from the posterior Dirichlet distribution

t\p	A	B	C	D
A	-2.70	-2.22	-0.33	0.83
B	-0.96	-0.49	1.41	2.57
C	-1.55	-1.07	0.82	1.98
D	-1.17	-0.69	1.20	2.36

Table 5: Example of clr-coordinates of the interaction component of a sample PT drawn from the posterior Dirichlet distribution

t\p	A	B	C	D
A	1.34	-3.17	1.35	0.47
B	0.39	0.83	-0.72	-0.50
C	-2.64	1.59	0.65	0.40
D	0.90	0.74	-1.28	-0.36

practical marks.

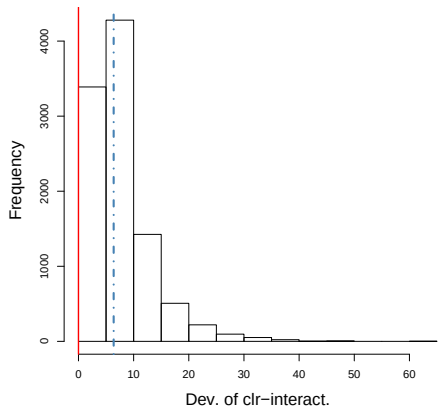

Figure 1: Histogram of posterior simplicial deviance (square norm of the clr-interaction) for final marks. Red line (solid): null value. Blue line (dashed): median.

The relative simplicial deviance $R^2_{\triangle}(\mathbf{P})$ may seem easier to interpret than the deviance, as $0 \leq R^2_{\triangle}(\mathbf{P}) \leq 1$, but this interpretation should be taken with caution as this parameter is not marginal invariant. Figure 2 shows the histogram corresponding to the relative simplicial deviance of the posterior sample. A zero-line is also included for a visual comparison. In this case, the majority of the relative deviance values are around 0.3, being near to its median value, reassuring the interpretation of lack of independence.

The simplicial deviance is an overall measure of dependence, but often more detail is needed. The cell values of the interaction table (e.g. Table 6) provide this detail, but the direct interpretation of the values may be confusing due to its compositional character. In order to obtain a detailed description of interaction using the appropriate scale, the clr-coordinates of interaction PT in the posterior sample have been computed. Figure 3 shows the histograms of the cell interactions as a summary of the obtained results. For visual comparison, a zero line has been added to each histogram. Visually, a zero line near the median of the histogram indicates no interaction added by that cell (e.g. histogram corresponding to cell 4). If the zero line is *far* from the center of the histogram (e.g. histogram corresponding to cell 1), that cell may be adding interaction to the deviance (Equation 6), and should be studied.

However, for an easier understanding of the importance of each cell, the interaction array of

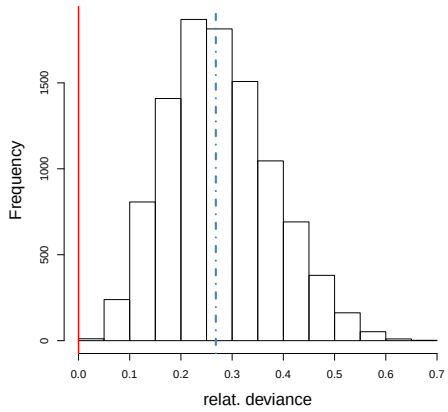

Figure 2: Histogram of relative simplicial deviance (square norm of clr-interac / total) corresponding to the posterior sample. Final marks. Red line (solid): null value. Blue line (dashed): median.

the cells has also been computed (e.g. Table 6), measuring the percentage of interaction added to the deviance by each cell, and including the sign of this interaction. The histogram of the signed interaction array of the posterior sample is shown in Figure 4. Visually, cells with interaction arrays clearly different from zero should be studied, as they are the influential ones. It seems that the most influential cells for the departure of independence are cells number 1, namely 'A in theory' vs 'A in practical' marks and number 3, 'A in theory' vs 'C in practical' marks, with more or less the same weight and opposite signs (see Figure 4, first row). The positive sign of the interaction array for cell number 1 means that the predicted probability for the cell is larger than the predicted by the independent table, while the predicted probabilities for cell 3 (negative sign) are lower than the probabilities predicted by the independent table. Other cells, as cell 5, are also influential, but with a lower weight. The hypothesis of null interaction has also been assessed by means of a Bayesian p-value based on a discrepancy (posterior probability of $v_{ij} \leq 0$ across the sample) (Table 7). If the zero value is central in the sample, i.e. the proportion of $v_{ij} \leq 0$ is near 0.5, the hypothesis is not rejected. Otherwise, small or large proportions, lead to the rejection of the null interaction hypothesis. For instance, for cell number 3, the Bayesian p-value is 0.956, and therefore the null hypothesis is clearly rejected. For cell number 1, the p-value is 0.101 and, although the value is low, the decision of rejection of null cell interaction is not so straightforward.

Table 6: Example of interaction array from the sample

t\p	A	B	C	D
A	6.27	0.54	-24.25	2.84
B	-34.89	2.42	8.79	1.92
C	6.37	-1.82	1.49	-5.74
D	0.77	-0.89	0.55	-0.46

Table 7: Assessment of null interaction hypothesis. Bayesian p-value based on discrepancy (posterior probability of $v_{ij} \leq 0$ across the sample)

t\p	A	B	C	D
A	0.101	0.297	0.956	0.455
B	0.773	0.131	0.223	0.697
C	0.261	0.873	0.175	0.746
D	0.880	0.914	0.115	0.052

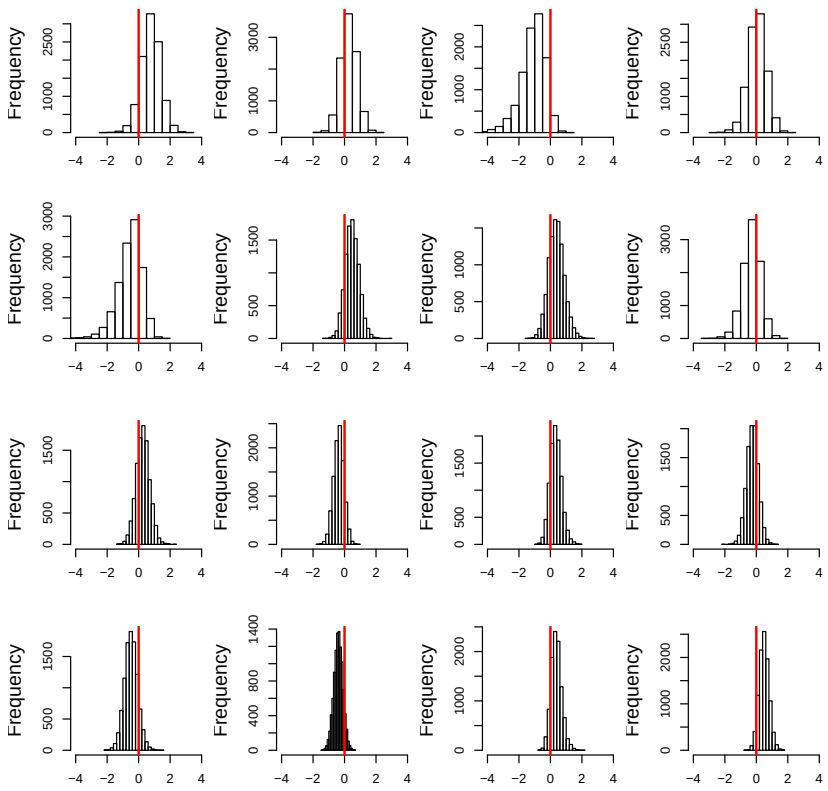

Figure 3: Histograms of clr-cell interactions for the posterior sample. Red lines (solid): null interaction

4.2. An independence test

Simplicial deviance, relative deviance and the interaction array are useful quantities to study dependence in a contingency table. However, it is usual to discuss independence in contingency tables by means of a test (e.g. Equation 8). In our example, are the marks for theory and practice in the exam independent? In the established theoretical context, that can be rephrased as, does the contingency table of marks, T, belong to the subspace of independent tables?

$$H_0 : T = \mathbf{P}_{ind} \in \mathcal{S}^D_{ind} \quad ; \quad H_1 : T = \mathbf{P} \notin \mathcal{S}^D_{ind}$$

The selected likelihood ratio test statistic (Equation 9), is based on the sample of estimates of the independent component $\mathbf{P}_{ind}$, $\widehat{\mathbf{P}}_{ind}$. For each table of the sample of the posterior distribution, its decomposition into independent and interaction component has been obtained in section 4.1. The proposed test statistic and its corresponding predictive p-value have been computed for each of these decompositions. This sample of p-values can be used to measure the uncertainty of the decision of the independence test. Figure 5 shows the histogram of these predictive p-values for the posterior sample of tables. It can be observed that there is variability in the sample of p-values, with a majority of small values, leading to the rejection of the independence hypothesis. However, the lack of uniformity of p-values and their relative scale are problematic for their interpretation (Robins *et al.* 2000). Therefore, the predictive p-values of the sample have been suitably transformed and combined, in order to obtain a summary p-value, α, with asymptotic uniform distribution. In this case, α is nearly 0, and the independence hypothesis has been rejected, as already pointed out by the deviance values. That is, it cannot be considered that the theory and practical marks of this exam are independent.

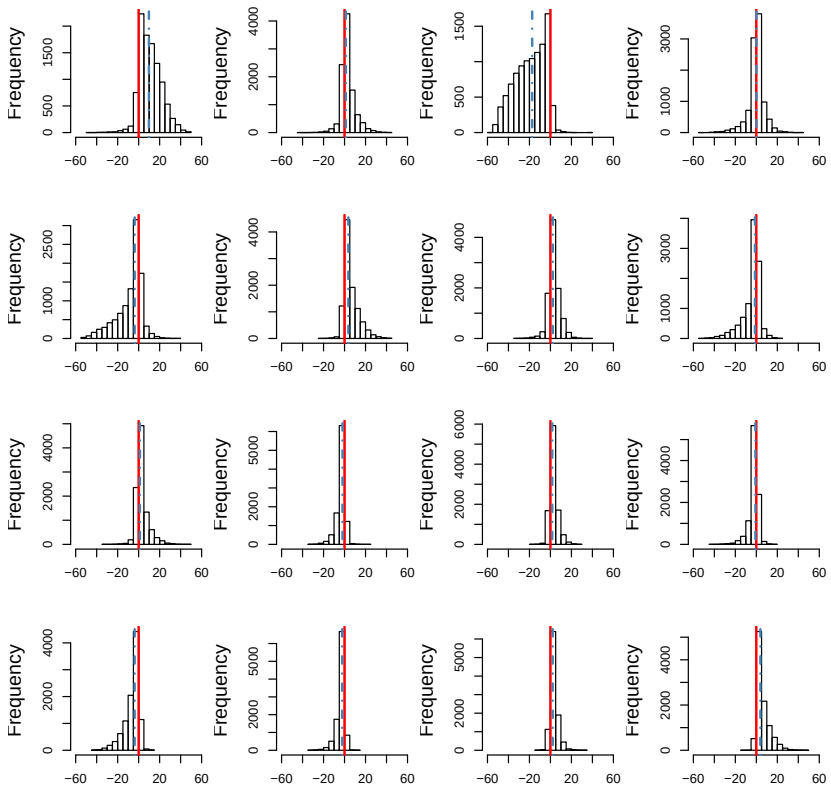

Figure 4: Histogram of the interaction array for each cell. Red line (solid): null value. Blue line (dashed): median

5. Conclusions

Contingency tables have been broadly studied, although only recently they have been treated from the compositional point of view. The orthogonal decomposition of multinomial contingency tables has been presented. Also, a Bayesian framework for the estimation of the parameters of contingency tables has been introduced as a novelty in the compositional treatment of these tables.

A two-way contingency table containing marks from an exam of a college-level statistics course has been studied as an example. Results show that theory and practical exam marks cannot be considered as independent as suggested by the table decomposition and their summary statistics. The Bayesian point of view allows considering uncertainty of estimators and summary statistics. Moreover, the Bayesian approach deals with small or null counts in the original table very efficiently. The multinomial probabilities of the table are assumed compositional. Contrarily, counts in the original contingency table are not reduced to frequencies thus avoiding zero replacements or imputations. This latter fact makes Bayesian estimation very useful in the context of compositional analysis of probability parameters.

Acknowledgements

This research has been partially funded by the Ministerio de Economía y Competividad under project "CODA-RETOS" (Ref. MTM2015-65016-C2-2-R); and by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya under the project "Compositional and Spatial Analysis" (COSDA) (Ref: 2014SGR551;2014-2016).

References

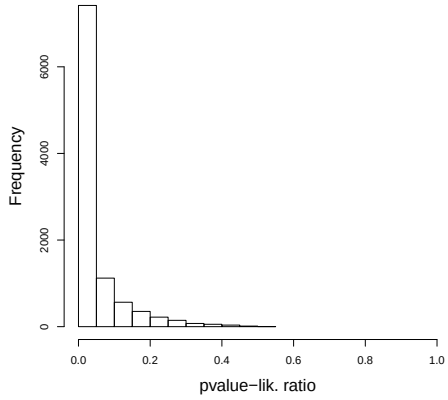

Figure 5: Predictive p-value of the multinomial likelihood ratio independence test. Final marks

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). ISBN 0-412-28060-4. 416 p.

Bayarri MJ, Berger JO (2000). "P-values for composite null models." *Journal of the American Statistical Association*, **95**, 1127–1142.

Chambers RL, Welsh AH (1993). "Log-linear Models for Survey Data with Non-ignorable Non-response." *J. R. Statist. Soc. B*, **55**(1), 157–170.

Darroch JN, Lauritzen SL, Speed TP (1980). "Markov fields and Log-linear interaction models for contingency tables." *The Annals of Statistics*, **8**(3), 522–539.

Egozcue JJ (2009). "Reply to "On the Harker variation diagrams; ..." by J. A. Cortés." *Mathematical Geosciences*, **41**(7), 829–834.

Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2008). "Compositional analysis of bivariate discrete probabilities." In J Daunis-i Estadella, JA Martín-Fernández (eds.), *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*, pp. 1–11. University of Girona, Girona. ISBN 978-84-8458-272-4, URL http://hdl.handle.net/10256/717.

Egozcue JJ, Pawlowsky-Glahn V, Templ M, Hron K (2015). "Independence in contingency tables using simplicial geometry." *Communications in Statistics- Theory and methods*, **44**(18), 3978–3996. doi:10.1080/03610926.2013.824980.

Everitt BS (1977). *The Analysis of Contingency Tables*. John Wiley & Sons, Inc. New York, New York, USA. ISBN 0-470-71135-3.

Fačevicová K, Hron K (2013). "Statistical analysis of compositional 2 X 2 tables." In K Hron, P Filzmoser, M Templ (eds.), *Proceedings of the 5th International Workshop on Compositional Data Analysis*. TU Wien, Wien. ISBN 978-3-200-03103-6.

Gallo M (2015). "Tucker3 Model for Compositional Data." *Communications in Statistics - Theory and Methods*, **44**(21), 4441–4453. doi:10.1080/03610926.2013.798664.

Gelman A, Meng XL, Stern H (1996). "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)." *Statistica Sinica*, **6**, 733–807.

Goodman LA (1996). "A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fisher, and Also Some Methods of Correspondence Analysis and Association Analysis." *Journal of the American Statistical Association*, **91**(433), 408–428.

Kenett RS (1983). "On an Exploratory Analysis of Contingency Tables." *The Statistician*, **32**(3), 395–403.

McCullagh P, Nelder JA (1983). *Generalized Linear Models*. Chapman and Hall, London, UK. 522 p.

Meng XL (1994). "Posterior predictive p-values." *Annals of Statistics*, **22**, 1142–1160.

Nelder JA (1974). "Log linear models for contingency tables: a generalization of classical least squares." *Appl. Statist.*, **23**, 323–329.

Nelder JA, Wedderburn RWM (1972). "Generalized linear models." *Journal of the Royal Statistical Society, series A*, **135**, 370–384.

Ortego MI (2015). *Estimación Bayesiana de cópulas extremales en procesos de Poisson*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Pawlowsky-Glahn V, Egozcue JJ (2001). "Geometric approach to statistical analysis on the simplex." *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15**(5), 384–398.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and analysis of compositional data*. Statistics in practice. John Wiley & Sons, Chichester UK. ISBN 9781118443064. 272 pp.

Robins JM, van der Vaart A, Ventura V (2000). "Asymptotic Distribution of p-Values in Composite Null Models." *Journal of the American Statistical Association*, **95**(452), 1143–1156.

Yule GU (1912). "On the Methods of Measuring Association Between Two Attributes." *Journal of the Royal Statistical Society*, **75**, 579–642.

Affiliation:

Maria Isabel Ortego
Civil and Environmental Engineering Department.
Campus Nord UPC. Edifici C2
Universitat Politècnica de Catalunya
08034 Barcelona, Spain
E-mail: ma.isabel.ortego@upc.edu

Juan José Egozcue
Civil and Environmental Engineering Department.
Campus Nord UPC. Edifici C2
E-mail: juan.jose.egozcue@upc.edu

Austrian Journal of Statistics
September 2016, Volume 45, 57–71.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.142

The Mathematics of Compositional Analysis

Carles Barceló-Vidal
University of Girona, Spain

Josep-Antoni Martín-Fernández
University of Girona, Spain

Abstract

The term *compositional data analysis* is historically associated to the approach based on the logratio transformations introduced in the eighties. Two main principles of this methodology are scale invariance and subcompositional coherence. New developments and concepts emerged in the last decade revealed the need to clarify the concepts of compositions, compositional sample space and subcomposition. In this work the mathematics of compositional analysis based on equivalence relation is presented. A logarithmic isomorphism between quotient spaces induces a metric space structure for compositions. The *logratio compositional analysis* is the statistical analysis of compositions based on this structure, consisting of analysing logratio coordinates.

Keywords: composition, compositional analysis, equivalence class, logratio, quotient space, simplex.

1. Introduction

The term *compositional data* (CoDa) was first introduced by Aitchison (1982) and later developed in Aitchison (1986). In these publications CoDa is identified with vectors of strictly positive components whose sum is always equal to one; that is, vectors of the unit simplex

$$\mathcal{S}^D = \{(w_1, \ldots, w_D)' : w_1 > 0, \ldots, w_D > 0; w_1 + \ldots + w_D = 1\}.$$

The term *compositional data analysis* (CoDA) has been implicitly associated with the methodology proposed by Aitchison (1986), which is based on applying the logratio transformations to the CoDa and describing, analysing and modelling them statistically from the logratios of their components. The main aim of this methodology is to free the CoDa from the constraints of the constant sum in order to be able to use the standard distributions in the real space to model the CoDa, e.g., the multivariate normal distribution. This strategy has two fundamental concepts in the so-called *principles of CoDA* (Aitchison 1986), namely, 'scale invariance' and 'subcompositional coherence'. From Aitchison (1986), "scale invariance merely reinforces the intuitive idea that a composition provides information only about relative values not about absolute values and therefore ratios of componentes are the relevant entities to study"; and "subcompositional coherence demands that two scientist, one using full composition and the other using subcompositions of these full compositions, should make the same inference about relations within the common parts". Later it was seen that the methodology initiated by Aitchison is more than a simple transformation of the CoDa, because it is in

fact a way to provide the simplex with a structure of Euclidean space. The interested reader can refer to Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras (2011) for further information.

The identification of the term CoDA with the methodology based on the logratio transformations developed by Aitchison has meant that other possible methods for analysing CoDa have made little impact. Watson and Philip (1989), Wang, Liu, Mok, Fu, and Tse (2007) or Scealy and Welsh (2011), for example, prefer to apply the techniques characteristic of directional data, given that they take the positive orthant of the unit hypersphere centred at the origin as the sample space of the CoDa. At that time this alternative method for analysing CoDa was cause for intense epistolary exchanges between D. F. Watson (and G. M. Philip) and J. Aitchison (see Aitchison 1990; Watson 1990; Aitchison 1991; Watson 1991). Recently, Scealy and Welsh (2014), have returned to the controversial questioning of the principles of CoDa, which they consider to be made to specifically exclude any methodology other than the one developed by Aitchison. As Scealy and Welsh (2014) recognise, the crux of the controversy lies in the definitions of composition and sample space in CoDA, both of which were introduced by Aitchison (1986) and based on constant sum vectors. The lack of clarity in the presentation of the properties scale invariance and subcompositional coherence is also a matter for discussion.

The main aim of this paper is to provide a precise and unequivocal definition of the concepts of composition, CoDa sample space and subcomposition, on which compositional analysis (CoAn) is based. Contrary to Scealy and Welsh (2014), we turn to mathematics to introduce these concepts with maximum precision. Thus, in Section 2 we define the quotient space of the compositions and we provide a precise definition of the concept of subcomposition. We also define what we understand by CoAn, distinguishing it from the traditional concept CoDA. In Section 3 we show how the logarithmic and exponential functions allow us to structure the sample space as a Euclidean space and to operate with the logratio coordinates of the data as if we were doing so in the real space. In the last section we compile the advantages and limitations of CoAn based on logratio coordinates and of the analysis based on transformations that take the positive orthant of the unit hypersphere as the sample space. Finally, we present the main conclusions.

2. The sample space in a compositional analysis

2.1. A composition is an equivalence class

We assume that our data and observations materialise in vectors $\mathbf{w} = (w_1, \ldots, w_D)'$ with strictly positive components, that is vectors from real space $\mathbb{R}^D_+$, the positive orthant of $\mathbb{R}^D$. Note that we are eluding to the case of zero values in the data. We consider the zero as a special value that deserves a particular analysis according to its nature (Palarea-Albaladejo and Martín-Fernández 2015); that is, the reason why a zero value is present in a CoDa set is informative and determines the approach to be applied. The interested reader is referred to Martín-Fernández, Palarea-Albaladejo, and Olea (2011) for further information. In the discussion we outline some of the approaches and discuss some kinds of zero.

Sometimes the observational vectors $\mathbf{w}$ are constant sum vectors. Typical examples are the data from time-use surveys where the sum equals to 24 in hours, 1440 in minutes or 100 in percentages. This case of CoDa is known as 'closed data'. In other situations, the components of the observational vectors are themselves meaningful, that is, they represent absolute magnitudes. However, in spite of that, we can decide to take only the relative information into account for our analysis. For example, in the analysis of household expenditure on D commodity groups, we can decide to analyze the distribution of the expenditure regardless of the total. In both scenarios we are implicitly assuming that the vectors $\mathbf{w}$ and $k\mathbf{w}$, with $k \in \mathbb{R}^+$, are providing us with the same compositional information, that is, the information given

by the ratios between the components. For example, the vectors $(0.3, 0.5, 0.2)$, $(30, 50, 20)$, $(7.2, 12, 4.8)$, and $(3/2, 5/2, 1)$ provide the same compositional information. In both cases we are assuming that our data are CoDa and our analysis will be a CoAn. Moreover, from a strictly mathematical point of view this implies that in a CoAn the sample space is not $\mathbb{R}_+^D$.

Definition 2.1. Two D-observational vectors $\mathbf{w}$ and $\mathbf{w}^*$ are *compositionally equivalent*, written $\mathbf{w} \sim \mathbf{w}^*$, if there is a positive constant k such that $\mathbf{w} = k\mathbf{w}^*$. This equivalence relation on $\mathbb{R}_+^D$ splits the space into equivalence classes, called *D-compositions* or, simply, *compositions*. The composition generated by an observational vector $\mathbf{w}$, i.e. the equivalence class of $\mathbf{w}$, is symbolized by $\underline{\mathbf{w}}$:

$$\underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\} \ .$$

Following Aitchison (1986), it is clear that a D-part composition can be geometrically interpreted as a ray from the origin in the positive orthant of $\mathbb{R}^D$ (Figure 1). Therefore, from a strictly mathematical point of view, the term CoAn is the equivalent of assuming that the sample space is the set of all D-compositions.

Definition 2.2. The set of all D-compositions, that is, the quotient space $\mathbb{R}_+^D/\sim$ is called the *D-compositional space* or, in brief, *compositional space*, and is symbolized by $\mathcal{C}^D$. We symbolize by ccl (from compositional class) the mapping from $\mathbb{R}_+^D$ to $\mathcal{C}^D$ which assigns each D-observational vector $\mathbf{w}$ to the composition $\underline{\mathbf{w}}$, i.e.,

$$\begin{array}{rccc} \text{ccl}: & \mathbb{R}_+^D & \longrightarrow & \mathcal{C}^D \\ & \mathbf{w} & \longmapsto & \underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\} \ . \end{array} \tag{1}$$

Property 2.1. Two D-observational vectors $\mathbf{w} = (w_1, \ldots, w_D)'$ and $\mathbf{w}^* = (w_1^*, \ldots, w_D^*)'$ are compositionally equivalent when the information provided by their ratios is the same, that is

$$\frac{w_i}{w_j} = \frac{w_i^*}{w_j^*} \quad \text{for each } i, j = 1, \ldots, D \ .$$

Any D-composition $\underline{\mathbf{w}}$ is completely determined by its ratios w_i/w_j of their components. Therefore, in a CoAn the relevant information provided by the observational vector $\mathbf{w}$ is found not in its components w_i, but rather in its ratios w_i/w_j. This is what we mean when we say that a composition only contains 'relative information' about its components. Note that all the observational vectors in the same ray (Figure 1) have the same ratios, providing the same relative information. That is, any point in the ray can be selected as representative of the equivalence class, and any statistical analysis has to provide the same information regardless of the representative selected. Importantly, if one applies a statistical method that does not take into account this essential attribute of the compositions, the application of different criteria to select the representatives will give different results and, likely, one will extract different conclusions.

To conclude, when we decide to do a CoAn we are assuming that the sample space of our data is the compositional space $\mathcal{C}^D$, which means in fact an acceptance of the 'scale invariance' principle.

2.2. Representatives of compositions

Any composition $\underline{\mathbf{w}}$ is determined by any observational vector $\mathbf{w}$ that belongs to the equivalence class. Thus, many different criteria can be used to select a representative of a composition. Each criterion gives rise to a different reference frame where projecting the compositions of $\mathcal{C}^D$. Here we present the most commonly-used criteria that facilitate the interpretation and have relevant mathematical properties.

Definition 2.3. The *linear criterion* selects the unit-sum vector $\mathbf{w}/\sum_{j=1}^{D} w_j$ to represent the composition $\underline{\mathbf{w}}$. We symbolize by $\mathrm{r_l}$ the mapping from $\mathcal{C}^D$ to the subset $\mathcal{S}^D$ of $\mathbb{R}_+^D$, that is

$$\begin{array}{rccc} \mathrm{r_l}: & \mathcal{C}^D & \longrightarrow & \mathcal{S}^D \subset \mathbb{R}_+^D \\ & \underline{\mathbf{w}} & \longmapsto & \mathbf{w}/\sum_{j=1}^{D} w_j \ , \end{array} \tag{2}$$

where $\mathcal{S}^D$ is the well-known *unit simplex*, historically considered as the sample space of CoDa.

The mapping $\mathrm{r_l}$ corresponds to the *constraining operator* or *closure operator* $\mathcal{C}$ introduced by Aitchison (1986). Geometrically, $\mathrm{r_l}(\underline{\mathbf{w}})$ is the intersection of the ray going from the origin through $\mathbf{w}$ and the hyperplane of $\mathbb{R}^D$ defined by the equation $w_1 + \ldots + w_D = 1$ (Figure 1). This criterion can be generalized to representatives with a sum equal to 100 or any other positive value.

Definition 2.4. We symbolize by $\mathrm{r_s}$ the mapping from $\mathcal{C}^D$ to the subset Sph_+^D of $\mathbb{R}_+^D$ which assigns to composition $\underline{\mathbf{w}}$ the intersection of the ray going from the origin through $\mathbf{w}$ and the unit hypersphere of $\mathbb{R}^D$ centred in the origin, i.e.,

$$\begin{array}{rccc} \mathrm{r_s}: & \mathcal{C}^D & \longrightarrow & \mathrm{Sph}_+^D \subset \mathbb{R}_+^D \\ & \underline{\mathbf{w}} & \longmapsto & \mathbf{w}/\|\mathbf{w}\| \ , \end{array} \tag{3}$$

where Sph_+^D is the strictly positive orthant of the unit hypersphere of $\mathbb{R}^D$ centred in the origin. We call this selection criterion the *spherical criterion* (Figure 1) because the representatives are unit-norm vectors using the classical Euclidean norm. This selection criterion was proposed by Watson and Philip (1989).

Definition 2.5. The *hyperbolic criterion*, $\mathrm{r_h}$, assigns to composition $\underline{\mathbf{w}}$ the intersection of the ray going from the origin through $\mathbf{w}$ and the hyperbolic surface Hip_+^D in $\mathbb{R}_+^D$ implicitly defined by the equation $\prod_{i=1}^{D} w_i = 1$:

$$\begin{array}{rccc} \mathrm{r_h}: & \mathcal{C}^D & \longrightarrow & \mathrm{Hip}_+^D \subset \mathbb{R}_+^D \\ & \underline{\mathbf{w}} & \longmapsto & \mathbf{w}/g(\mathbf{w}) \ , \end{array} \tag{4}$$

where $g(\mathbf{w}) = (\prod_{j=1}^{D} w_j)^{1/D}$ is the geometric mean of the components of vector $\mathbf{w}$ (Fig. 1).

Note that the function composition $\log \circ \mathrm{r_h}$ is equivalent to the *centred logratio* transformation (clr) introduced by Aitchison (1986): $\mathrm{clr}(\mathbf{w}) = \log(\mathbf{w}/g(\mathbf{w}))$.

The mappings $\mathrm{r_l}$, $\mathrm{r_s}$ and $\mathrm{r_h}$ can also be viewed as scale-invariant functions from $\mathbb{R}_+^D$ to $\mathcal{C}^D$. Recall that a function $f(\cdot)$ from $\mathbb{R}_+^D$ is said to be 'scale invariant' if for any positive constant k and for any observational vector $\mathbf{w}$, the function verifies $f(k\mathbf{w}) = f(\mathbf{w})$.

These criteria to select a representative of a composition can be extended to any surface defined in $\mathbb{R}_+^D$ using a bijective function. Indeed, making each composition correlate with the intersection of the corresponding ray with the surface is sufficient.

2.3. Subcompositions

In a CoAn, attention is usually focused on a determinate subset of the components of our observations of $\mathbb{R}_+^D$. For example, in time-use surveys we might only be interested in those activities that are different from the sleeping hours. If the analysis to be carried out on the components selected from our observations must also be compositional, then the sample space also needs to interpret it as a quotient space. This brings us to the need to introduce the concept of *subcomposition*.

Definition 2.6. Given a composition $\underline{\mathbf{w}} \in \mathcal{C}^D$, any composition obtained from the selection of two or more components of the D-observational vector $\mathbf{w}$ is termed a *subcomposition* of $\underline{\mathbf{w}}$. More precisely, let s be the number of selected components, with $2 \leq s < D$, and $i_1 < \ldots < i_s$ the sub indexes of these components (we implicity assume that the sub indexes

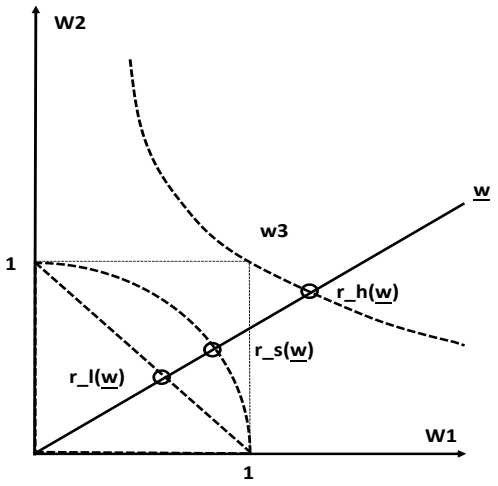

Figure 1: Linear, spherical and hyperbolic selection criteria (case $D = 2$).

of the D-observational vectors are $1, \ldots, D$). Let $\mathbf{S}$ be the $s \times D$ matrix with the ones in the positions $(1, i_1), \ldots, (s, i_s)$ of the matrix and zeros in the remaining positions. Making a subcomposition can be viewed as the transformation sub_S from $\mathcal{C}^D$ to $\mathcal{C}^s$ given by

$$\begin{array}{rccc} \text{sub}_S: & \mathcal{C}^D & \to & \mathcal{C}^s \\ & \underline{\mathbf{w}} & \to & \underline{\mathbf{Sw}}\,. \end{array} \tag{5}$$

The symbol $\mathbf{w}_S$ indicates the observational subvector $\mathbf{Sw} = (w_{i_1}, \ldots, w_{i_s})'$, and $\underline{\mathbf{w}_S}$ represents the final subcomposition which belongs to the compositional space $\mathcal{C}^s$. The transformation sub_S is compatible with the equivalence relation $\sim$, that is, equivalent observational vectors are transformed into equivalent subvectors. Importantly, the selected components $(w_{i_1}, \ldots, w_{i_s})'$ provide the same relative information regardless they belong to $\underline{\mathbf{w}}$ or they form the subcomposition $\underline{\mathbf{w}_S}$. This 'subcompositional coherence' is an inherent attribute of the compositions rather a required principle. The formation of a subcomposition $\underline{\mathbf{w}_S}$ from a D-composition $\underline{\mathbf{w}}$ can be geometrically interpreted as the orthogonal projection of the ray associated to $\underline{\mathbf{w}}$ onto the subspace of $\mathbb{R}_+^D$ generated by the positive coordinate axes associated to the components in the subcomposition. Figure 2 shows the subcompositions for the case $D = 3$ and the relationship with the corresponding representatives.

3. The Euclidean compositional space

Any statistical analysis with data from the sample space $\mathcal{C}^D$ needs this space to have an algebraic and metric structures. Remember that such basic concepts as the mean and the variance of a set of data are based on the algebraic and metric structure of the sample space of the data. The strategy that we develop is to define an isomorphism between $\mathcal{C}^D$ and another Euclidean space using the logarithmic function. Despite this isomorphism may not be the unique feasible isomorphism, the rest of the possibilities are still unknown to us.

3.1. A quotient Euclidean space in the real space

The well known classical Euclidean space $\mathbb{R}^D$ is based on the addition and subtraction operations. Because we need to connect the relative information provided by the ratios of components with an existing Euclidean space, the logarithmic function becomes a useful option.

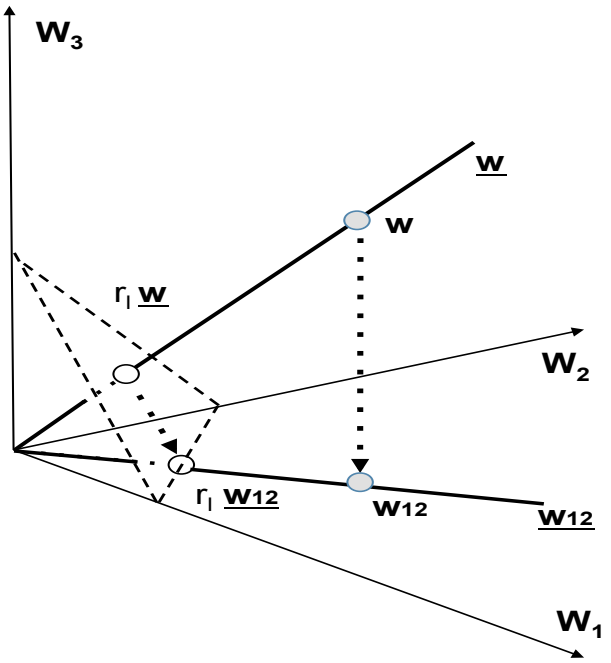

Figure 2: Geometrical interpretation in $\mathbb{R}^3_+$ of a subcomposition $\underline{\mathbf{w}_{12}}$ of a composition $\underline{\mathbf{w}} \in \mathcal{C}^3$. Filled circles are the observational vectors. Empty circles their corresponding linear representatives.

Indeed, the logarithmic transformation from $\mathbb{R}^D_+$ to $\mathbb{R}^D$ suggests defining in $\mathbb{R}^D$ an equivalence relation in correspondence with the compositional equivalence relation $\sim$ defined in $\mathbb{R}^D_+$. Note that if $\mathbf{w} \sim \mathbf{w}^*$, then $\log \mathbf{w} - \log \mathbf{w}^*$ of $\mathbb{R}^D$ is a multiple of the vector of unities $\mathbf{1}_{\mathrm{D}} = (1, \ldots, 1)' \in \mathbb{R}^D$.

Definition 3.1. Two vectors $\mathbf{z}$ and $\mathbf{z}^*$ in $\mathbb{R}^D$ are *equivalent*, written $\mathbf{z} \equiv \mathbf{z}^*$, if a constant λ exists such that $\mathbf{z}^* - \mathbf{z} = \lambda \mathbf{1}_{\mathrm{D}}$. The equivalence class $\{\mathbf{z} + \lambda \mathbf{1}_{\mathrm{D}} : \lambda \in \mathbb{R}\}$ generated by the vector $\mathbf{z}$ in $\mathbb{R}^D$ is denoted by $\underline{\underline{\mathbf{z}}}$. The set of all these classes is the quotient space $\mathbb{R}^D/\equiv$, denoted by $\mathcal{L}^D$. We denote by ocl (from onesclass) the mapping from $\mathbb{R}^D$ to $\mathcal{L}^D$ which assigns each vector $\mathbf{z} \in \mathbb{R}^D$ to the class $\underline{\underline{\mathbf{z}}}$

$$\begin{array}{rccc} \mathrm{ocl}: & \mathbb{R}^D & \to & \mathcal{L}^D \\ & \mathbf{z} & \to & \underline{\underline{\mathbf{z}}}\,. \end{array} \tag{6}$$

Figure 3 shows that the classes $\underline{\underline{\mathbf{z}}}$ can be geometrically interpreted by straight lines parallel to $\mathbf{1}_{\mathrm{D}}$. A simple criterion for selecting a representative of an equivalence class $\underline{\underline{\mathbf{z}}}$ is to assign the intersection point of the straight line associated to this class and the orthogonal hyperplane by the origin

$$\mathrm{V}_D = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}'\mathbf{1}_{\mathrm{D}} = 0\}\,. \tag{7}$$

Definition 3.2. We denote by $\mathrm{r}_{\mathrm{V}_D}$ the one-to-one mapping which assigns each class $\underline{\underline{\mathbf{z}}}$ to this representative

$$\begin{array}{rccc} \mathrm{r}_{\mathrm{V}_D}: & \mathcal{L}^D & \to & \mathrm{V}_D \\ & \underline{\underline{\mathbf{z}}} & \to & \mathbf{z} - \frac{\sum_{j=1}^{D} z_j}{D}\mathbf{1}_{\mathrm{D}} = \mathbf{H}_D\mathbf{z}\,, \end{array} \tag{8}$$

where $\mathbf{H}_D$ is the $D \times D$ *centering matrix*, that is $\mathbf{H}_D = \mathbf{I}_D - D^{-1}\mathbf{J}_D$ ($\mathbf{I}_D$ is the identity matrix of order $D \times D$, and $\mathbf{J}_D = \mathbf{1}_{\mathrm{D}}\mathbf{1}_{\mathrm{D}}'$).

Definition 3.3. The sum of two classes $\underline{\underline{\mathbf{z}}}$ and $\underline{\underline{\mathbf{z}^*}}$ in $\mathcal{L}^D$ is defined as $\underline{\underline{\mathbf{z}}} + \underline{\underline{\mathbf{z}^*}} = \underline{\underline{\mathbf{z} + \mathbf{z}^*}}$, and the product of an equivalence class $\underline{\underline{\mathbf{z}}}$ by a constant $\alpha \in \mathbb{R}$ is defined by $\alpha\underline{\underline{\mathbf{z}}} = \underline{\underline{\alpha\mathbf{z}}}$.

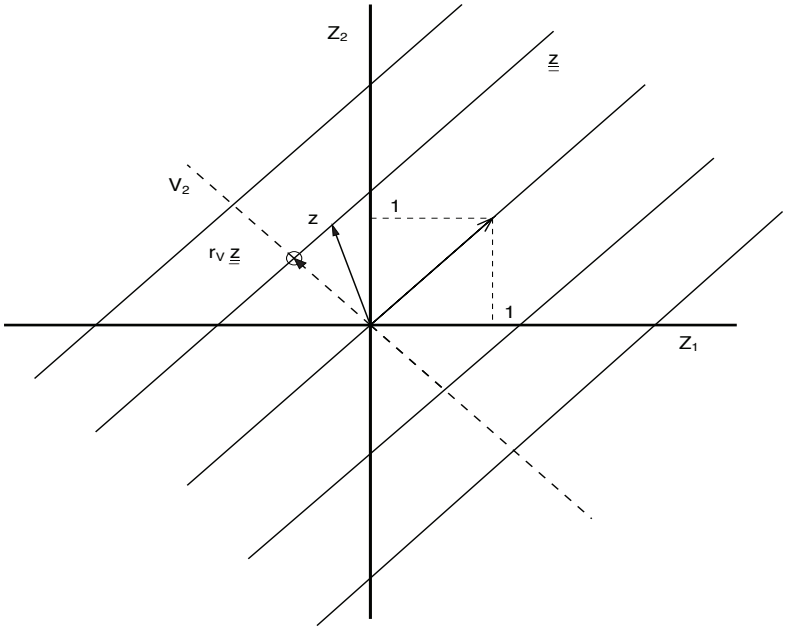

Figure 3: Selection of the representative $r_{V_D}\underline{\underline{z}}$ (empty circle) for an equivalence class $\underline{\underline{z}}$ in $\mathcal{L}^2 = \mathbb{R}^2/\equiv$. Dashed line is the orthogonal hyperplane to vector $\mathbf{1}_D$.

With these definitions, the quotient space $\mathcal{L}^D$ becomes a real vector space. The class of $\underline{\underline{\mathbf{0}_D}}$ is the neutral element and the opposite of $\underline{\underline{z}}$ is the class $\underline{\underline{-\mathbf{z}}}$. Moreover, the mapping r_{V_D} is an isomorphism between the vector space $(\mathcal{L}^D, +, \cdot)$ and the subspace V_D of $\mathbb{R}^D$ (Equation 7). Since the dimension of V_D is $D-1$, the dimension of the vector space $\mathcal{L}^D$ will be also equal to $D-1$.

The vector space structure defined in $\mathcal{L}^D$ is coherent with a subcompositional analysis because one can define subvectors in the space V_D and reproduce them in $\mathcal{L}^D$ using the inverse mapping ${r_{V_D}}^{-1}$. More precisely, the mapping sub_S (Equation 5) corresponds to orthogonal projection of the hyperplane V_D (Equation 7) onto the subspace of $\mathbb{R}^D$ defined implicitly by

$$\{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}'\mathbf{1}_D = 0; z_{j_1} = 0; \ldots, z_{j_{(D-s)}} = 0\} ,$$

where $j_1, \ldots, j_{(D-s)}$ are the sub indexes of the no-selected components in the subvector.

Given that the elements of $\mathcal{L}^D$ can be interpreted as straight lines parallel to vector $\mathbf{1}_D$, one can define the distance between the two classes $\underline{\underline{z}}$ and $\underline{\underline{z^*}}$ of $\mathcal{L}^D$ as the Euclidean distance between these two straight lines in $\mathbb{R}^D$. This distance is equal to the length of the difference vector $r_{V_D}(\underline{\underline{z}}) - r_{V_D}(\underline{\underline{z^*}})$ (Figure 4).

Following this strategy, it is possible to reproduce the Euclidean structure defined on $V_D \subset \mathbb{R}^D$ on $\mathcal{L}^D$.

Definition 3.4. For each $\underline{\underline{z}}, \underline{\underline{z^*}} \in \mathcal{L}^D$, we define the $\mathcal{L}$-inner product $<\underline{\underline{z}}, \underline{\underline{z^*}}>_{\mathcal{L}}$ as the usual inner product $<r_{V_D}\underline{\underline{z}}, r_{V_D}\underline{\underline{z^*}}>$ in $\mathbb{R}^D$.

It follows that $<\underline{\underline{z}}, \underline{\underline{z^*}}>_{\mathcal{L}} = \mathbf{z}'\mathbf{H}_D\mathbf{z}^*$. Then it is possible to define a norm and a distance in $\mathcal{L}^D$ from the $\mathcal{L}$-inner product.

Definition 3.5. The $\mathcal{L}$-norm of an equivalence class $\underline{\underline{z}} \in \mathcal{L}^D$ is given by

$$\|\underline{\underline{z}}\|_{\mathcal{L}} = (<\underline{\underline{z}}, \underline{\underline{z}}>_{\mathcal{L}})^{1/2} = (\mathbf{z}'\mathbf{H}_D\mathbf{z})^{1/2}, \tag{9}$$

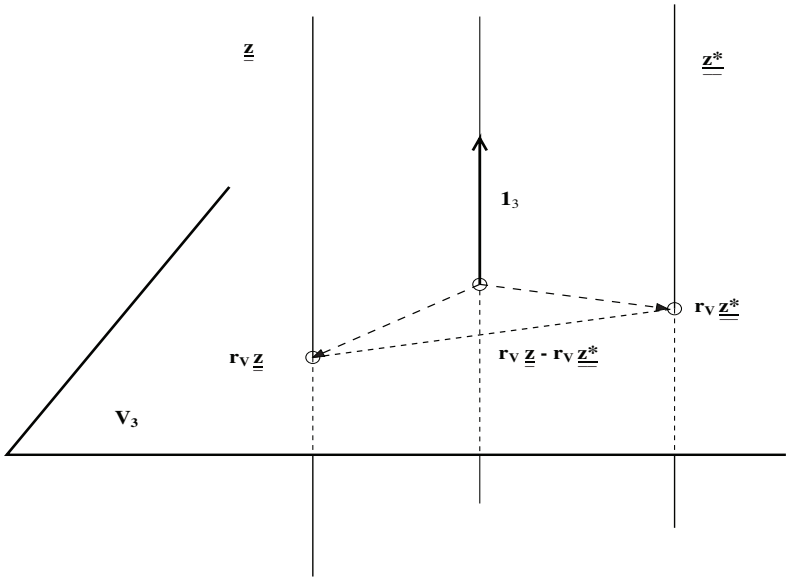

Figure 4: Two equivalence classes $\underline{\underline{\mathbf{z}}}$ and $\underline{\underline{\mathbf{z}}}^*$ of $\mathcal{L}^D$, its corresponding representatives $\mathrm{r}_{\mathrm{V}_D}\,\underline{\underline{\mathbf{z}}}$ and $\mathrm{r}_{\mathrm{V}_D}\,\underline{\underline{\mathbf{z}}}^*$, and the distance between them (case D=3).

and the $\mathcal{L}$-distance between two classes $\underline{\underline{\mathbf{z}}}$ and $\underline{\underline{\mathbf{z}}}^*$ in $\mathcal{L}^D$ is given by

$$d_{\mathcal{L}}(\underline{\underline{\mathbf{z}}}, \underline{\underline{\mathbf{z}}}^*) = \|\underline{\underline{\mathbf{z}}} - \underline{\underline{\mathbf{z}}}^*\|_{\mathcal{L}} = \left[(\mathbf{z}-\mathbf{z}^*)'\mathbf{H}_D(\mathbf{z}-\mathbf{z}^*)\right]^{1/2}. \tag{10}$$

Since $d_{\mathcal{L}}(\underline{\underline{\mathbf{z}}}, \underline{\underline{\mathbf{z}}}^*) = d(\mathrm{r}_{\mathrm{V}_D}\,\underline{\underline{\mathbf{z}}}, \mathrm{r}_{\mathrm{V}_D}\,\underline{\underline{\mathbf{z}}}^*)$, the following property holds.

Property 3.1. From the definitions (9) and (10), the quotient space $\mathcal{L}^D$ becomes an Euclidean space isometric to the subspace V_D of $\mathbb{R}^D$.

3.2. The logarithmic isomorphism between the quotient spaces

The logarithmic and exponential transformations from $\mathbb{R}_+^D$ to $\mathbb{R}^D$ are compatible with the equivalence relations $\sim$ and $\equiv$ defined in $\mathbb{R}_+^D$ and $\mathbb{R}^D$, respectively, i.e.,

$$\mathbf{w} \sim \mathbf{w}^* \text{ in } \mathbb{R}_+^D \iff \log\mathbf{w} \equiv \log\mathbf{w}^* \text{ in } \mathbb{R}^D\ ,$$

and

$$\mathbf{z} \equiv \mathbf{z}^* \text{ in } \mathbb{R}^D \iff \exp\mathbf{z} \sim \exp\mathbf{z}^* \text{ in } \mathbb{R}_+^D\ .$$

Therefore, these transformations can be extended to the quotient spaces $\mathcal{C}^D$ and $\mathcal{L}^D$.

Definition 3.6. We will symbolize by logc the transformation from $\mathcal{C}^D$ to $\mathcal{L}^D$, i.e.,

$$\begin{array}{rccc} \mathrm{logc}: & \mathcal{C}^D & \longrightarrow & \mathcal{L}^D \\ & \underline{\mathbf{w}} & \longmapsto & \underline{\underline{\log\mathbf{w}}}\ , \end{array} \tag{11}$$

and by expc the inverse transformation from $\mathcal{L}^D$ to $\mathcal{C}^D$, i.e.,

$$\begin{array}{rccc} \mathrm{expc}: & \mathcal{L}^D & \longrightarrow & \mathcal{C}^D \\ & \underline{\underline{\mathbf{z}}} & \longmapsto & \underline{\exp\mathbf{z}}\ . \end{array} \tag{12}$$

The representative in V_D of the equivalence class $\underline{\underline{\log \mathbf{w}}}$ is

$$\mathrm{r}_{\mathrm{V}_D}(\underline{\underline{\log \mathbf{w}}}) = \mathbf{H}_D \log \mathbf{w} = \log \frac{\mathbf{w}}{g(\mathbf{w})} \ ,$$

where $g(\mathbf{w})$ is the geometric mean of the vector $\mathbf{w}$.

Importantly, the function composition $\mathrm{r}_{\mathrm{V}_D} \circ \mathrm{logc}$ is equivalent to the transformation clr (Aitchison 1986). This one-to-one correspondence between $\mathcal{C}^D$ and $\mathcal{L}^D$ allows a real vector space isomorphic to $\mathcal{L}^D$ to be defined in $\mathcal{C}^D$.

Definition 3.7. In correspondence with the sum in $\mathcal{L}^D$, the inner operation $\otimes$ in $\mathcal{C}^D$ is defined as

$$\underline{\mathbf{w}} \otimes \underline{\mathbf{w}^*} = \mathrm{expc}\left(\underline{\underline{\log \mathbf{w}}} + \underline{\underline{\log \mathbf{w}^*}}\right) = \mathrm{expc}\left(\underline{\underline{\log \mathbf{w} + \log \mathbf{w}^*}}\right) = \underline{(w_1 w_1^*, \ldots, w_D w_D^*)'} \ ,$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^D$.

Similarly, in correspondence with the product by a constant in $\mathcal{L}^D$, the external operation $\odot$ in $\mathcal{C}^D$ is defined as

$$\alpha \odot \underline{\mathbf{w}} = \mathrm{expc}\left(\alpha \, \underline{\underline{\log \mathbf{w}}}\right) = \mathrm{expc}\left(\underline{\underline{\alpha \, \log \mathbf{w}}}\right) = \underline{(w_1^\alpha, \ldots, w_D^\alpha)'} \qquad (\underline{\mathbf{w}} \in \mathcal{C}^D) \ (\alpha \in \mathbb{R}) \ .$$

The operations $\otimes$ and $\odot$ are respectively the perturbation and power operations introduced by Aitchison (1986).

Therefore, $(\mathcal{C}^D, \otimes, \odot)$ becomes a real vector space of dimension $D-1$, isomorphic to the quotient space $\mathcal{L}^D$ and to the subspace V_D of $\mathbb{R}^D$. In the commutative group $(\mathcal{C}^D, \otimes)$, the composition $\underline{\mathbf{1}_\mathrm{D}} = \underline{(1, \ldots, 1)'}$ is the neutral element, and the inverse composition $\underline{\mathbf{w}}^{-1}$ of $\underline{\mathbf{w}}$ is the composition $\underline{\mathbf{w}^{-1}} = \underline{(1/w_1, \ldots, 1/w_D)'}$.

Moreover, the structure of real vector space of $(\mathcal{C}^D, \otimes, \odot)$ is compatible with the concept of subcomposition.

Property 3.2. The mapping sub_S defined in Equation 5 is a linear function between the vector spaces $(\mathcal{C}^D, \otimes, \odot)$ and $(\mathcal{C}^s, \otimes, \odot)$. Therefore, it holds that

$$\mathrm{sub}_S(\underline{\mathbf{w}} \otimes \underline{\mathbf{w}^*}) = \underline{\mathbf{w}_S} \otimes \underline{\mathbf{w}_S^*} \quad \text{and} \quad \mathrm{sub}_S(\alpha \odot \underline{\mathbf{w}}) = \alpha \odot \underline{\mathbf{w}_S} \ , \tag{13}$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^D$ and $\alpha \in \mathbb{R}$.

3.3. The compositional space as an affine Euclidean space

Because $(\mathcal{C}^D, \otimes, \odot)$ is a real vector space, it can be viewed as an affine space when the group $(\mathcal{C}^D, \otimes)$ operates on $\mathcal{C}^D$ as a group of transformations.

Definition 3.8. Given a composition $\underline{\mathbf{p}} \in \mathcal{C}^D$, the *perturbation* associated to $\underline{\mathbf{p}}$ is the transformation from $\mathcal{C}^D$ to $\mathcal{C}^D$ defined by

$$\underline{\mathbf{w}} \to \underline{\mathbf{p}} \otimes \underline{\mathbf{w}} \qquad (\underline{\mathbf{w}} \in \mathcal{C}^D) \ .$$

Then we say that $\underline{\mathbf{p}} \otimes \underline{\mathbf{w}}$ is the composition which results when the *perturbation* $\underline{\mathbf{p}}$ is applied to the composition $\underline{\mathbf{w}}$.

Perturbations in the compositional space play the same role as translations play in the real space. Like them, the set of all perturbations in $\mathcal{C}^D$ is a commutative group isomorphic to $(\mathcal{C}^D, \otimes)$. Thus, the composition of two perturbations $\underline{\mathbf{p}_1}$ and $\underline{\mathbf{p}_2}$ is the perturbation associated to $\underline{\mathbf{p}_1} \otimes \underline{\mathbf{p}_2}$. Furthermore, the perturbation associated to $\underline{\mathbf{1}_\mathrm{D}}$ is the identity perturbation which does not produce any change when applied to a composition. Also, for any given perturbation

$\underline{\mathbf{p}}$ there is the inverse perturbation $\underline{\mathbf{p}}^{-1}$ which undoes the changes produced by $\underline{\mathbf{p}}$. Finally, given two compositions $\underline{\mathbf{w}}'$ and $\underline{\mathbf{w}}^* \in \mathcal{C}^D$, a unique perturbation $\underline{\mathbf{p}}$ exists which transforms $\underline{\mathbf{w}}$ on $\underline{\mathbf{w}}^*$. This perturbation is

$$\underline{\mathbf{p}} = \underline{\mathbf{w}}^* \otimes \underline{\mathbf{w}}^{-1} = \underline{\left(\frac{w_1^*}{w_1}, \ldots, \frac{w_D^*}{w_D}\right)'} ,$$

the *perturbation difference* between $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$. Thus, the measurement of the 'difference' between two compositions is defined from the ratios between the components of compositions.

The one-to-one transformations logc (Equation 11) and expc (Equation 12) between $\mathcal{C}^D$ and $\mathcal{L}^D$ allow the real Euclidean structure defined on $\mathcal{L}^D$ to be transfered to $\mathcal{C}^D$.

Definition 3.9. The *compositional inner product* of two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$ will be equal to

$$<\underline{\mathbf{w}}, \underline{\mathbf{w}}^*>_{\mathcal{C}} = <\underline{\log \mathbf{w}}, \underline{\log \mathbf{w}^*}>_{\mathcal{L}} = (\log \mathbf{w})' \mathbf{H}_D \log \mathbf{w}^* .$$

Importantly, $<\underline{\mathbf{w}}, \underline{\mathbf{w}}^*>_{\mathcal{C}} = <\operatorname{clr} \mathbf{w}, \operatorname{clr} \mathbf{w}^*>$, i.e., the standard inner product of the clr-transformed vectors.

From this inner product in $\mathcal{C}^D$ we can define a norm and a distance in the compositional space.

Definition 3.10. The *compositional norm* of a composition $\underline{\mathbf{w}} \in \mathcal{C}^D$ will be given by

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}} = (<\underline{\mathbf{w}}, \underline{\mathbf{w}}>_{\mathcal{C}})^{1/2} = \left[(\log \mathbf{w})' \mathbf{H}_D \log \mathbf{w}\right]^{1/2} ,$$

and the *compositional distance* between two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$ of $\mathcal{C}^D$ is given by

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*) = \left[(\log \mathbf{w}^* - \log \mathbf{w})' \mathbf{H}_D (\log \mathbf{w}^* - \log \mathbf{w})\right]^{1/2} .$$

The distance $d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*)$ defined on $\mathcal{C}^D$ is equivalent to the Aitchison distance (Aitchison, Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2000) that can be expressed as the typical Euclidean distance between the corresponding clr-transformed vectors.

Property 3.3. In relation to subcompositions, the distance $d_{\mathcal{C}}$ satisfies what is known as *subcompositional dominance*, according to which

$$d_{\mathcal{C}}(\underline{\mathbf{w}_S}, \underline{\mathbf{w}_S^*}) \leq d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}^*}) ,$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}}^* \in \mathcal{C}^D$ and for any subcomposition S.

Proof. It is sufficient to demonstrate that the compositional norm of a composition $\underline{\mathbf{w}}$ is greater or equal to the compositional norm of a subcomposition $\underline{\mathbf{w}_S}$ obtained by removing one of its components. If, without lack of generality, we assume that $\underline{\mathbf{w}_S} = \underline{(w_1, \ldots, w_{D-1})'}$, then it holds that

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}^2 = \|\underline{\mathbf{w}_S}\|_{\mathcal{C}}^2 + \frac{1}{D(D-1)} \left(\log w_1 + \ldots + \log w_{D-1} - (D-1) \log w_D\right)^2 ,$$

and $\|\underline{\mathbf{w}}\|_{\mathcal{C}} \geq \|\underline{\mathbf{w}_S}\|_{\mathcal{C}}$.

The subcompositional dominance property of the Euclidean space $\mathcal{C}^D$ correlates with the traditional property at the real space $\mathbb{R}^D$, according to which the distance between the orthogonal projections of two points on any subspace is never greater than the original distance between the points. In practical terms, this property also admits the following interpretation: given two observational vectors $\mathbf{w}_S$ and $\mathbf{w}_S^*$, if one adds supplementary components to both vectors to form, respectively, the vectors $\mathbf{w}$ and $\mathbf{w}^*$, then the difference between the new vectors must be at least equal to the difference between the initial vectors.

Palarea-Albaladejo, Martín-Fernández, and Soto (2012) present examples to illustrate that other usual distances, like the typical Euclidean or the angular distances, do not verify this property. As a consequence, when one applies these distances or one calculates related statistics (e.g., correlation coefficient), some misleading results can be obtained.

Since $\mathcal{C}^D$ is a real vector space of dimension $D-1$, any composition $\underline{\mathbf{w}}$ could be identified with its $D-1$ coordinates relative to a basis of $\mathcal{C}^D$. In practice, we can obtain a basis of $\mathcal{C}^D$ from a basis of the subspace V_D of $\mathbb{R}^D$. Indeed, if $\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}$ is a basis of V_D then $\mathrm{expc}(\mathrm{r}_{\mathrm{V}_D}^{-1}\,\mathbf{v}_1), \ldots, \mathrm{expc}(\mathrm{r}_{\mathrm{V}_D}^{-1}\,\mathbf{v}_{D-1})$ is a basis of $\mathcal{C}^D$, and the coordinates of a composition $\underline{\mathbf{w}}$ relative to this basis coincide with the coordinates of $\mathrm{r}_{\mathrm{V}_D}(\underline{\log\,\mathbf{w}})$ relative to $\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}$.

Definition 3.11. Let $\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}$ be a basis of V_D, and let $\mathbf{V}$ be the $D \times (D-1)$ matrix $[\mathbf{v}_1 : \ldots : \mathbf{v}_{D-1}]$. Then the coordinates of the composition $\underline{\mathbf{w}}$ relative to the basis $\mathrm{expc}(\mathrm{r}_{\mathrm{V}_D}^{-1}\,\mathbf{v}_1), \ldots, \mathrm{expc}(\mathrm{r}_{\mathrm{V}_D}^{-1}\,\mathbf{v}_{D-1})$ are the components of the vector $(\mathbf{FV})^{-1}\mathbf{F}\log\mathbf{w}$, where $\mathbf{F} = [\mathbf{I}_{D-1} : -\mathbf{1}_{D-1}]$.

Note the expression of the coordinates of $\underline{\mathbf{w}}$ will depend on the matrix $\mathbf{V}$ we selected. These coordinates are usually known as *logratio coordinates* because they are always expressed in terms of logarithms of ratios of components. For example, for

$$\mathbf{V} = \begin{bmatrix} 1-1/D & -1/D & -1/D & \ldots & -1/D \\ -1/D & 1-1/D & -1/D & \ldots & -1/D \\ -1/D & -1/D & 1-1/D & \ldots & -1/D \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1/D & -1/D & -1/D & \ldots & 1-1/D \\ -1/D & -1/D & -1/D & \ldots & -1/D \end{bmatrix}, \tag{14}$$

the coordinates of $\underline{\mathbf{w}}$ relative to $\mathbf{V}$ are equal to $(\log(w_1/w_D), \ldots, \log(w_{D-1}/w_D))'$. In this case, the logratio coordinates coincide with the *additive logratio* transformation (alr) introduced by Aitchison (1986).

When one is making a statistical analysis it is recommendable to select orthonormal basis in $\mathcal{C}^D$ because the metrics properties are preserved under a change of basis. This fact guarantees the invariance of the results under a change of basis. To select an orthonormal basis it suffices that the matrix $\mathbf{V}$ verifies the two identities $\mathbf{V}'\mathbf{V} = \mathbf{I}_{D-1}$ and $\mathbf{V}\mathbf{V}' = \mathbf{H}_D$. In this case, the mapping that assigns composition $\underline{\mathbf{w}}$ to its logratio coordinates is the *isometric logratio* transformation $\mathrm{ilr}_{\mathbf{V}}$ relative to matrix $\mathbf{V}$ (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003), that is

$$\begin{array}{rccl} \mathrm{ilr}_{\mathbf{V}}: & \mathcal{C}^D & \longrightarrow & \mathbb{R}^{d-1} \\ & \underline{\mathbf{w}} & \longmapsto & \mathrm{ilr}_{\mathbf{V}}\,\underline{\mathbf{w}} = (\mathbf{FV})^{-1}\mathbf{F}\log\mathbf{w}\ . \end{array}$$

In practice, it is very useful to select a basis that facilitates the interpretation of the logratio coordinates. Egozcue and Pawlowsky-Glahn (2005) describe a stepwise procedure to make an orthonormal basis of $\mathcal{C}^D$ from sequential binary partitions of components of the observational vectors of $\mathbb{R}_+^D$.

4. Final remarks and conclusions

Because all Euclidean spaces of the same dimension are isometric, the sample space of CoDa $\mathcal{C}^D$ is isometric to $\mathbb{R}^{D-1}$. This fact allows all the statistical procedures that we naturally apply on the real space $\mathbb{R}^{D-1}$ to be applied to CoDa. The isomorphism presented in this article is based on the logaritmic function. From a theoretical point of view, other approaches could be possible but are unknown to us. With our approach, the compositional quotient space $\mathcal{C}^D$ has an algebraic and a metric structure induced by the isomorphism. Consequently, it suffices to work with the logratio coordinates of the compositions with respect to an orthonormal

basis on $\mathcal{C}^D$ (Mateu-Figueras, Pawlowsky-Glahn, and Egozcue 2011). That is, our CoAn is in essence a *logratio* CoAn, that is an analysis of CoDa based on the logarithm of the information provided by the ratios.

The fact that our analysis focuses on ratios means that it can be applied directly to the original data of $\mathbb{R}^D_+$, to the simplex $\mathcal{S}^D$, to the strictly positive orthant of the unit hypersphere Sph^D_+, to the hyperbolic surface Hip^D_+ or to any other representative. Moreover, when working with logratio coordinates all of the statistical procedures that are defined in $\mathbb{R}^{D-1}$, both descriptive and inferential, are transferred to the space $\mathcal{C}^D$. The application of CoAn leads to the assumption that the group of perturbations is the operating group on the compositional space, in the same manner as we assume that the translations is the operating group in the real space. This is the keystone of the methodology introduced by Aitchison (1986). In fact, it means accepting that the 'difference' between two compositions $\underline{\mathbf{w}} = \underline{(w_1, \ldots, w_D)'}$ and $\underline{\mathbf{w}^*} = \underline{(w_1^*, \ldots, w_D^*)'}$ is based on the ratios w_j^*/w_j between parts instead of on the arithmetic differences $w_j^* - w_j$, according to the 'relative scale' property. Therefore, for example, the difference between the compositions $\underline{(0.980, 0.010, 0.010)'}$ and $\underline{(0.970, 0.002, 0.028)'}$ is more than three times greater than the distance between $\underline{(0.300, 0.200, 0.500)'}$ and $\underline{(0.200, 0.300, 0.500)'}$. The relative scale property of CoAn justifies the choice of the logarithm transformation to measure the difference between two compositions.

The CoAn applies only in the open orthant $\mathbb{R}^D_+$. That is, the components of the observational vectors must be strictly positive. This limitation is certainly a difficulty because often the observations contain zeros. However, when the zeros are rounded zeros or are count zeros they can be preprocessed using techniques inspired by techniques for missing data (Palarea-Albaladejo and Martín-Fernández 2015) that make a replacement by a small value. When the zero is an essential zero, that is, the zero value is a *true* value, it makes no sense to replace the zero by a small value. In this case, the analysis should take into account the presence and absence of zeros, that is, the pattern of zeros. Both descriptive and inferential analysis should be performed among the groups defined by the pattern of zeros.

Some researchers, for example Watson and Philip (1989), consider that the appropriate group to operate on compositions is the rotations on the sphere and not the perturbations on which the logratio CoAn is based. Watson and Philip (1989) represent a composition $\underline{\mathbf{w}}$ from the components of the unitary vector $\mathbf{w}/||\mathbf{w}||$, that is, from the cosine of the different angles that $\mathbf{w}$ forms with the axes of coordinates. Then, the angle formed by two observation vectors $\mathbf{w}$ and $\mathbf{w}^*$ is taken as the appropriate measure from which to define the distance between the two compositions. Others, for example Wang *et al.* (2007) and Scealy and Welsh (2011), also apply the methodology of Watson and Philip (1989) after applying the scale-invariant transformation $\mathbf{w} \to (\mathbf{w}/\sum_{j=1}^D w_j)^{1/2}$ to the observations. Thus, they work with the components of the unitary vector $(\mathbf{w}/\sum_{j=1}^D w_j)^{1/2}$ rather than the coordinates of the vector $\mathbf{w}/||\mathbf{w}||$. From these approaches, which are based on the representation of the compositions on the positive orthant of the unit hypersphere centred at the origin, the authors apply the statistical analysis (characteristic) of *directional statistics*, based on the von Mises-Fisher distribution. As stated in Aitchison (1982)'s final discussion, the problems of this approach derives from the fact that the von Mises-Fisher distribution is defined on the whole unit hypersphere and not only on the positive orthant. This leads to problems when the components of $\mathbf{w}$ are too close to 0. Aitchison (1982) also points out the difficulties that CoAn based on the spherical representation of the compositions encounters when dealing with problems related to independence and regression. Neither is it possible from this representation to easily relate the statistics that describe a set of compositions $\underline{\mathbf{w}_1}, \ldots, \underline{\mathbf{w}_n}$ of $\mathcal{C}^D$ to the statistics of the subcompositions $\underline{\mathbf{w}_{S,1}}, \ldots, \underline{\mathbf{w}_{S,n}}$.

To conclude, the most relevant results shown in this article are:

- A *composition* is an equivalence class and its sample space is the quotient space $\mathcal{C}^D$. Geometrically, the compositions are semi-straight lines by the origin of the positive orthant of the space $\mathbb{R}^D_+$. We refer any analysis of these equivalence classes as *compositional*

analysis (CoAn). Regardless the use of the logarithm function or a transformation, when an analyst decides to do a CoAn he or she is assuming that the sample space of the data is the compositional space $\mathcal{C}^D$, which means in fact an acceptance of the 'scale invariance' principle of CoDA.

- The logarithmic and exponential transformations provide the space $\mathcal{C}^D$ with an Euclidean space structure. We denominate *logratio CoAn* the compositional analysis developed from this structure of $\mathcal{C}^D$. It agrees with the methodology introduced by Aitchison (1982), based on a logratio relative scale of measurement of the difference between two compositions.

- The logratio CoAn allows us to carry out the standard statistical analyses on the logratio coordinates.

- The logratio CoAn allows us to apply the subcompositional analysis in a natural and intuitive way, giving results which are coherent with those obtained from the whole compositions.

- The logratio CoAn has the drawback of being unable to operate directly with compositions with zero values. Applying preprocessing techniques to replace rounded and count zeros is then recommended. A statistical analysis in the presence of essential zeros must take into account the groups defined by the pattern of zeros.

- When the techniques for analysing directional data are restricted to compositions, they must be considered to be a CoAn. Even though these analyses do not have the problem of the zeros it is still impossible to guarantee that coherent results will always be obtained in inferential studies (e.g., confidence regions), that is, strictly contained within the positive orthant, because the sample space of these analyses is the whole sphere. Moreover, this approach does not guarantee that a subcompositional analysis will produce results that concur with the results of the analysis of the whole composition.

5. Acknowledgements

This work has been partially financed by the Ministerio de Economía y Competitividad (Ref: MTM2015-65016-C2-1-R) and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (Ref: 2014 SGR 551).

References

Aitchison J (1982). "The Statistical Analysis of Compositional Data (with Discussion)." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **44**, 139–177.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison J (1990). "Comment on *Measures of Variability for Geological Data*, by D.F. Watson and G.M. Philip." *Mathematical Geology*, **22**, 223–226.

Aitchison J (1991). "Delusions of Uniqueness and Ineluctability." *Mathematical Geology*, **23**, 275–277.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000). "Logratio Analysis and Compositional Distance." *Mathematical Geology*, **32**, 271–275.

Egozcue JJ, Barceló-Vidal C, Martín-Fernández JA, Jarauta-Bragulat E, Díaz-Barrero JL, Mateu-Figueras G (2011). "Elements of simplicial linear algebra and geometry." In Pawlowsky-Glahn and Buccianti (2011), pp. 141–157. 378 p.

Egozcue JJ, Pawlowsky-Glahn V (2005). "Groups of Parts and their Balances in Compositional Data Analysis." *Mathematical Geology*, **37**, 795–828.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*, **35**, 279–300.

Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011). *Dealing with Zeros*. John Wiley & Sons, Ltd.

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011). "The Principle of Working on Coordinates." In Pawlowsky-Glahn and Buccianti (2011), pp. 31–42. 378 p.

Palarea-Albaladejo J, Martín-Fernández JA (2015). "zCompositions - R Package for Multivariate Imputation of Nondetects and Zeros in Compositional Data sets." *Chemometrics and Intelligent Laboratory Systems*, **143**, 85–96.

Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012). "Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data." *Journal of Classification*, **29**, 144–169.

Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons. 378 p.

Scealy JL, Welsh AH (2011). "Regression for Compositional Data by Using Distributions Defined on the Hypersphere." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 351–375.

Scealy JL, Welsh AH (2014). "Colours and cocktails: Compositional Data Analysis 2013 Lancaster Lecture." *Australian & New Zealand Journal of Statistics*, **56**, 145–169.

Wang H, Liu Q, Mok HM, Fu L, Tse WM (2007). "A Hyperspherical Transformation Forecasting Model for Compositional Data." *European Journal of Operational Research*, **179**, 459–468.

Watson DF (1990). "Reply to *Comment on Measures of Variability for Geological Data*, by D.F. Watson and G.M. Philip." *Mathematical Geology*, **22**, 227–231.

Watson DF (1991). "Reply to *Delusions of Uniqueness and Ineluctability*, by J. Aitchison." *Mathematical Geology*, **23**, 279–279.

Watson DF, Philip GM (1989). "Measures of Variability for Geological Data." *Mathematical Geology*, **21**, 233–254.

Affiliation:

Carles Barceló-Vidal
Department of Computer Science, Applied Mathematics and Statistics
University of Girona
E-17071 Girona, Spain
E-mail: carles.barcelo@udg.edu

Josep-Antoni Martín-Fernández
Department of Computer Science, Applied Mathematics and Statistics
University of Girona
E-17071 Girona, Spain
E-mail: josepantoni.martin@udg.edu
URL: http://ima.udg.edu/~jamf/

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 45 *Submitted:* 2015-10-26
September 2016 *Accepted:* 2016-04-20

Austrian Journal of Statistics
September 2016, Volume 45, 73–87.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.122

Compositional Uncertainty Should Not Be Ignored in High-Throughput Sequencing Data Analysis

Gregory B. Gloor
The University of Western Ontario, Canada

Jean M. Macklaim
The University of Western Ontario, Canada

Michael Vu
The University of Western Ontario, Canada

Andrew D. Fernandes
YouKaryote Genomics, Canada

Abstract

High throughput sequencing generates sparse compositional data, yes these datasets are rarely analyzed using a compositional approach. In addition, the variation inherent in these datasets is rarely acknowledged, but ignoring it can result in many false positive inferences. We demonstrate that examination of point estimates of the data can result in false positive results, even with appropriate zero replacement approaches, using an *in vitro* selection dataset with an outside standard of truth. The variation inherent in real high-throughput sequencing datasets is demonstrated, and we show that this variation can be approximated, and hence accounted for, by Monte-Carlo sampling from the Dirichlet distribution. This approximation when used by itself is itself problematic, but becomes useful when coupled with a log-ratio approach commonly used in compositional data analysis. Thus, the approach illustrated here that merges Bayesian estimation with principles of compositional data analysis should be generally useful for high-dimensional count compositional data of the type generated by high throughput sequencing.

Keywords: Bayesian estimation, centred log-ratio, transcriptome, metagenome, 16S rRNA gene sequencing, ALDEx2, R.

1. Introduction

High throughput sequencing studies, that generate as outputs thousands to billions of sequence tags, are becoming the norm in the life sciences. That these experiments generate compositional data can be understood with two statements. First, the total number of sequence tags obtained in an experiment are of no importance. Second, the sequence tags are binned into features where the difference between features is exponential and best explained by log ratios. These features can represent genes as in 16S rRNA gene sequencing, transcriptomics and metagenomics or single-nucleotide variant abundances after differential growth experiments. The experimentalist is interested in knowing which features, if any, are differentially abundant between two or more distinct groups. Furthermore, all experiments of this type explicitly or implicitly examine sub-compositions. Finally, each individual experimental design is analyzed using different sets of underlying assumptions that are derived from historical dogma, despite having the same underlying data structure. Fernandes, Reid,

Macklaim, McMurrough, Edgell, and Gloor (2014) demonstrated that tools developed for one experimental design (e.g. RNA-Seq) do not translate well to other experimental designs (e.g. 16S RNA gene sequencing).

These data are necessarily sparse and complex. There are often hundreds or thousands of features, and the high cost of these experiments prevents the collection of sufficient sequence tags to ensure that all features are covered by at least one sequence tag. Thus, the treatment of features with zero counts is a pervasive problem when treating these data as compositions (Lovell, Müller, Taylor, Zwart, Helliwell, Pawlowsky-Glahn, and Buccianti 2011). It is assumed that features with zero counts across all samples are removed because they are uninformative. For the remainder where one approach is to delete features where one or more samples have zero counts (Lovell *et al.* 2011; Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler 2015). This removes the problem of zero count features at the expense of potentially excluding the most important features from consideration. Another approach, is to replace the zero counts with an expected value calculated in some way. Several approaches with differing underlying assumptions are in use, and Martín-Fernández, Hron, Templ, Filzmoser, and Palarea-Albaladejo (2014) suggested that a Bayesian-Laplace approach to be the most reasonable. Regardless of the method used to treat zero count features, these analyses use maximum-likelihood approaches to determine feature abundance prior to analyses.

We have found that the variation due to sampling alone (technical variation) in compositional datasets derived from high-throughput sequencing is large and inversely related to the number of reads mapping to a fragment, in agreement with theory (Fernandes, Macklaim, Linn, Reid, and Gloor 2013). Ignoring this technical variation can lead to false positive inferences regarding differential abundance if the data are not treated as compositions. We have found that a two-step procedure incorporating a Bayesian estimate of feature abundance along with analyses conducted after a centred-log-ratio transformation markedly improves specificity with no loss of sensitivity, and that the increase in specificity derived almost entirely from the exclusion of low-count (including zero count) features (Fernandes *et al.* 2014).

Our paper explores how the analyses differ when the value of zero is assigned using different approaches with, and without Bayesian estimation of the technical variation. Our initial work showed that a uniform prior added to all values was able to encompass the estimated technical variation in a sparse dataset (Fernandes *et al.* 2013). However, we observed that this approach slightly overestimated technical variation of low count and zero count features, suggesting that this approach had less than optimum power.

We will compare uniform priors that replace 0, uniform priors added to all values, and the prior estimation methods from the zCompositions package (Palarea-Albaladejo and Martín-Fernández 2015) that produce non-uniform estimates of the actual zero value. We will examine a real differential growth experimental dataset for which an objective standard of truth is known. We argue that these results are generalizable across other datasets including RNA-seq datasets and 16S rRNA gene sequencing experiments.

2. Statement of the problem

High throughput sequencing is a technology that delivers thousands to millions of reads that correspond uniquely to genes or other features in a genome, or to bins that represent sequence variants. Figure 1A shows several different study designs that are common in the literature. Regardless of design a very large number of molecules, shown in the orange box in Figure 1A are randomly sampled to produce a library that is then sequenced. The sequencing instrument delivers a much smaller random sample of the actual input and the act of sequencing converts the data from unconstrained to constrained proportional data because the instrument delivers a fixed number of sequence reads. This hard upper bound means that all such analyses generate compositional data regardless of the actual study design. In general, these experiments aim to ask the question, "what gene or feature has a different abundance between groups A and B?"

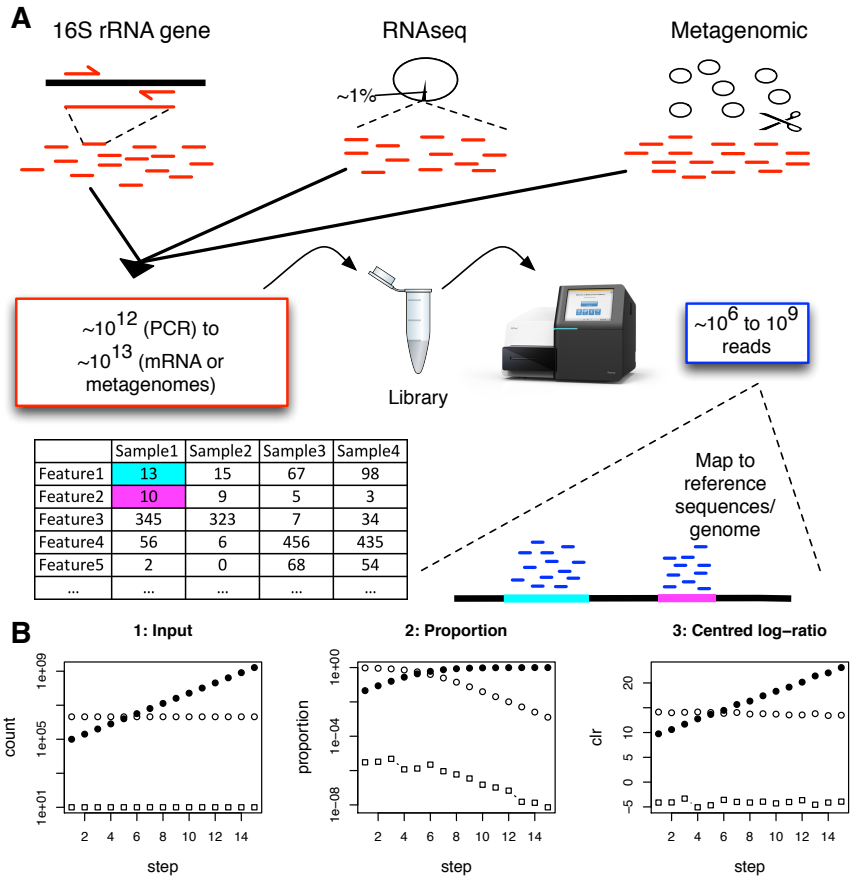

Figure 1: High-throughput sequencing affects the shape of the data. Panel A illustrates the workflow by which high throughput sequencing samples the DNA or RNA from an environment. There are many more molecules that are sampled than can be incorporated into the library, or that can be sequenced on the instrument. The capacity of the instrument itself determines the number of reads observed. The orange box shows the number of molecules in typical initial samples, and the blue box shows the maximum number of reads that are obtained from the instrument. These reads are assigned to features such as genes or operational taxonomic units or other bins, and a table of the reads per feature is output. Panel B illustrates how the data is distorted during the process. The input DNA or RNA usually has no fixed sum and is randomly sampled sequentially during the library preparation and sequencing steps. The output from the instrument is compositional because the instrument can deliver only a fixed upper limit of reads, regardless of the number of molecules in the input. Panel B.1 shows the number of reads in the input tube for 15 steps where the open square and circular features are held at a constant number and the black feature is increasing in abundance by 2-fold each step. Panel B.2 shows the output in proportions (or ppm) after random sampling to a constant sum, as occurs on the sequencer. Panel B.3 shows the shape of the data following centre log-ratio transformation. Note that panels B.1 and B.2 have the y axis on a logarithmic scale, and that the natural scale for centred log-ratio data is logarithmic.

Figure 1B shows how sequencing distorts the data. Many processes examined by high throughput sequencing can be thought of as linear compositional processes. Consider a mixture of many distinctive molecules in vector $x = [x_1, x_2, ...x_n]$ over time or space increments i. For each increment we can determine the abundance of each molecule using Equation 1:

$$x_i = x_0 \times 2^{(\lambda i)} \tag{1}$$

where λ is the incremental rate. If $\lambda = 0$ for all but one of the members of vector x and

$\lambda = 1$ for one member, then one member will double in abundance at each increment and all remaining members will be unchanged. Figure 1B.1 shows such a thought experiment where the values are plotted as counts of molecules. Producing and sequencing a library generates a set of counts per gene that are scaled by the maximum number of reads delivered by the machine. In other words, the counts for gene x_i are per-gene probabilities p_i and are formally equivalent to a random multivariate Poisson sample of the original group of DNA molecules. We can model this process by multimomial (fixed effect) or negative binomial (random effects), implying that the posterior of the model parameters is approximately Dirichlet according to Equation 2:

$$[p_1, p_2, ...p_n] \sim \text{Dirichlet}[x_1, x_2, ...x_n]. \tag{2}$$

A single Dirichlet instance generates a single Bayesian estimate of the underlying posterior probabilities for each feature, and multiple samples generate a full posterior distribution (Holmes, Harris, and Quince 2012; La Rosa, Brooks, Deych, Boone, Edwards, Wang, Sodergren, Weinstock, and Shannon 2012; Fernandes *et al.* 2013). Figure and panel 1B.2 shows the posterior values for a single Dirichlet instance from the counts in Panel 1B.1. Here we can see that the constant sum constraint resulting from the finite read limit of the instrument severely distorts the underlying shape of the data. Figure 1B.3 demonstrates that applying the centred log-ratio transform of Aitchison (1986) to the vector of probabilities p in Panel 1B.1

$$clr(\mathbf{p}) = [log_2 \frac{p_1}{g(p)}, log_2 \frac{p_2}{g(p)}, ...log_2 \frac{p_n}{g(p)}] \tag{3}$$

reconstitutes the essential shape of the data, with the actual data points now showing some variability because of random sampling. In Equation 3, $g(p)$ denotes the geometric mean of the vector p. This transformation is convenient because it reconstitutes the essential shape of the original data, and because there is a one to one mapping between the values in the original and in the transformed dataset. Furthermore, this transformation can easily be interpreted for the experimentalist because it is simply a ratio between the abundance of a gene or feature in the sample and the average abundance of all genes or features in the sample. Of particular note is that $g(p)$ cannot be calculated when 0 values are present, and it is the influence of different means of estimating 0 values that are the primary focus of this report. It needs to be emphasized that an the biological underpinnings of these experiments rarely supports a true observation of 0 samples because the background gene expression level or bacterial abundance can easily fall below the limit of detection.

2.1. Data from high-throughput sequencing are highly variable

Data from high throughput sequencing experiments are often thought of as point estimates despite being random samples of the input molecules, and despite several experiments showing that sequencing the same DNA library will produce somewhat different count tables at the same sequencing depth (Marioni, Mason, Mane, Stephens, and Gilad 2008; Bottomly, Walter, Hunter, Darakjian, Kawane, Buck, Searles, Mooney, McWeeney, and Hitzemann 2011; Gierliński, Cole, Schofield, Schurch, Sherstnev, Singh, Wrobel, Gharbi, Simpson, Owen-Hughes, Blaxter, and Barton 2015). Figure 2 shows an example of this variability. Marioni et al. 2008 did an experiment where two aliquots of the same RNA-seq library were run in duplicate, and the resulting reads were mapped to the $>$ 20000 genes in the human genome. Replicate runs did not return exactly the same number of reads per gene: for example, when the genes in one replicate contained zero counts, the same genes in the other replicate often had non-0 reads. This imprecision extends across the range of per-gene counts as shown for a few replicate read values in Figure 2. This imprecision is proportionally larger for small count values, and smaller for large count values. For example, the range of counts observed in replicate B when genes in replicate A contain one count span the range of 0-14 in this example: a difference of

over 10-fold. By comparison, when genes in replicate A contain 64 counts the corresponding genes in replicate B span counts from 38-91: a difference of less than 50%. See Figure 1 of Fernandes *et al.* (2013) for a demonstration that the proportional error does indeed span the entire range of expression values in this dataset.

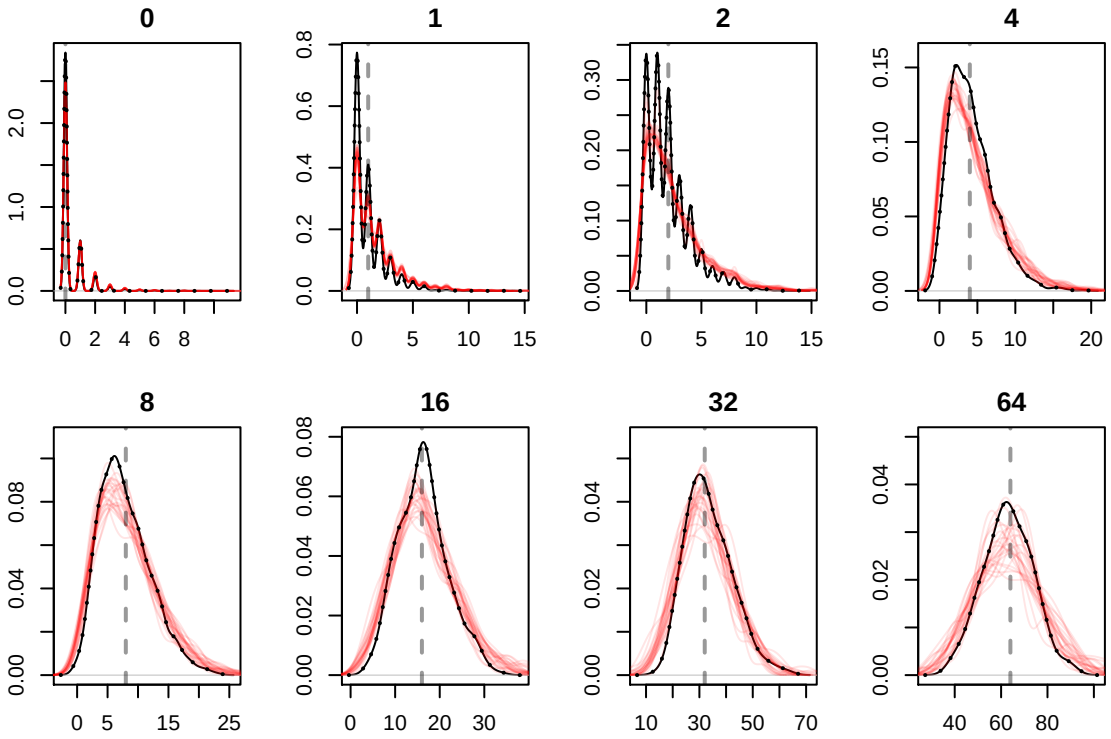

Figure 2: Technical replicate variation, and density plots of the estimation of that variation for an RNA-seq experiment. The black dotted lines show the density of the technical variation of features from one replicate of an RNA-seq dataset when compared to another as a function of the counts in the first replicate. The count value of the first replicate is give above each plot, and the location of this value is shown as the dotted grey vertical line. The red lines show density plots of the inferred technical generated through 25 random instances drawn from a Dirichlet distribution. Data are from the Marioni *et al.* (2008) dataset.

The imprecision can be modelled by Monte-Carlo sampling from a Dirichlet distibution (Fernandes *et al.* 2013, 2014) as in Equation 2. Figure 2 shows density plots comparing the distribution of true technical variation to the distribution of the estimated technical variation obtained by drawing instances from the Dirichlet distribution. That is for a vector of counts x, $x_{Dir} = \text{Dirichlet}[x] \times \sum x$. Sampling multiple Dirichlet instances thus returns a distribution of the posterior probabilities of each feature in the vector x, and conserves probability. These plots show that the Dirichlet instances slightly over-estimate the tails of the distributions, although these conclusions need to be tempered by the lack of datapoint for technical replicates containing double-digit counts. We conclude that drawing Dirichlet instances is an acceptable method to model posterior probabilities in these datasets.

2.2. False positive results because of unaccounted variation

One problem when analyzing such data is that the available datasets – whether derived from 16S rRNA gene sequencing, transcriptomics, or other experiments – are exploratory and so generally lack a standard of truth. This makes it difficult to develop and test tools without modelling a dataset. While modelled datasets have some allure because the parameters can be closely controlled, we prefer to examine the behaviour of real biological datasets because they often have unanticipated error and less predictable behaviour than modelled datasets.

McMurrough, Dickson, Thibert, Gloor, and Edgell (2014) generated a selective growth dataset, hereafter called the 'selex' dataset, for which a standard of truth for many variables is known, and that can be inferred for many others. This dataset compares the growth of a set of 1600 sequence variants in the I-LtrI endonuclease under two conditions. The first condition is a non-restrictive condition where the growth of all variants is unconstrained. The second condition is restrictive for growth, unless the I-LtrI endonuclease is active and can cleave and inactivate the gene encoding *Ccdb*, a DNA gyrase toxin. The gyrase toxin is dose-dependent so cleavage of a fraction of the plasmids containing the gene confers slower growth (Smith and Maxwell 2006), and under the conditions of the assay, the toxin would be bacteriostatic if no cleavage occurred. Thus in this experimental design the difference between inactive variants between the two conditions would be one of dilution alone, and *no variant should become less abundant during the experiment.* Variants that cleave the toxin gene would confer a growth advantage, and would become more abundant over the time of the assay. Furthermore, McMurrough *et al.* (2014) showed that the *in vitro* enzymatic activity of the endonuclease is strongly correlated with the output of the selective growth experiment.

The abundance of each variant in the mixture can be modelled by Equation 1. At time zero if each variant is contained in vector $n_0 = [n_1, n_2, n_3 ... n_{1600}]$, over time increments, the change in abundance in the non-selected growth conditions can be modelled with $\lambda = 1$ and the variation in λ being small. The experimental conditions allowed for approximately 16 doublings, or time increments. Therefore at the last increment of the non-selected time series, we anticipate that the initial relationships between the abundances of each of the 1600 variants will be essentially unchanged. In contrast, the selected variants are under strongly differential selection. Here the most active variants will have $\lambda \approx 1$, that is, these variants grow at the same rate in the selected and unselected conditions. The least active variants will have $\lambda = 0$, that is, these variants will not change in actual abundance during selection, but will become relatively less abundant when compared to their active counterparts. Inactive variants are known to be by far the most prevalent in the samples. Intermediate positive values of λ are expected, and no negative values are expected. Finally, it is possible for individual samples to demonstrate differences in apparent λ under selection. This can occur if a variant is partially active, and cleaves different proportions of the toxin genes in a particular cell by chance. This event is heritable and so would allow cells carrying the same variant to grow at slightly different rates. Thus, the sample in which this occurred would have an apparent increase in λ for that variant in that sample.

The question we wish to address with this dataset is: can we identify from the growth experiment alone which variants are likely to be active? Active variants will have had a maximum of 16 cell doublings becoming much more abundant, inactive variants will stay at the same abundance and variants with partial activity will become only somewhat more abundant. In addition, we wanted to know the effect on our inference of the different approaches to estimating the zero values. We first examined the dataset using a biplot to show the relationship between the samples and the variants.

Figure 3 shows the density and distributions of 0 values in this dataset as summarized by the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015). We can see that this dataset will be very challenging to analyze because the control samples do not contain any 0 values, but most of the variants in the experimental samples contain 0 values in several samples. This high density of 0 values comes about because the number of sequence reads was insufficient, and not because we expected a 0 value in any of the variants. Thus, we must impute the most likely value of 0 in each sample before analysis.

A compositional biplot generated with the compositions R package (van den Boogaart and Tolosana-Delgado 2008) following zero replacement using the CZM approach from the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015) is shown in Figure 4. The first two components of this biplot explained 52.4% and 10.4% of the variance in the data, indicating that this is a good summary of such a complex dataset. The selected and non-selected samples separate clearly on the first component, and this separation is associated

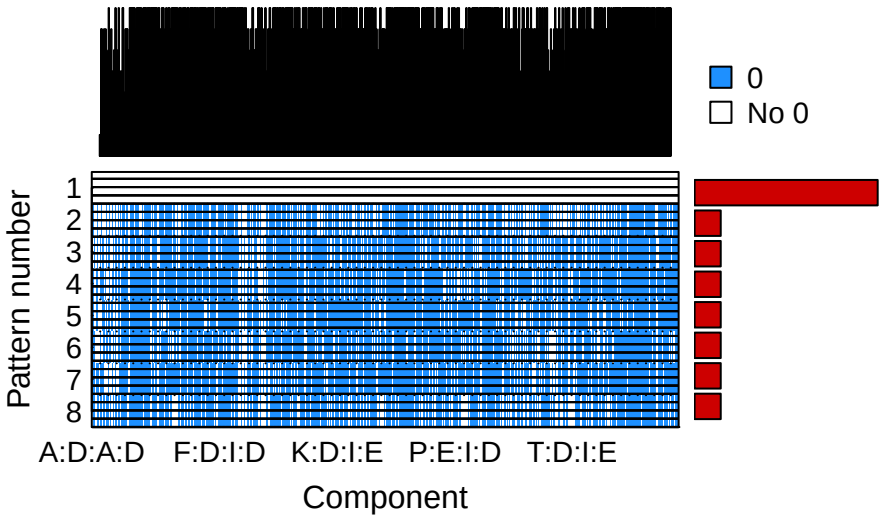

Figure 3: Characteristics of the McMurrough *et al.* (2014) dataset summarized by the zCompositions package. The top panel shows that all parts contain 0 values, and the bottom panel shows that there are 8 patterns in the data. Seven of the samples contain no 0 values, these are the control samples grown in the absence of selection. The samples derived from the selective growth display individual seven different patterns for 0 values likely due to random sampling.

with the variants on the right side, such as A:E:G:E, G:E:G:E, G:E:G:E, etc. McMurrough *et al.* (2014) demonstrated these to be the highly active variants. However, it is difficult to quantitate the magnitude of the abundance change of the variants from this analysis. This figure also shows that the non-selected samples, which cluster on the left side, are essentially redundant since the links between them are exceedingly short. The selected samples on the right side are much more diverse. The differences between the selected samples is largely on component 2, and appears to be driven by the abundance of a small number of variants that also exhibit variation from the bulk on component 2. Inspection of the data finds that this diversity is found in only a few variables such as I:E:V:E, P:D:M:E, A:E:M:E etc, and that these variables separate the X1 and X2 sample sets: interestingly, these sets are from identical experiments performed with different batches of the same cell type. Examination of the underlying count table shows that these variants are indeed different in abundance between the X1 and X2 sets. For example the I:E:V:E variable has 17933 reads in sample X2_DS but has zero reads in sample X1_DS. This is an example of a single stochastic event that conferred a growth advantage to this variant in this sample. It is important to note that this large stochastic variation has important consequences when examining datasets because such variation is not unusual in real biological datasets.

Not accounting for sampling results in many false positive identifications

The selex dataset is unique because we have a validated truth for some of the features that differentiate the conditions (McMurrough *et al.* 2014). In this dataset we have unambiguously biochemically identified variants that are active and those that are not. Based on this prior information, we expect that approximately 60 variants would exhibit substantial activity in this assay, and so substantial deviation from this number would indicate many false positive results.

One approach that is widely used in the literature is to reduce the values from the count table to proportions or normalized counts through Maximum Likelihood approaches, then to conduct univariate statistical tests for each variant, and (sometimes) to correct for multiple

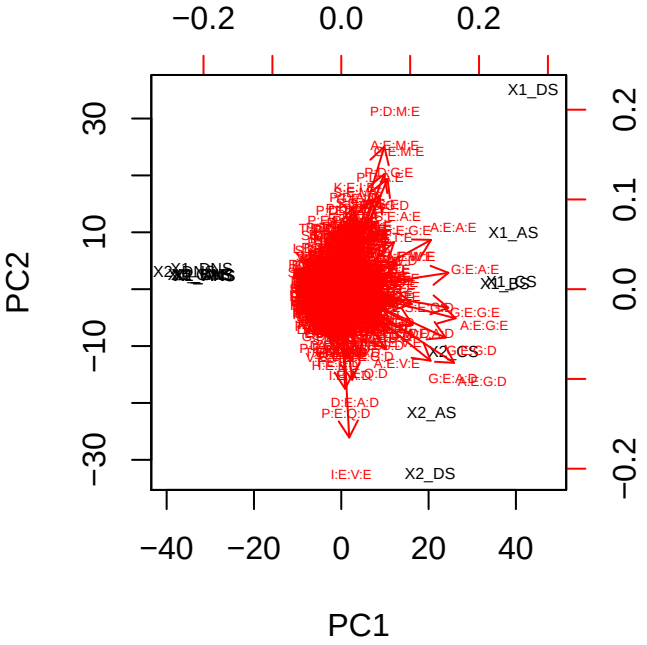

Figure 4: Biplot showing the relationship between samples and variables in the selex dataset. Zero values were adjusted using the count zero multiplicative approach using the zCompositions R package, data were clr-transformed with a base 2 logarithm, and the plot was generated using standard functions from the R statistical programming language. Samples ending with 'NS' are the control non-selected growth samples and form a group at the left side, and samples ending in 'S' are from the selected growth samples and are on the right side. The vast majority of the variables cluster around the centre of the dataset. The selected replicates on the right exhibit some variability. The differences between samples are driven by variation in a small number of variables.

testing (see for example the 16S rRNA analysis methods in Hsiao, McBride, Hsien, Sharon, Hyde, McCue, Codelli, Chow, Reisman, Petrosino, Patterson, and Mazmanian (2013) and this review Di Bella, Bao, Gloor, Burton, and Reid (2013)). These approaches often treat zero values as actual zeros, making no adjustments. Applying this simple method using an unpaired Wilcoxon test, and applying the Benjamini-Hochberg correction (Benjamini and Hochberg 1995) to the resulting P values reveals that 1593 of 1600 variants are identified as having a differential abundance between the selected and non-selected conditions with an adjusted P value cutoff of 0.05. This is clearly at odds with the known biology of the underlying dataset.

A potentially more rigorous, yet still simple approach is to replace the zero values in this dataset using one of the available methods that are implemented in the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015) and then to treat the data as compositions by applying Equation 3 before performing univariate statistical tests. Recall from Figure 1 that this transformation recapitulates the essential shape of the data, and there is a one to one mapping of variant counts to centred log-ratio values. The range of values for zero replacement by different methods are given in the 'Zero assignment' column of Table 1. Interestingly, three of these zero replacement methods returned values greater than 1 for some of the zero values, this likely was a result of the very large difference between the selected and non-selected count values in the two groups. In addition, we applied two other approaches

to deal with zero values. The first, labeled uniform replacement, replaces all zero values with 0.5 but does not adjust other values in the dataset. This is akin to adding a pseudo count to the zero values. The second labelled uniform prior, applies a uniform prior adjustment to *all values* in the dataset. For this we use the minimally informative Jeffrey's prior for the D parts of $Dirichlet[0.5_D]$. The Comp 1 and Comp 2 columns in Table 1 show the percentage variation explained by a compositional biplot using each of these zero adjustments in the first two principle components. Only the biplot that used the square-root Bayesian multiplicative method appears to result in a transformation that explains substantially less of the variation in the dataset.

Table 1: Numbers of distinguishing features identified in the selective growth experiment observed with different approaches to assign prior expectations to zero count features.

Zero assignment	Prior	Comp 1	Comp 2	Point	Dir
Count zero	0.325452 - 0.325910	0.524	0.104	874	91
Geometric Bayesian	0.061279 - 4.890273	0.504	0.108	355	82
Square root	0.006854 - 3.102299	0.452	0.118	1008	DNR
Bayes-Laplace	0.030497 - 4.883747	0.480	0.108	435	133
Uniform replacement	0.5	0.556	0.098	958	74
Uniform Prior	0.5	0.528	0.102	868	84

The utility of these zero replacement approaches to detect univariate differences in this experiment was tested by closing the vectors after prior assignment, applying the centred log-ratio transform to each sample and then subjecting the features to unpaired Wilcoxon tests. Again P values were adjusted using the Benjamini-Hochberg method and an adjusted P value of 0.05 was used as the threshold for significance. Table 1 'Point' shows the results of this approach. Here we see that all of the methods substantially improve upon the naive approach, with between one-quarter and two-thirds of the variables being identified as differential. In this dataset, the square root Bayesian multiplicative method provides the largest number of positive identifications, and the Geometric Bayesian multiplicative correction provides the smallest number of positive identifications, although no method is able to strongly distinguish the known small number of true positives from a much larger number of false positives.

Accounting for sampling reduces false positive identifications

One substantial shortcoming of these approaches is that the inherent technical variation in the dataset is not taken into account. It is becoming an accepted practice to account for the sampling using Dirichlet multinomial mixture models, where each sample is represented by a vector of probabilities, rather than point estimates (Holmes *et al.* 2012). For example, Ding and Schloss (2014) recently used this approach to partition microbiomes into different community states in a robust manner. This approach thus generates a Bayesian posterior estimate of the probabilities associated with each count prior to analysis.

This approach was tested by generating 128 Dir instances of the selex dataset using Equation 2 with Jeffrey's prior, and then conducted per-variant Wilcoxon tests on each instance. The mean Benjamini-Hochberg adjusted P value for each variant was tabulated, and again the cutoff used was an adjusted P value of 0.05. Surprisingly, this approach, which takes into account the inferred technical variation, again resulted in 1593 of the 1600 variants as being differentially abundant between the selected and non-selected groups. This is more than the 868 variants detected when variation was not taken into account but the centred log-ratio transform was applied, and equivalent to the naive method accounting for neither variation nor the compositional nature of the data. Thus, simple averaging across inferred the technical variates is not sufficient to screen out false positive variants in this dataset.

Finally, we combined the Bayesian posterior estimated from 128 Dirichlet instances of the

data and the centred log-ratio transformation of the posterior and used this as the input to significance tests. This method is implemented in the ALDEx2 R package for the analysis of high throughput sequencing datasets (Fernandes *et al.* 2013, 2014), and is available at Bioconductor.

As implemented, the ALDEx2 package uses the uniform zero replacement value of 0.5. One purpose of this investigation was to determine if using one of the more rigorous zero replacement models from the zCompositions package would increase our selectivity because, as shown in Table 1, these adjustments output non-uniform estimates of the underlying value of zero based on abundances of the same feature in different samples.

We applied the same seven methods to adjust the value of zero in the selex dataset, and an overview of the results are shown in the 'Dir' column of Table 1. We again used Wilcoxon tests on the two groups and corrected the resulting P values using the Benjamini-Hochberg approach. Significance was assumed if the mean adjusted P value across all 128 instances was less than 0.05. In this analysis the substituted values of zero in the adjusted datasets serve as prior estimates of the range of values that zero could assume in each of the Monte-Carlo Dirichlet instances. The square root Bayesian multiplicative approach was incompatible this approach because many of the values that replaced zero generated Dirichlet posterior estimates that were not distinguishable from zero. Modelling uniform priors indicated that this occurred when the prior was less than approximately 0.05. The remaining six approaches were compatible with the approach, and resulted in substantially smaller numbers of variants being identified as significantly different between the selected and non-selected groups. In this analysis, the Geometric Bayesian multiplicative, uniform replacement and uniform prior approaches were approximately similar, the count zero multiplicative approaches was nearly as selective, and the Bayes-Laplace approach was least selective.

Figure 5 shows a variance-variance plot of the output from an analysis using the uniform prior replacement with a value of 0.5. Note that in this plot the vast majority of variants have an estimated between group difference of approximately zero, that only a small number have a positive between group difference, and no variants have a strong negative between group difference. This fits with the experimental design where variants could increase in abundance if the endonuclease was active, but not decrease in abundance if it was not. In this plot the variants with a mean Benjamini-Hochberg adjusted P value determined by an unpaired Wilcoxon test are indicated by the large grey dots. Variants that were tested for enzymatic activity *in vitro* are indicated by coloured central dots. Variants that had near wild type enzymatic activity *in vitro* are in the sector marked as > 8. There were four variants that had partial enzymatic activity *in vitro*. Many variants were tested for growth in pure culture. Variants AEAE, SEGE, ADGD and GDAD exhibited variable, partial growth under these conditions, with the GDAD variant exhibiting the weakest growth. Thus, there is a strong relationship between the observed results in this experiment, and the results observed *in vitro*.

There is remarkable concordance between the data viewed in this way, and the same data viewed as a point estimate in the compositional biplot. The biplot shows that the most distinguishing variants between the selected and non-selected groups, i.e., the variants that drive the separation on principle component 1, are those in the upper left quadrant of Figure 5. In addition, the variants that drive the separation on principle component 2, are those that exhibit the largest within-condition difference. For example, the GEME, AEME, PDME, IEVE, PEQD, and DEAD variants that were strongly separated on component 2 on the biplot, are among those with the largest within group difference on the variance-variance plot.

Finally, we examined the effect of the different zero replacement methods on the shape of the variance-variance plot to determine why these different approaches deliver slightly different results after Dirichlet sampling log-ratio transformation. As shown in Figure 6 all the prior estimation methods delivered similar differences between conditions for the true positive variants. These all exhibited an increase in abundance of about 2^{16} relative to their mean abundance in the unselected group. In particular, the CZM plot was remarkably similar to the

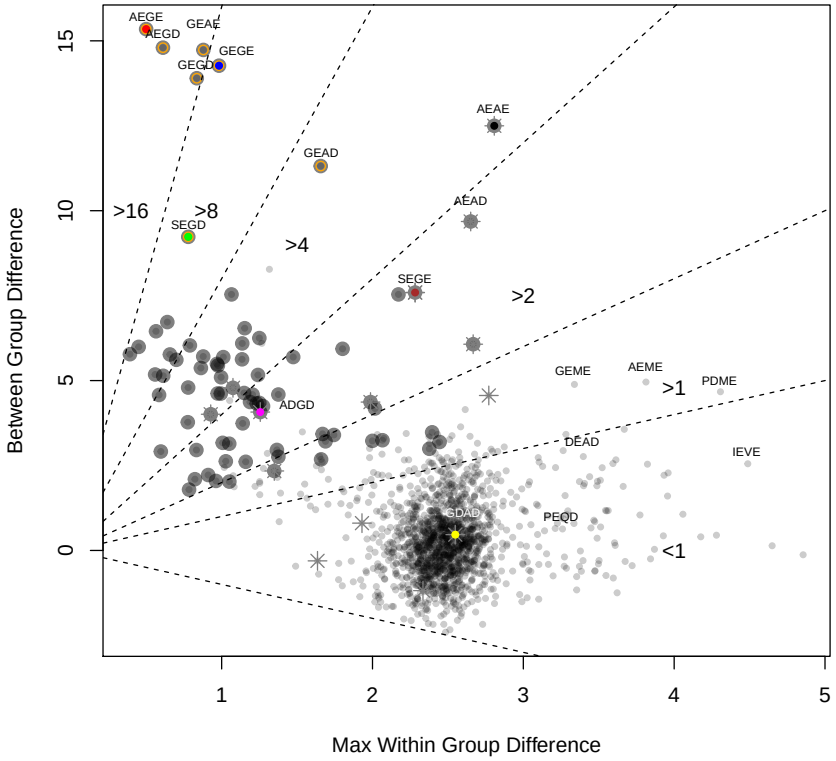

Figure 5: Variance-variance plot showing the median maximum centred log-ratio scaled difference within each group plotted vs. median between group difference for each variant. Dotted lines represent the approximate location of effect sizes, which is calculated as the median between to within group difference. Variants are coloured if their activity was validated *in vitro*, have a star if they failed to grow reliably in individual culture *in vitro*. Variants that exhibit a significant increase in abundance using the Wilcoxon test with a mean Benjamini-Hochberg adjusted P value of > 0.05 are shown as large grey dots. The analysis was done with a uniform prior of 0.5 applied to the dataset. Also shown are the six variants that were outliers on component 2 of the clr biplot in Figure 4.

plot that used a uniform prior of 0.5, with the major difference between the two approaches being a slight broadening of the within-group difference. This is perhaps not surprising since the prior values of zero using this approach are non-uniform in a narrow range of near 0.325. In contrast the non-uniform prior values for zero count variants from both the GBM and BL ranged over much larger values. The vast majority of values were between zero and one, but the GBM method had an average of 197.4 zero replacements that were greater than one, and the BL had an average of 55 replacements that were greater than one. Examination of the variance-variance plots of these two approaches showed that between-group difference for many variants was not significant, but tended to be strongly negative. This result is incompatible with the known biology of the experiment, where no variant is expected to become less abundant than average in the selected dataset. Therefore, in this dataset, the geometric Bayesian multiplicative and the Bayes-Laplace substitution methods are distorting the underlying data. This distortion likely contributes to the greater number of variants identified as significantly different between the selected and non-selected groups.

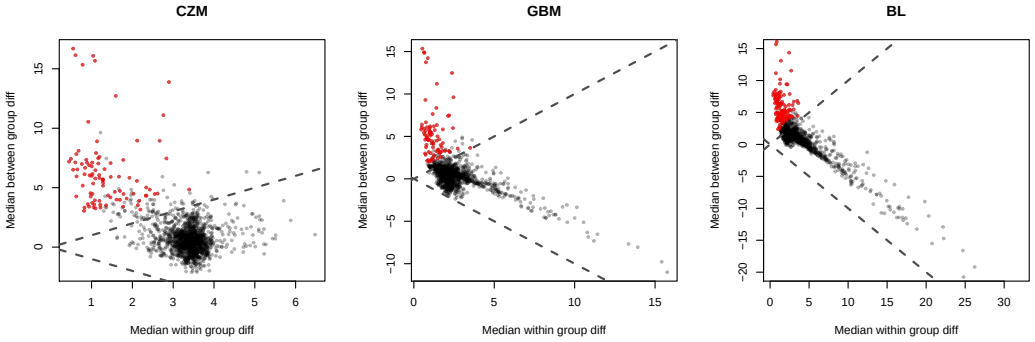

Figure 6: Variance-variance plots showing how the prior values for zero determined by the count zero multiplicative (CZM), geometric Bayesian multiplicative (GBM) and Bayes-Laplace (BL) methods alter the variation of the data. Red dots represent those that are called differential, black dots are not differential and the lines represent effect sizes of 1 and -1. The cutoff used was a Benjamini-Hochberg adjusted P value of 0.05 from an unpaired Wilcoxon test.

3. Discussion

High throughput sequencing datasets are different from other types of datasets to which compositional approaches are often applied, but fall into the general class of 'count compositional' data. However, it is useful to remember that high throughput sequencing datasets result from random sampling of a large number of DNA fragments, and that the act of sequencing these DNA fragments on the instrument results in data that has the constant sum constraint. The estimation of the true abundance of genes or features with low counts exhibits a very large proportional error.

It is tempting to imagine that the large number of counts observed for a given sample, ranging from the thousands to billions, provides great precision in estimating the true values of the genes or OTUs (parts) being examined. This is an erroneous assumption because there are hundreds to thousands of parts in each sample, and many of the parts will be represented by zero reads in some samples. Thus, it is more useful to think of these data as *one instance of the data observed from a single random sample*. When thinking about the data in this way, the reads per part in each sample can be represented as prior values for a Bayesian estimation of their posteriors. The posterior distribution of the underlying abundance of each feature can be estimated by generating multiple instances of the data by sampling from a Dirichlet distribution.

As noted above, these datasets are necessarily very sparse, but in many cases the sparsity is informative. For example, in 16S rRNA gene sequencing it is difficult to argue that a particular taxonomic group would *never* be observed if we generated sufficient sequencing reads. As another example, gene expression is stochastic, and the number of transcripts for a given gene is observed not to be zero when large populations of cells are sampled, even if the gene in question is 'not expressed' (Munsky, Neuert, and van Oudenaarden 2012).

The centred log-ratio approach is intrinsically attractive in a biological context for two reasons. First, it can be intuitively explained to biologists as being similar to quantitative PCR, a familiar technique where the ratio between the gene of interest and a gene assumed to be at a constant level is determined. The centred log-ratio approach merely extends this analogy to the ratio between the gene of interest and all other genes in the system. Second, biologists understand that many of the processes that they study, cell growth, enzyme kinetics, etc, are exponential processes. Less well understood is that the underlying data is not 'set in stone' but actually represents a snapshot of what would have been observed had the experiment been done again.

A common criticism of using log-ratio approaches when analyzing such sparse data is the problem of zero observed counts. Structural zeros, those features that contain zero in every sample, are always excluded, and do not cause problems. However, count zeros that occur in one condition but not the other are problematic because log-ratio transformations cannot be performed when the underlying data contains one or more features with a zero value (Aitchison 1986). Much work has been put into this problem because of the prevalence of features with values of zero are common in many kinds of datasets. Several approaches have been developed to determine the best point estimate of the actual underlying value of zero in these datasets (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015), and they are implemented in zCompositons R package (Palarea-Albaladejo and Martín-Fernández 2015). Less work has been done modelling this in a Bayesian framework where the distribution of probable values for each variable are taken into account.

Here we have examined the effect of using various approaches to estimating the value of zero on both point estimates and Bayesian distributions derived from Dirichlet multinomial sampling. We have found that point estimates, whether modelled as proportions or centre log-ratio transformed values, cannot distinguish features that differ between conditions in a problematic dataset. We found that estimating the technical variation alone is also unsuitable. However, the combination of estimating technical variation and the centre log-ratio transformation provides a large increase in selectivity. We further observe that methods that generate priors in a narrow range give outputs that closely mimic a dataset derived from a differential growth experiment, and that methods that generate priors with broad ranges generate posterior distributions that are different from the known underlying distribution.

The selex dataset is an extreme example of the type of data that is analyzed by high throughput sequencing. It has a small number of features that exhibit a marked difference in abundance between conditions, and is very sparse. Other experimental designs will have much smaller difference in abundance of features. For example, in the case of RNA-seq it is more common to examine differential abundance of a small number of genes that are themselves relatively rare in the cell, and from carefully controlled experiments where the total number of input molecules is similar between conditions. This would be akin to comparing steps 1 and 2 in Figure 1B.1, where no gene or set of genes perturbs the system signfificantly. In this simple case, any approach would likely give reasonable answers. However, comparing gene expression between cells from different tissues, or gene expression in RNA from environmental samples, would introduce extreme distortions in the underlying data and could give false positive and false negative results (Fernandes *et al.* 2013; Macklaim, Fernandes, Di Bella, Hammond, Reid, and Gloor 2013; Fernandes *et al.* 2014). In the case of 16S rRNA gene sequencing experiments, it is likely that many conditions would have wildly divergent underlying abundances because bacterial growth is an exponential process, and such samples are more difficult to analyze.

Acknowledgements

This work was supported by a grant to Greg Gloor by the National Science and Engineering Research Council of Canada.

Correspondence addresses of author(s) should be added at the end of the manuscript.

References

Aitchison J (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall.

Benjamini Y, Hochberg Y (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Bottomly D, Walter NAR, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R (2011). "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-seq and Microarrays." *PLoS One*, **6**(3), e17820. doi:10.1371/journal.pone.0017820.

Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G (2013). "High Throughput Sequencing Methods and Analysis for Microbiome Research." *J Microbiol Methods*, **95**(3), 401–14. doi:10.1016/j.mimet.2013.08.011.

Ding T, Schloss PD (2014). "Dynamics and Associations of Microbial Community Types Across the Human Body." *Nature*, **509**(7500), 357–60. doi:10.1038/nature13178.

Fernandes AD, Macklaim JM, Linn T, Reid G, Gloor GB (2013). "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PLoS ONE*, **8**(7), e67019.

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB (2014). "Unifying the analysis of High-throughput Sequencing Datasets: Characterizing RNA-seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome*, **2**, 15. doi:10.1186/2049-2618-2-15.

Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G, Owen-Hughes T, Blaxter M, Barton GJ (2015). "Statistical Models for RNA-seq Data Derived from a Two-condition 48-replicate Experiment." *Bioinformatics*. doi:10.1093/bioinformatics/btv425.

Holmes I, Harris K, Quince C (2012). "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics." *PLoS One*, **7**(2), e30126. doi:10.1371/journal.pone.0030126.

Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK (2013). "Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders." *Cell*, **155**(7), 1451–63. doi:10.1016/j.cell.2013.11.024.

La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD (2012). "Hypothesis Testing and Power Calculations for Taxonomic-based Human Microbiome Data." *PLoS One*, **7**(12), e52078. doi:10.1371/journal.pone.0052078.

Lovell D, Müller W, Taylor J, Zwart A, Helliwell C, Pawlowsky-Glahn V, Buccianti A (2011). "Proportions, Percentages, ppm: Do the Molecular Biosciences Treat Compositional Data Right?" *Compositional Data Analysis: Theory and Applications*, pp. 193–207.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015). "Proportionality: a Valid Alternative to Correlation for Relative Data." *PLoS Comput Biol*, **11**(3), e1004075. doi:10.1371/journal.pcbi.1004075.

Macklaim MJ, Fernandes DA, Di Bella MJ, Hammond JA, Reid G, Gloor GB (2013). "Comparative Meta-RNA-seq of the Vaginal Microbiota and Differential Expression by *Lactobacillus iners* in Health and Dysbiosis." *Microbiome*, **1**, 15. doi:doi:10.1186/2049-2618-1-12.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008). "RNA-seq: an Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Res*, **18**(9), 1509–17. doi:10.1101/gr.079558.108.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2014). "Bayesian-multiplicative Treatment of Count Zeros in Compositional Data Sets." *Statistical Modelling*, **doi:10.1177/1471082X14535524**, 1:25.

McMurrough TA, Dickson RJ, Thibert SMF, Gloor GB, Edgell DR (2014). "Control of Catalytic Efficiency by a Coevolving Network of Catalytic and Noncatalytic Residues." *Proc Natl Acad Sci U S A*, **111**(23), E2376–83. doi:10.1073/pnas.1322352111.

Munsky B, Neuert G, van Oudenaarden A (2012). "Using Gene Expression Noise to Understand Gene Regulation." *Science*, **336**(6078), 183–7. doi:10.1126/science.1216379.

Palarea-Albaladejo J, Martín-Fernández JA (2015). "zCompositions — R Package for Multivariate Imputation of Left-censored Data under a Compositional Approach." *Chemometrics and Intelligent Laboratory Systems*, **143**(0), 85 – 96. ISSN 0169-7439. doi:http://dx.doi.org/10.1016/j.chemolab.2015.02.019. URL http://www.sciencedirect.com/science/article/pii/S0169743915000490.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Smith AB, Maxwell A (2006). "A Strand-passage Conformation of DNA Gyrase Is Required to Allow the Bacterial Toxin, CcdB, to Access Its Binding Site." *Nucleic Acids Res*, **34**(17), 4667–76. doi:10.1093/nar/gkl636.

van den Boogaart KG, Tolosana-Delgado R (2008). ""compositions": A Unified R Package to Analyze Compositional Data." *Computers & Geosciences*, **34**(4), 320 – 338. ISSN 0098-3004. doi:http://dx.doi.org/10.1016/j.cageo.2006.11.017. URL http://www.sciencedirect.com/science/article/pii/S009830040700101X.

Affiliation:

Gregory B. Gloor
Department of Biochemistry
The University of Western Ontario
London, Ontario, Canada
E-mail: ggloor@uwo.ca
URL: http://www.academicbiography.uwo.ca/profile.php?&n=ggloor

ISSN 1026–597X

Contents

Page