

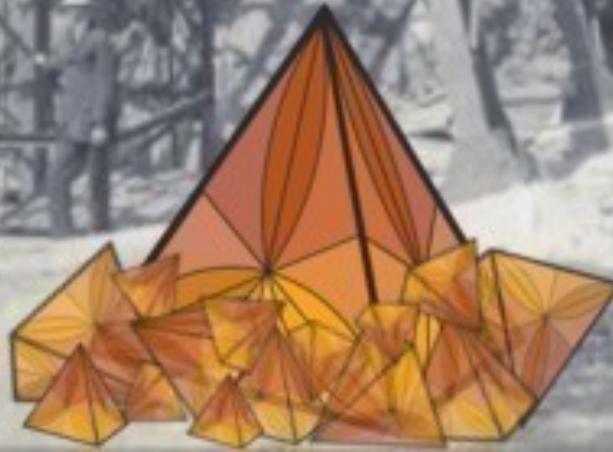
The 7th International Workshop on Compositional Data Analysis

Proceedings book

Karel Hron
and
Raimon Tolosana Delgado (eds.)

C
W
O
R
K
D
2
0
1
7
a

Abbadia
San Salvatore
Siena (Italy)



Foreword

Dear Friends and Colleagues,

We warmly welcome you to Abbadia San Salvatore, for the Seventh International Workshop on Compositional Data Analysis (CoDaWork 2017). As many of you know, this workshop series was established in 2003 as a forum of discussion for people concerned with the statistical treatment and modeling of compositional data or other constrained data sets, and the interpretation of models or applications involving them. The main aim of the CoDaWork series is to bring together specialist researchers, data analysts, postgraduate students, as well as those with a general interest in the field, to summarize and share their contributions and recent developments.

The presented Proceedings consist of 24 papers or extended abstracts of a wide range, from theoretical issues to various applications of compositional data analysis. Different methodological aspects are used in the contributions that reflect well rich history of the field. The first comprehensive introduction to the analysis of compositional data using the logratio methodology is the seminal monograph *The Statistical Analysis of Compositional Data* (1986) by John Aitchison. Since then, much work has been devoted to the theoretical development of the field and to its applications in practice. The *staying-in-the simplex* approach that uses the specific algebraic-geometric structure of the simplex to express compositions using proper coordinates as standard real data, turned out to provide a solution to a number of statistical problems with an application potential in many fields where compositional data naturally arise. However, experience from the last three decades shows that not only a comprehensive theory, but rather fruitful applications justify the recent position of the logratio methodology as a powerful alternative to the other existing approaches to compositional data analysis. Successful application areas are, for example, those from the contributions in these Proceedings: geochemistry, mining, analytical chemistry, ecology, sociology, economics, medicine, and particularly, the biological -omics analytics. Indeed, an increasing interest is devoted to logratio processing of gene expression and omics data, being naturally of relative scale. These and other large-scale compositional data sets bring as well topics related to Big Data and Data Analytics. Further theoretical developments focus on more complex concepts of compositional data analysis, represented, e.g., by continuous compositions (probability densities) and compositional tables, or on the meaning of subcompositional coherence. A rich tasting of the above mentioned topics is contained in the presented Proceedings, reflecting various interests of CoDaWork 2017 participants.

In order to further propagate leading ideas of CoDaWork 2017, authors of selected papers will be asked after CoDaWork 2017 to submit an extended version of the paper to a special issue of the Austrian Journal of Statistics. The papers for this special issue will be peer-reviewed. The selection will be done by the CoDaWork Scientific Program Committee.

The proceedings papers reflect the highlights of CoDaWork 2017 and the compositional data analysis in general. Accordingly, the Scientific Committee has endeavored to provide a balanced and stimulating program that will appeal to the diverse interests of the participants. We keep to the tradition in the series of workshops on compositional data analysis of no parallel sessions, in order to enable a knowledge transfer as well as to stimulate a fruitful discussion among researchers from various fields of interest in the compositional data community.

To stir scientific community life between CoDaWorks, the CoDa-Association (www.codaa.org) was established during CoDaWork 2015 in L'Escala (Spain). The CoDa-Association is an organization which brings together scientists interested in developing methods for compositional data modelling and their applications. The Association is devoted primarily to support young researchers interested in compositional data analysis, as well as to the organisation of CoDaWork and

other CoDA-related scientific events. Its scope is by no means closed! Please, become an active member and help to strengthen the community spirit. Practitioners interested in compositional data analysis are also invited to visit the website www.compositionaldata.com, owned by the research group from University of Girona, where they will find a forum for the exchange of information, material and ideas.

We acknowledge the hard work of all organizers and the support of our hosts and sponsors, and particularly the Università degli Studi di Firenze (DST, Department of Earth Sciences), IAMG (International Association for Mathematical Geosciences), the Politecnico di Milano (MOX, Dept. of Mathematics), the Università di Milano Bicocca (DEMS, Dep. of Economics, Management and Statistics), the Università di Napoli, “L’Orientale” (Department of Human and Social Sciences), the Italian Geochemical Society (SoGeI), the Parco Nazionale Museo delle Miniere dell’Amiata, the Abbadia San Salvatore Council, and the Research Group on CoDa (Universitat de Girona, Dept. IMAE).

We wish you a productive, stimulating workshop and a memorable stay in Abbadia San Salvatore.

Abbadia San Salvatore, June 5, 2017

Karel Hron and Raimon Tolosana-Delgado
(Editors)

Publisher: CoDA, <http://www.coda-association.org/en/>

ISBN: 978-84-947240-0-8

Index

Bear, J. and D. Billheimer (ZER-1)	
<i>Zeros and Subcompositionally Coherent Estimators</i>	1
Blondes, M.S., W.H. Craddock, J.L. Shelton and M.A. Engle (GEO-2-3)	
<i>Characterization of Crustal Gas Systems using Compositional Data Analysis of Noble Gas Isotopes and Gas Compositions</i>	11
Coleman, S.Y. (SOC-2-1)	
<i>Analysing activities in a classroom – Remembrances of John Aitchison in Hong Kong with applications to a Service Provider</i>	21
Edjabou, V.M.E., J. A. Martín-Fernández, A. Boldrin and T.F. Astrup (P3-4)	
<i>Compositional data analysis of household waste recycling centres in Denmark</i>	27
Egozcue, J. J. and V. Pawlowsky-Glahn (GENER-1)	
<i>Compositional data: simple questions, difficult answers</i>	35
Engle, M.A., A. Buccianti, R. Olea and M.S. Blondes (GEO-2-2)	
<i>Merging key concepts in the chemistry of natural waters with compositional data analysis</i>	47
Erb, I. , T. Quinn, D. Lovell, and C. Notredame (P1-10)	
<i>Differential proportionality - an normalization-free to differential gene expression</i>	57
Graf, M. (STA-2-1)	
<i>A distribution on the simplex of the Generalized Beta type</i>	71
Gulban, O.F. and F. De Martino (P1-7)	
<i>Application of Aitchison metrics on magnetic resonance imaging data with multiple contrasts at ultra high field (7 Tesla) to investigate compositional characteristics of brain tissues in living humans</i>	91
Hingley, P. (TIM-2)	
<i>Forecasting patent filings at the European Patent Office (EPO) using compositional data analysis techniques</i>	97
Kenett, R.S., J.A. Martín-Fernández and M. Vives-Mestres (SOC-1-1)	
<i>Association rules and compositional data analysis: implications to big data</i>	107
LaRoche, D.D., Billheimer, D., Sinari, S., Michels, K. and LaFleur, B.J. (P1-5)	
<i>Quality control metrics for extraction-free targeted RNA-Seq: methods afforded by a compositional framework</i>	117
Liu, X.C, W.L. Wang, and Y.R. Pei (P3-5)	
<i>Compositional data analysis on the the stream sediment geochemical data at the Duolong mineral district, Tibet, China</i>	133
Monti, G.S. and S. Migliorati (ENV-2-3)	
<i>Compositional approach to the analysis of species abundance data</i>	141

Monti, G.S., G. Mateu-Figueras, M. I. Ortego, V. Pawlowsky-Glahn and J.J. Egozcue (P1-9)	
<i>Modified Multivariate Kolmogorov-Smirnov Test of Goodness of Fit</i>	152
Morais, J. , C. Thomas-Agnan and M. Simioni (SOC-1-2)	
<i>Interpreting the impact of explanatory variables in compositional models</i>	159
Ordóñez-Calderón, J.C. , S. Gelcich and J.F. Oliveira (GEO-1-2)	
<i>Applied Data Analytics on Multi-element Geochemistry for Pre-mining Characterization of Geological and Geometallurgical Attributes: Examples from the Rosemont Cu-Mo-Ag Skarn Deposit, Tucson, Arizona</i>	181
Parent, S.E. and L.E. Parent (ENV-1-1)	
<i>Balance designs revisit indices commonly used in agricultural science and eco-engineering</i>	195
Pawlowsky-Glahn, V. , J.J. Egozcue and M. Planes-Pedra (P3-3)	
<i>Survey data on perceptions of contraceptive measures as compositional tables</i>	229
Sinari, S., Dean Billheimer, and Edward J Bedrick (MAT-2)	
<i>Subcompositional coherence and the compositional complex</i>	239
Speranza, A. , R. Caggiano, S. Margiotta and V. Summa (ENV-1-3)	
<i>Compositional data analysis of element concentrations of simultaneous size segregated PM measurements</i>	251
Washburne, A. (OMI-1-3)	
<i>Phylofactorization - theory and challenges</i>	261
Wu, Jia R., Jean M. Macklaim, Briana L. Genge, and Gregory B. Gloor (P1-8)	
<i>Finding the centre: corrections for asymmetry in high-throughput sequencing datasets</i>	283
Ziembik, Z., A. Dolhańczuk-Śródka, and T. Majcherczyk (P3-7)	
<i>Analysis of gamma radioactive isotopes content in surface soil layers near Longyearbyen, Spitsbergen</i>	299

Zeros and Subcompositionally Coherent Estimators

John Bear¹and Dean Billheimer²

¹University of Arizona, Tucson, AZ, USA; *jbear@email.arizona.edu*

²Statistical Consulting Lab, University of Arizona, Tucson, AZ, USA

Abstract

Subcompositional coherence demands that we reach the same inferential conclusion about subcompositions under two different operations:

1. taking a subcomposition of the data, making calculations, and performing inference, or
 2. performing calculations on the full set of data, taking subcompositions of the results, and then performing inference.
- (Aitchison (1986), Aitchison and Egozcue (2005), Egozcue (2009), Egozcue and Pawlowsky-Glahn (2011).)

Some researchers have argued that subcompositional coherence might not be maintainable when modeling compositional data sets containing essential zeros, Butler and Glasbey (2008), Scealy and Welsh (2014). Others have argued that perhaps it is possible to get “close enough” to subcompositional coherence, Greenacre (2011).

We offer formal criteria for subcompositional coherence using functions of location and dispersion estimators. We show that a given statistical model can have both subcompositionally coherent and noncoherent estimators.

Note that we are extending the definition of subcompositional coherence from a property of a probability model to one of parameter estimators.

We illustrate with examples, including examples containing essential zeros. The benefit of these explicit criteria is that they make clearer the relationships between zeros, parameter estimators, and subcompositional coherence of models.

Key words: subcomposition, subcompositional coherence, zeros.

1 Subcompositional Coherence

One of the motivations for using the logistic normal distribution, mentioned repeatedly in the literature, is a property called *subcompositional coherence*. Although there are multiple definitions, the spirit of the idea is this, from Aitchison and Egozcue (2005),

“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.”

(1)

Aitchison writes about the importance of the property in Aitchison (1994), Aitchison (1999), Aitchison *et al.* (2002), and Aitchison and Egozcue (2005). Others have argued for it as well, Egozcue (2009), Egozcue and Pawlowsky-Glahn (2011), and others. There is controversy though. Butler and Glasbey (2008) claimed, “Any approach that attempts to describe zero and non-zero proportions by using a common model will inevitably break these principles [scale invariance and subcompositional coherence], because ratios of proportions are infinite along the boundaries of the simplex.” Scealy and Welsh (2014), also object: “We show that this Principle [subcompositional coherence] is based on implicit assumptions and beliefs that do not always hold. Moreover, it is applied selectively because it is not actually satisfied by the log-ratio methods it is intended to justify.”

We posit that part of the reason for the controversy is the lack of an operational definition, and we propose one. We take the statement in (1) to mean that when estimating the mean and covariance parameters of a subcomposition of a set of data, the answer should be the same whether one computes the estimates on the subcomposition, or computes the estimates based on the full data set, and then takes the appropriate subparts, corresponding to the subcomposition. We state this more formally in section (4), but first we illustrate the idea with some examples from an artificial data set.

We have two goals for this paper. One is to state this important property more precisely. The second is to show that it is not just a property of distributions, but of estimators.

2 Examples

Suppose we have compositional data on how much money 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.

	spices	lentils	rice
1	0.1598	0.0000	0.8402
2	0.1687	0.0000	0.8313
3	0.1576	0.0000	0.8424
4	0.0000	0.3726	0.6274
5	0.0000	0.3727	0.6273
6	0.0000	0.3747	0.6253
7	0.1166	0.3285	0.5549
8	0.1083	0.3350	0.5567

Table 1: Some artificial data

2.1 Naive Estimator, $\hat{\mu}$

This section illustrates an estimator which produces different results depending on order of operations, whether the subcomposition is done before or after calculation of the mean. We consider D-part compositions, and we use $d = D - 1$. Let $\mathbf{X}_{n \times D} = [x_{ti}]$ be a collection of n compositions, each row being a composition, possibly with zeros, with the D^{th} part always strictly positive. We start by defining a naive estimator $\hat{\mu} = [\hat{\mu}_i], i \in \{1, 2, \dots, D\}$.

$$\hat{\mu}_i = \frac{1}{n} \sum_{t=1}^n x_{ti}. \quad (2)$$

First we compute $\hat{\mu}$ for the full set: (spices: 0.089, lentils: 0.223, rice: 0.688). Take the subcomposition of $\hat{\mu}$ with the 1st and 3rd columns: (spices: 0.089, rice: 0.688), and renormalize: (spices: 0.114, rice: 0.886).

When we start by taking the {1,3} subcomposition of the data first, and then calculating the mean, the result changes. The raw {1,3} subset of the data is in Table (2), and the renormalized subcompositions are in Table (3). Computing $\hat{\mu}$ from the values in Table (3) the result is (spices:

	spices	rice		spices	rice
1	0.1598	0.8402		1	0.1598
2	0.1687	0.8313		2	0.1687
3	0.1576	0.8424		3	0.1576
4	0.0000	0.6274		4	0.0000
5	0.0000	0.6273		5	0.0000
6	0.0000	0.6253		6	0.0000
7	0.1166	0.5549		7	0.1736
8	0.1083	0.5567		8	0.1629

Table 2: Raw Subset

Table 3: Subcomposition (renormalized)

0.103, rice: 0.897). This value is different from the one above. For this mean estimator, the values are not invariant to the order of operation.

	spices	rice
mean first, subcomposition second	0.114	0.886
subcomposition first, mean second	0.103	0.897

Table 4: Result is not order invariant for $\hat{\mu}$

2.2 Simple ALR Estimator $\hat{\mu}^*$

Next we define a simple estimator of the mean, $\hat{\mu}^* = (\hat{\mu}_1^*, \hat{\mu}_2^*, \dots, \hat{\mu}_d^*)^T$, based on the additive logratio transformation. Let n_i be the number of elements of the i^{th} column of \mathbf{X} that are nonzero. For $i \in \{1, 2, \dots, d\}$, and $t \in \{1, 2, \dots, n\}$, define

$$\hat{\mu}_i^* = \frac{1}{n_i} \sum_{\{t: x_{ti} \neq 0\}} \log(x_{ti}/x_{tD}). \quad (3)$$

In the case where there are no zeros in the data, this estimator is just the usual maximum likelihood estimator for the mean of the additive logistic normal distribution.

In the spices-lentils-rice example, $\hat{\mu}^{*T} = (\log(\text{spices}/\text{rice}): -1.635, \log(\text{lentils}/\text{rice}): -0.523)$. For ease of interpretation, we convert the estimate back to a composition with the alr^{-1} transformation:

(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. (The alr^{-1} transformation is defined in the appendix.)

When we calculate $\hat{\mu}$ first, before taking a subcomposition, we get: $(\log(\text{spices}/\text{rice}))$: -1.626, $\log(\text{lentils}/\text{rice})$: -0.517). The subvector without lentils is $(\log(\text{spices}/\text{rice}))$: -1.626). Translating back to a composition with alr^{-1} gives (spices: 0.164, rice: 0.836).

If we take a subcomposition first, as in Table 3, and then compute $\hat{\mu}$, we get $(\log(\text{spices}/\text{rice}))$: -1.626). Translating back to a composition with alr^{-1} gives (spices: 0.164, rice: 0.836), which is the same as in the previous paragraph. The point is that $\hat{\mu}$ is invariant to the order of operations.

3 Covariance Estimators

Covariance estimators can also be, but need not be, invariant to the order of operations. As with the examples of mean estimators, we give an example of an estimator that is not invariant to order of operation, and one which is. For the estimator which is not invariant to order of operations, we use the examples of Aitchison (1986) (pp. 52-55). He defines a “crude covariance matrix,” $\hat{\mathbf{K}} = [\kappa_{ij} : i, j = 1, \dots, D]$. $\kappa_{ij} = \text{Cov}(x_i, x_j)$, $(i, j = 1, \dots, D; t = 1, \dots, N; n = N - 1)$. He defines the estimator in the way that is common for multivariate *noncompositional* data.

$$\hat{\mathbf{K}} = [\hat{\kappa}_{ij}]; \quad \hat{\kappa}_{ij} = n^{-1} \sum_{t=1}^N (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j), \quad (i, j = 1, \dots, D); \quad \bar{x}_i = N^{-1} \sum_{t=1}^N x_{ti}. \quad (4)$$

He shows that the order of operations matters. Taking a subcomposition of the data and then estimating $\hat{\mathbf{K}}$ gives a different result than estimating $\hat{\mathbf{K}}$ from the full data, and then projecting onto the subspace determined by the subcomposition. The next few tables (Tables 5, 6, 8) are covariances of subcompositions of Hongite (25×5), taken from Aitchison (1986), p. 55. The covariances are $10^{-3} \times$ the values in the tables. Tables 7 and 9 are the result of removing the relevant rows and columns from Table 5.

Table 5: Covariance of {A,B,C,D,E} Full Composition

	A	B	C	D	E
A	2.98	3.30	-5.98	-0.07	-0.23
B	3.30	13.99	-14.22	-1.38	-1.69
C	-5.98	-14.22	17.77	1.02	1.41
D	-0.07	-1.38	1.02	0.55	-0.13
E	-0.23	-1.69	1.41	-0.13	0.64

Table 6: Take subcomposition, then estimate, for {A,D,E}

	A	D	E
A	2.54	-1.15	-1.39
D	-1.15	1.46	-0.31
E	-1.39	0.31	1.70

Table 7: Estimate covariance, then take {A,D,E} subcomposition

	A	D	E
A	2.98	-0.07	-0.23
D	-0.07	0.55	-0.13
E	-0.23	-0.13	0.64

If we compare Table 6 with Table 7, we see they are quite different, and similarly for Table 8 and 9. For Tables 8 and 9, if we compare the (D,E) entries we see there is even a sign change. The point Aitchison made, is that for this estimator of covariance, finding the estimate, and then taking a subcomposition yields very different results than taking the subcomposition first, and then calculating the estimate.

Table 8: Take subcomposition then estimate, for {B,D,E}

	B	D	E
B	2.87	-1.48	-1.39
D	-1.48	1.02	0.45
E	-1.39	0.45	0.94

Table 9: Estimate covariance, then take {B,D,E} subcomposition

	B	D	E
B	13.99	-1.38	-1.69
D	-1.38	0.55	-0.13
E	-1.69	-0.13	0.64

3.1 Simple ALR Covariance Estimator, $\overset{*}{\Omega}$

Next we define a simple covariance estimator, $\overset{*}{\Omega}$, based on $\overset{*}{\mu}$, and show that, with some restrictions, it is order invariant.

Define $\overset{*}{\Omega}_{d \times d} = [\overset{*}{\sigma}_{ij}]$, where

$$\overset{*2}{\sigma}_{ii} = \frac{1}{n_i - 1} \sum_{\{t: x_{ti} \neq 0\}} (\log(x_{ti}/x_{tD}) - \overset{*}{\mu}_i)^2, \text{ and} \quad (5)$$

$$\overset{*}{\sigma}_{ij} = \frac{1}{n_{ij} - 1} \sum_{\{t: x_{ti} \neq 0 \& x_{tj} \neq 0\}} (\log(x_{ti}/x_{tD}) - \overset{*}{\mu}_i)(\log(x_{tj}/x_{tD}) - \overset{*}{\mu}_j). \quad (6)$$

t indexes the rows; i indexes the columns.

n_{ij} is the number of points where both x_{ti} and x_{tj} are not 0.

$\overset{*}{\sigma}_{ij}$ is based on n_{ij} pairs, but

$\overset{*}{\mu}_i$ and $\overset{*}{\mu}_j$ are based on n_i and n_j observations, respectively.

If there are no zeros in the data, this estimator reduces to the usual covariance estimator for the additive logistic normal distribution. This estimator, $\overset{*}{\Omega}_{d \times d}$, has some undesirable properties. For one, it is not guaranteed to be positive definite. For another, 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.

Both $\overset{*}{\mu}_i$ and $\overset{*2}{\sigma}_{ii}$ depend only on the i^{th} and D^{th} columns of the data. Covariances, $\overset{*}{\sigma}_{ij}$, depend only on the i^{th}, j^{th} , and D^{th} columns. These two facts together entail that we get the same result whether we estimate the entire covariance matrix first and then remove some rows and columns to get the covariance for the subcomposition, or take the subcomposition first, and then estimate the covariance matrix, *provided that the subcomposition contains the D^{th} (reference) column*.

4 Extending Coherence to Estimators

Here we offer an extension of the definition of *subcompositionally coherent* in the spirit of Aitchison and Egoozcue (2005), applied specifically to estimators. We designate a subcomposition with a set $W \subset \{1, 2, 3, \dots, D\}$ of indices. Without loss of generality we can order the indices from least to greatest:

$$W = \{j_1, j_2, \dots, j_J, j_{J+1} = D\} \text{ where } 0 < j_1 < j_2 < \dots < j_J < D. \quad (7)$$

Next we define two different selection matrices, \mathbf{B}_W based on the set W , and \mathcal{B}_W based on $(J+1) \times D$ based on $W \setminus \{D\} = \{j_1, j_2, \dots, j_J\}$.

For $p \in \{1, 2, \dots, J+1\}$, and $m \in \{1, 2, \dots, D\}$, with $W = \{j_1, j_2, \dots, j_J, j_{J+1}\}$,

we define the elements of $\mathbf{B}_W = [B_{p,m}]$ to be $B_{p,j_p} = 1$ and $B_{p,m \neq j_p} = 0$. (8)

For $p \in \{1, 2, \dots, J\}$, and $m \in \{1, 2, \dots, d\}$, with $W = \{j_1, j_2, \dots, j_J\}$,

we define the elements of $\mathcal{B}_W = [b_{p,m}]$ to be $b_{p,j_p} = 1$ and $b_{p,m \neq j_p} = 0$. (9)

Given these definitions, with $\mathbf{X}_{n \times D}$ a set of compositional data, $\mathbf{X}\mathbf{B}_W^T$ is a matrix where columns correspond to indices in W , and $\mathbf{X}\mathcal{B}_W^T$ is a matrix where columns correspond to indices in $W \setminus \{D\}$.

In extending the notion of subcompositional coherence, we address two cases: (1) the case where a reference component is crucial in the distribution, as with the additive logistic normal, and with approaches like those of Greenacre (2011), Leininger *et al.* (2013), and Stewart *et al.* (2014); (2) cases where there is no relevant reference component, as in the Dirichlet distribution, and the approaches of Butler and Glasbey (2008), and Scealy and Welsh (2014).

4.1 Case 1: With a reference component

Let $\text{Loc}_d(\cdot)$ be an estimator of the location parameter, and $\ddot{\mu}$ be its estimate (10), and let $\text{Var}_d(\cdot)$ be an estimator of the covariance parameter, and $\ddot{\Omega}$ be its estimate (11).

$$\begin{aligned} \text{Loc}_d : \mathcal{S}^{n \times d} &\rightarrow \mathbb{R}^{d \times 1} \\ \underbrace{\mathbf{X}_{n \times D}}_{J \times 1} &\mapsto \underbrace{\ddot{\mu}_{d \times 1}}_{J \times 1} \end{aligned} \quad (10)$$

$$\begin{aligned} \text{Var}_d : \mathcal{S}^{n \times d} &\rightarrow \mathbb{R}^{d \times d} \\ \underbrace{\mathbf{X}_{n \times D}}_{J \times J} &\mapsto \underbrace{\ddot{\Omega}_{d \times d}}_{J \times J} \end{aligned} \quad (11)$$

We define a location estimator, Loc_d , as *subcompositionally coherent* if Condition (12) holds. Condition (12) says that you get the same result whether you take subcomposition of the data first, and then estimate the mean, or do things in the other order, estimating the mean first, and then taking the subcomposition. The closure operator (\mathcal{C}) is defined in the appendix.

We define a variance estimator, Var_d , as *subcompositionally coherent* if Condition (13) holds. Condition (13) says that taking a subcomposition and then estimating the covariance gives the same result as estimating the full covariance, and then projecting onto the subspace of the subcomposition.

$$\underbrace{\text{Loc}_J(\mathcal{C}(\underbrace{\mathbf{X}_{n \times D} (\mathbf{B}_{D \times (J+1)}^T}))}_{J \times 1} = \underbrace{\mathcal{B}(\text{Loc}_d(\underbrace{\mathbf{X}_{n \times D}}_{J \times 1}))}_{J \times 1}. \quad (12)$$

$$\underbrace{\text{Var}_J(\mathcal{C}(\underbrace{\mathbf{X}_{n \times D} (\mathbf{B}_{D \times (J+1)}^T}))}_{J \times J} = \mathcal{B}(\underbrace{\text{Var}_d(\underbrace{\mathbf{X}_{n \times D}}_{J \times J})}_{d \times d}) \mathcal{B}^T. \quad (13)$$

We define *subcompositional coherence* as requiring that the conditions in Equations (12, 13) hold.

The simple mean estimator, $\ddot{\mu}^*$, is defined in such a way (Equation (3)) that the estimate $\ddot{\mu}_i^*$ is independent of the other $\ddot{\mu}_j^*, j \neq i$. Hence it satisfies Condition 12, whether or not there are zeros.

Similarly, our simple covariance estimator, $\ddot{\Omega}^*$, by the way it is defined in Equations (5) and (6), satisfies Condition (13). Both of these estimators have been investigated further in the context of

a model that is a mixture of logistic normal distributions, in Bear and Billheimer (2016). For that model, these estimators can be appropriate.

One fairly strong limitation of these constraints (12, 13) is that they require the Dth reference component to be strictly positive, and included in the subcomposition.

4.2 Case 2: No reference component

The conditions in Equations (12, 13) have selection matrices of two different dimensions. These are needed to be compatible with models which use one part of the composition as a reference component, like the logistic normal, or the approach by Leininger *et al.* (2013). For the approaches that do not use a reference component, like the Dirichlet, and the approaches by Scealy and Welsh (2014) and Butler and Glasbey (2008) the conditions for subcompositional coherence of estimators would be as in Equations (15) and (17). For this case we relax a constraint on W . Now we no longer require that $j_{(J+1)} = D$ when we construct the selection matrix, \mathbf{B} .

$$\begin{aligned} \text{Loc}_D : \mathcal{S}^{n \times D} &\rightarrow \mathbb{R}^{D \times 1} \\ \underbrace{\mathbf{X}_{n \times D}}_{(J+1) \times 1} &\mapsto \underbrace{\ddot{\boldsymbol{\mu}}_{D \times 1}}_{(J+1) \times 1} \end{aligned} \quad (14)$$

$$\underbrace{\text{Loc}_{(J+1)}(\mathcal{C}(\mathbf{X}_{n \times D} \mathbf{B}^T))}_{(J+1) \times 1} = \underbrace{\mathbf{B}_{(J+1) \times D} (\text{Loc}_D(\mathbf{X}_{n \times D}))}_{D \times 1}. \quad (15)$$

$$\begin{aligned} \text{Var}_D : \mathcal{S}^{n \times D} &\rightarrow \mathbb{R}^{D \times D} \\ \underbrace{\mathbf{X}_{n \times D}}_{(J+1) \times (J+1)} &\mapsto \underbrace{\ddot{\boldsymbol{\Omega}}_{D \times D}}_{(J+1) \times (J+1)} \end{aligned} \quad (16)$$

$$\underbrace{\text{Var}_{(J+1)}(\mathcal{C}(\mathbf{X}_{n \times D} \mathbf{B}^T))}_{(J+1) \times (J+1)} = \underbrace{\mathbf{B}_{(J+1) \times D} (\text{Var}_D(\mathbf{X}_{n \times D}))}_{(J+1) \times (J+1)} \mathbf{B}_{D \times (J+1)}^T. \quad (17)$$

5 Conclusion

The claim about subcompositional coherence in statement (1) has at its core the idea of inference, of two inferences being the same. In statistics, inferences are commonly the result of calculations based on estimates of means and variances. We have used this fact to recharacterize compositional coherence in terms of calculations of estimates of means and variances.

An analysis which uses subcompositionally coherent mean and covariance estimators for drawing inferences can be said to be subcompositionally coherent. The issue is not just about which probability model is used. The same probability model can have estimators which are coherent, and others which are not. The latent Gaussian model for compositions proposed by Butler and Glasbey (2008), and the model based on the hypersphere proposed by Scealy and Welsh (2014) have been conjectured not to be subcompositionally coherent. Now there is a way, in theory, to check.

In spite of the word, *incoherence*, in the title of the paper by Greenacre (2011), analyses using the power transformation he proposed might be subcompositionally coherent, if appropriate estimators are used, and now there is a way to check.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.
- Aitchison, J. (1994). *Principles of compositional data analysis*, pages 73–81. In Anderson *et al.* (1994).
- Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology*, **31**(5), 563–580.
- Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, **37**(7), 829–850.
- Aitchison, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2002). Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis. *Archaeometry*, (44), 295–304.
- Anderson, T. W., Olkin, I., and Fang, K., editors (1994). *Multivariate analysis and its applications*. Institute of Mathematical Statistics, Hayward, CA.
- Bear, J. and Billheimer, D. (2016). A logistic normal mixture model for compositional data allowing essential zeros. *Austrian Journal of Statistics*, **45**(4).
- Butler, A. and Glasbey, C. (2008). A latent Gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**(5), 505–520.
- Egozcue, J. (2009). Reply to “on the Harker variation diagrams;...” by J.A. Cortes. *Mathematical Geosciences*, **41**(7), 829–834.
- Egozcue, J. J. and Pawlowsky-Glahn, V. (2011). *Basic Concepts and Procedures*, pages 12–28. In Pawlowsky-Glahn and Buccianti (2011).
- Greenacre, M. (2011). Measuring subcompositional incoherence. *Mathematical Geosciences*, **43**, 681–693.
- Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander Jr, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**(3), 314–334.
- Pawlowsky-Glahn, V. and Buccianti, A., editors (2011). *Compositional Data Analysis: Theory and Applications*. Wiley.
- Scealy, J. and Welsh, A. (2014). Colours and cocktails: Compositional data analysis 2013 Lancaster lecture. *Australian & New Zealand Journal of Statistics*, **56**(2), 145–169.
- Stewart, C., Iverson, S., and Field, C. (2014). Testing for a change in diet using fatty acid signatures. *Environmental and Ecological Statistics*, **21**(4), 775–792.

Appendix

The inverse of the logratio transformation is the logistic transformation, alr^{-1} , and is defined as follows, with $D = d + 1$. If $\mathbf{y} = (y_1, y_2, \dots, y_d)$, then $\text{alr}^{-1}(\mathbf{y}) = \mathbf{x} = (x_1, x_2, \dots, x_D)$ where:

$$x_i = \exp(y_i) / \{\exp(y_1) + \dots + \exp(y_d) + 1\} \quad (i = 1, \dots, d), \quad (18)$$

$$x_D = 1 - x_1 - x_2 - \dots - x_d \quad (19)$$

$$= 1 / \{\exp(y_1) + \dots + \exp(y_d) + 1\}. \quad (20)$$

Closure

Definition: The *closure* operation, \mathcal{C} , transforms a basis vector into a composition. Let

$$\mathbf{z} = (z_1, z_2, \dots, z_J), \text{ where } z_i \geq 0, \text{ and } s = \sum_{i=1}^J z_i.$$

Then $\mathbf{x} = \mathcal{C}(\mathbf{z}) = \mathbf{z}/s$ is a composition. Sometimes we refer to closure as *renormalization*. We also use closure to renormalize an entire matrix of compositions. Where \mathbf{z}_t is the t^{th} row of matrix \mathbf{Z} ($t \in \{1, 2, \dots, n\}$),

$$\text{define } \mathcal{C}(\underset{n \times J}{\mathbf{Z}}) = \begin{bmatrix} \mathcal{C}(\mathbf{z}_1) \\ \mathcal{C}(\mathbf{z}_2) \\ \vdots \\ \mathcal{C}(\mathbf{z}_n) \end{bmatrix}. \quad (21)$$

Characterization of Crustal Gas Systems using Compositional Data Analysis of Noble Gas Isotopes and Gas Compositions

M.S. Blondes^{1,2}, W.H. Craddock¹, J.L. Shelton¹, and M.A. Engle^{1,3}

¹U.S. Geological Survey, Reston, VA 20192, USA

²Email: mblondes@usgs.gov

³University of Texas at El Paso, El Paso, TX 79968, USA

Abstract

Noble gas isotopes and bulk gas concentrations are regularly used as tracers of fluid flow in Earth's crust. Common interpretive approaches include multiple plots of isotopic ratios (e.g. $^3\text{He}/^4\text{He}$, $^{20}\text{Ne}/^{22}\text{Ne}$, $^{21}\text{Ne}/^{22}\text{Ne}$, $^{40}\text{Ar}/^{36}\text{Ar}$) and gas compositions (e.g. CH₄ and other hydrocarbons, CO₂, He, Ar, Ne, N₂, O₂, H₂, and H₂S) to constrain the source and movement of the gases. The first order goal is often to understand the proportions of mantle, crustal, and atmospheric sources, so that more detailed geochemical processes can be interpreted. However, it would be helpful to have a simple tool to check whether all plots in a given study are internally consistent. Here we use a multivariate and compositional data analysis (CoDa) approach to examine all relevant noble gas isotopes and gas compositions in one analysis as a means to more clearly interpret mantle, crustal, and atmospheric gas sources in crustal gas systems, and to check for internal consistency in noble gas plots. Specifically, we present signature isometric log ratio (ilr) balances that can be used to differentiate between these sources.

Key words: Noble gas isotopes, CoDa, natural gas, ilr coordinates

1 Introduction

Paired noble gas isotopes and bulk gas concentrations are regularly used as tracers of fluid flow in Earth's crust (e.g. Burnard, 2013; Ballentine and Burnard, 2002; Ballentine et al., 2002). Noble gases (He, Ne, Ar, Kr, Xe) are ideal conservative tracers because they are chemically inert, yet their different physical properties (e.g. solubility and diffusivity) make the proportions between noble gases sensitive to physical processes. The ratios of noble gas isotopes are highly variable due to primordial partitioning of elements between terrestrial reservoirs (mantle, crust, and atmosphere), the subsequent radiogenic or nucleogenic production of certain isotopes, and any later stage mixing between reservoirs. Variations in noble gas isotope ratios have been used to interpret oil and gas accumulations (Ballentine et al., 1991; Prinzhofner, 2013), mantle source reservoirs (Graham, 2002), and the suitability of saline reservoirs for CO₂ storage (Gilfillan et al., 2008; Holland and Gilfillan, 2013) among many other applications.

When examining crustal gas systems, noble gas isotope ratios (typically ³He/⁴He, ²⁰Ne/²²Ne, ²¹Ne/²²Ne, ⁴⁰Ar/³⁶Ar and sometimes Kr and Xe isotopic ratios) are used independently, or else in combination with concentrations of major gas components (typically CH₄ and other hydrocarbons, CO₂, He, Ar, Ne, N₂, O₂, H₂, and H₂S) to constrain the source and movement of the gases. Standard data analysis techniques usually include identifying trends in multiple plots of isotope ratio vs. isotope ratio, or isotope ratio vs. gas concentration ratio. Each plot represents a different incoherent subcomposition, but each plot also tries to answer the same general question: What are the proportions of mantle, crustal, and atmospheric sources in the crustal gas accumulation? In the traditional approach multiple plots are needed, which make comparisons between subcompositions difficult and cumbersome.

Craddock et al. (in revision) used principal component analysis of the four most common noble gas isotope ratios (³He/⁴He, ²⁰Ne/²²Ne, ²¹Ne/²²Ne, ⁴⁰Ar/³⁶Ar) to show that, at least for high CO₂ and high N₂ + He reservoirs in the U.S. Colorado Plateau and Rocky Mountain Provinces, variation can be simply defined by one principal component that represents crust vs. mantle and a secondary one that represents the incorporation of air or air-saturated-water (ASW). Here we provide a multivariate and compositional data analysis (CoDa) approach (e.g. Aitchison, 1986), building upon previous work combining isotopes and concentrations (Tolosana-Delgado et al., 2005; Puig et al., 2011; Blondes et al., 2016), as a check on the internal consistency of noble gas isotope and gas composition plots, particularly to interpret mixing between mantle, crustal, and atmospheric sources of gas in the crust.

2 Data

Noble gas isotope ratio and major gas composition data were compiled from published reports containing data for hydrocarbon, volcanic, high CO₂, geothermal, and groundwater reservoirs. Only data that had the full specific suite of gas concentrations (hydrocarbons, CO₂, and N₂), noble gas isotope ratios (³He/⁴He, ²⁰Ne/²²Ne, ²¹Ne/²²Ne, and ⁴⁰Ar/³⁶Ar), and the concentration of at least one noble gas isotope from each ratio (⁴He, ²⁰Ne, ⁴⁰Ar) were included in this analysis. Further, only fields or volcanoes with at least two points were included in order to better differentiate small-scale process-related variation with regional variation. Methane (CH₄) and any measured higher hydrocarbons were summed to calculate a total hydrocarbon percentage of the gas. All gas concentrations were converted to volume fractions. Noble gas isotope ratios and noble gas isotope concentrations were used to determine the volume fraction of each isotope part. The remaining analysis was done using ten volume fractions as parts: ³He, ⁴He, ²⁰Ne, ²¹Ne, ²²Ne, ³⁹Ar, ⁴⁰Ar, N₂, CO₂, and hydrocarbons. This data subset includes high CO₂ reservoirs on the U.S. Colorado Plateau and Rocky Mountain Provinces (Gilfillan et al., 2008; Shelton et al., 2016); conventional hydrocarbon reservoirs in Italy (Elliot et al., 1993), Hungary (Ballentine et al., 1991), and Poland (Kotarba et al., 2014); a shale gas reservoir in the U.S. Michigan Basin (Wen et al., 2015); and hot springs from volcanoes in West Africa (Aka et al., 2001) and Antarctica (Kusakabe et al., 2009). All

Blondes et al.

this is page 3

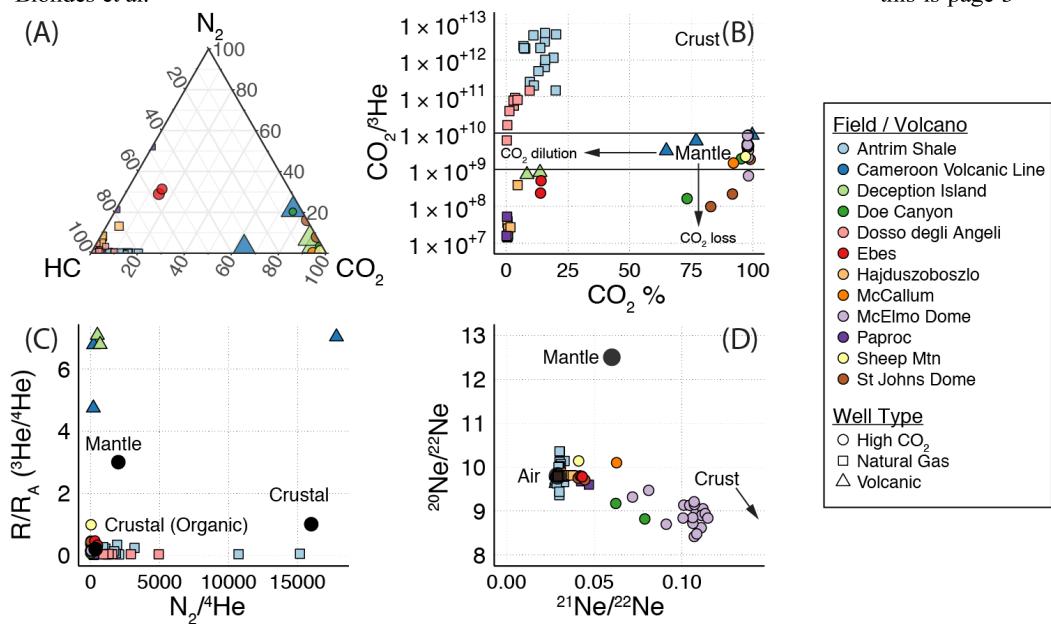


Figure 1: Traditional Noble gas and gas composition plots. Colors represent different gas fields or volcanoes, and shapes represent well types. Large black circles are endmember compositions. Plots created using R packages ggplot2 (Wickham et al., 2016a), ggtern (Hamilton, 2016), and cowplot (Wilke and Wickham, 2016). (A) Ternary diagram showing N₂, CO₂, and total hydrocarbon (HC) percentages, after Giggenbach (1993); (B) CO₂/³He vs. CO₂ %, after Gilfillan et al. (2008). The horizontal lines represent the range of mantle-sourced CO₂/³He; (C) R/R_A vs. N₂/⁴He. R/R_A is the ³He/⁴He ratio normalized to that of air. Endmembers are from Jenden et al. (1988); (D) Triple Neon plot. Endmembers are from Prinzhofer et al. (2013).

calculations were made using R (R Core Team, 2016) in RStudio (RStudio Team, 2016) using the package dplyr (Wickham et al., 2016b).

3 Traditional Methods

Figure 1 shows examples of plots that are used in noble gas isotope studies. The first is a ternary diagram (Figure 1A) showing that the subset of data used has two dominant clusters: one high CO₂ cluster that contains both the volcanic samples and the high CO₂ gas reservoirs, and one hydrocarbon cluster that contains most of the natural gas samples and a few high CO₂ wells. Figure 1B is a CO₂/³He vs CO₂ % plot and is typically used to define magmatic CO₂ and regions of CO₂ loss, CO₂ dilution, and crustal CO₂. Figure 1C is a R/R_A (defined as ³He/⁴He in the sample versus the same ratio in the atmosphere) vs. N₂/⁴He plot that is used to discern sources of nitrogen in the crust (Jenden et al., 1988). Figure 1D is a “triple neon” plot, often used to differentiate between crust, mantle, and air sources of subsurface gases. Certain interpretations are consistent between these plots. For example, both the Antrim Shale gases from the Michigan Basin and the Dosso degli Angeli gases from the Po Basin indicate extensive input from ASW (Figures 1B, 1C, and 1D). However, depending on the plot, the McElmo Dome high CO₂ fields may appear to have a mantle (Figure 1B) or a more crustal (Figure 1D) signature. As a check on the internal consistency of these interpretations, it would be ideal to examine all gas compositions and noble gas isotope ratios in a single analysis.

4 A multivariate approach using clr-biplots

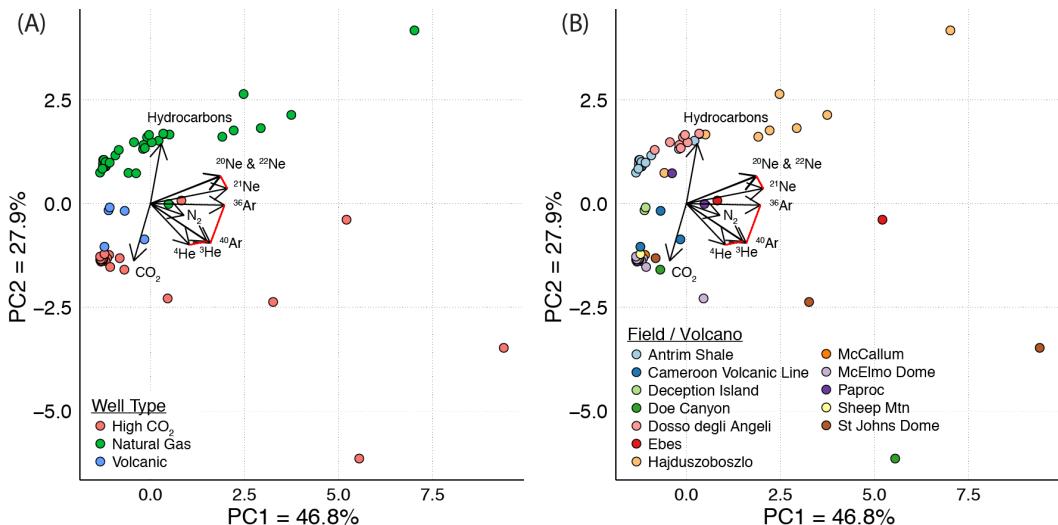


Figure 2: clr-biplots: points represent samples and rays represent the compositional parts. Links, in red, are shown between noble gas isotope pairs. Plots created using R packages ggplot2 (Wickham et al., 2016a) and gridExtra (Auguie and Antonov, 2016). (A) Data color-coded by Well Type; (B) Data color-coded by Field or Volcano.

One approach to check for internal consistency is the centered log ratio (clr) biplot (e.g. Aitchison and Greenacre, 2002). Figure 2 shows clr biplots for the crustal gas data subset. A clr biplot differs from a traditional biplot in that the length of the links between the ray end points approximate the relative log ratio between those two variables. The links between isotope pairs are highlighted in red. Since the links represent the log ratio variance between two parts, these links symbolize variation in the noble gas ratios for the data set. The first principal component is dominated by variation between hydrocarbons + CO₂ and N₂ + the noble gases. This is consistent with the N₂-He-Ar gas association seen in gas reservoirs on the Colorado Plateau (Craddock et al., in revision). The second principal component is dominated by variation between hydrocarbons and CO₂. There are two evident trends in the data. The first is a near linear trend of natural gas wells toward the top of Figure 2A. The second is a trend of the high CO₂ wells toward the bottom of Figure 2A. The volcanic wells plot in between these two trends. Though these trends exist, it is difficult to make interpretations with the noble gas isotopes using this type of plot. Since the log ratio variance between two noble gas isotopes is so small relative to log ratio variances between the major gas components, the links are small and provide little information (e.g. Blondes et al., 2016; Tolosana-Delgado et al., 2005).

5 An isometric log ratio approach for sourcing crustal gases

A better approach for noble gas isotopes and gas compositions may be to develop signature isometric log ratio (ilr) coordinates that represent air vs. crust vs. mantle contributions. The ilr transformation (Egozcue et al., 2003) not only removes the closure constraint for compositional data and has subcompositional coherence (whereby interpretations based on only some of the compositions should not conflict with those based on the entire data set), but creates orthonormal axes where distances are conserved (allowing proper linear regressions or cluster analysis, for example) and that may be more geologically intuitive.

In order to create ilr coordinates that represent air vs. crust vs. mantle contributions, it is necessary to make informed decisions about the sources of the different isotope and gas parts. Figure 3 shows a schematic Venn diagram of the three main source components of geologic gases. These are used to

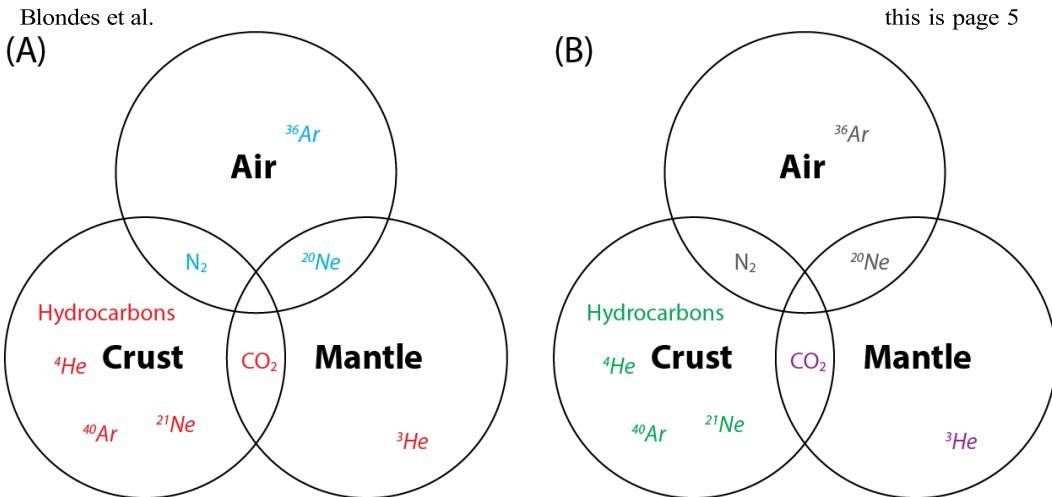


Figure 3: Schematic Venn diagram showing choice of parts for the sequential binary partition (Table 1) used to determine the orthonormal coordinates defined in Equations 2 and 3, and shown in Figure 4. Major gas composition parts are shown as normal text.

Isotope parts are shown in italics. (A) Schematic to represent z_1 (Eq. 2) the Air vs. Not Air axis in Figure 4. Numerator parts are cyan. Denominator parts are red; (B) Schematic to represent z_2 (Eq. 3) the Crust vs. Mantle axis in Figure 4. Numerator parts are green. Denominator parts are purple. Unused parts are grey.

help define a sequential binary partition to create orthonormal coordinates. When present in large quantities, some gases or gas isotopes are uniquely indicative of gas sourcing from the mantle, crust, or air (Ballentine et al., 2002; Holland and Gilfillan, 2013). Gases with unique source reservoirs include ^3He (mantle); ^4He , ^{21}Ne , ^{40}Ar , hydrocarbons (crust); and ^{36}Ar (air). Other gases (e.g. CO_2 , N_2 , ^{20}Ne) may be sourced from multiple reservoirs, and the importance of different reservoirs may change across tectonic settings. For example, CO_2 has a dominantly mantle source in volcanic settings or regions with a thermally activated crustal gas charge, yet may be dominantly crustal in regions where the thermally activated gas charge derives from burial heating of organic material. N_2 is the dominant component in air, yet it is also derived from clays and organic-rich sediments in the crust (Krooss et al., 1995). For the purposes of creating a plot that represents as many tectonic settings as possible, CO_2 was assumed to be a mantle gas and both N_2 and ^{20}Ne was assumed to be a gas prevalent in air. Note that because ^{22}Ne does have a distinct source, it is not included in the analysis. Thus we start with a $D = 9$ -part composition, which has $D - 1$ balances (Table 1). Each coordinate, z_i , is defined by the balance (adapted from Egozcue et al., 2005):

$$z_i = \sqrt{\left(\frac{r_i \times s_i}{r_i + s_i}\right)} \ln \frac{g(b_{r_i})}{g(b_{s_i})} \quad \text{for } i = 1 \dots D - 1 \quad (1)$$

where r_i is the number of +1 parts in balance b_i , $g(b_{r_i})$ is the geometric mean of the +1 parts of balance b_i , s_i is the number of -1 parts in balance b_i , and $g(b_{s_i})$ is the geometric mean of the -1 parts of balance b_i . Table 1 shows the balances that correspond to Figure 3. The ilr coordinates that correspond to Table 1 and Figure 3 are therefore:

$$z_1 = \sqrt{\left(\frac{3 \times 6}{3 + 6}\right)} \ln \frac{\left([^{36}\text{Ar}][^{20}\text{Ne}][\text{N}_2]\right)^{1/3}}{\left([^{3}\text{He}][^{4}\text{He}][^{21}\text{Ne}][^{40}\text{Ar}][\text{CO}_2][\text{HC}]\right)^{1/6}} \quad (2)$$

$$z_2 = \sqrt{\left(\frac{4 \times 2}{4 + 2}\right)} \ln \frac{\left([^{4}\text{He}][^{21}\text{Ne}][^{40}\text{Ar}][\text{HC}]\right)^{1/4}}{\left([^{3}\text{He}][\text{CO}_2]\right)^{1/2}} \quad (3)$$

Table 1: ilr Balances used to create coordinates (Equations 2 and 3) for Figure 4. Note that because only the first two coordinates are plotted, the remaining balances are not shown. All colors represent source components defined in Figure 3.

Balance	${}^3\text{He}$	${}^4\text{He}$	${}^{20}\text{Ne}$	${}^{21}\text{Ne}$	${}^{36}\text{Ar}$	${}^{40}\text{Ar}$	CO_2	N_2	HC
b_1	-1	-1	+1	-1	+1	-1	-1	+1	-1
b_2	-1	+1	0	+1	0	+1	-1	0	+1
...									
b_{D-1}									

Figure 4 shows a plot of the first ilr coordinate, z_1 , on the y-axis and the second coordinate z_2 , on the x-axis. The x-axis represents the crust vs. mantle gas component and the y-axis represents mixing with air or ASW. There are two dominant trends. The first is an apparent mantle trend. This trend includes all volcanic samples as well as many of the high CO_2 samples. An apparent crustal trend is dominated by the natural gas samples, as well has some high CO_2 samples. Both range from minimal air incorporation (bottom left of plot) to significant air incorporation (top of plot). This conforms to what we know about the specific geologic settings of the samples. The Cameroon Volcanic Line and Deception Island are both regions of mafic (and mantle-derived) volcanism (Aka et al., 2001; Kusakabe et al., 2009). The gas samples, however, were collected from hot springs and near surface waters not directly from the volcano, so we would expect significant mixing with air saturated water. All of the gases sampled from hydrocarbon reservoirs (Antrim Shale, Dosso degli Angeli, Hajduszoboszlo, Paproc) plot on the crustal trend, which is not surprising given that these are dominantly methane, a gas derived from sedimentary organic material in the crust (Elliot et al., 1993; Ballantine et al., 1991; Kotarba et al., 2014; Wen et al., 2015). It is also quite reasonable that these reservoirs have a significant air component because recharge fluids are known to circulate to their depths. The high CO_2 wells span both trends. The fields from the Colorado Plateau with abundant samples (McElmo Dome and St. John’s Dome) fall mostly along the mantle trend; any

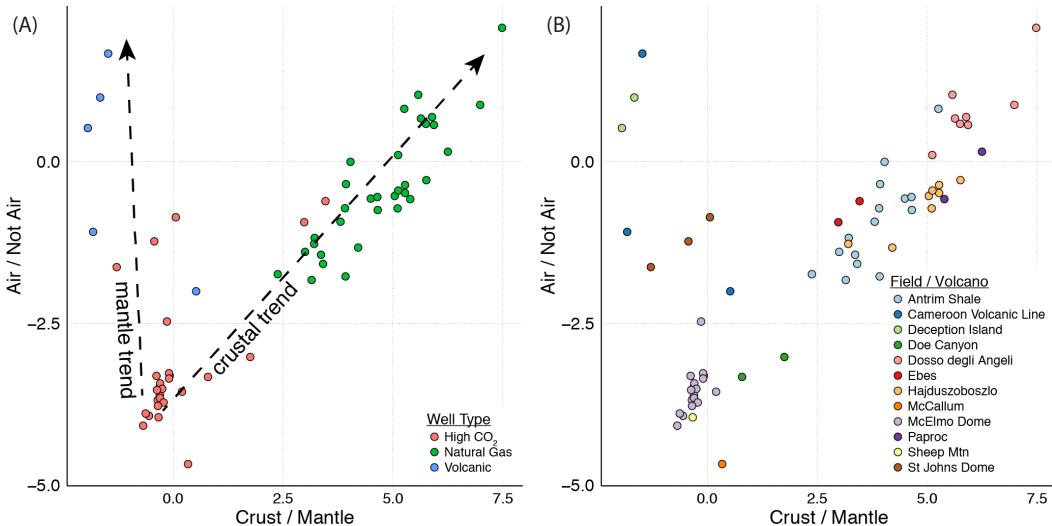


Figure 4: Air / Not Air vs. Crust / Mantle, using ilr coordinates. Air / Not Air and Crust vs. Mantle are represented by ilr coordinates z_1 and z_2 , respectively. The axes are defined by the schematic in Figure 3, the balances in Table 1, and the coordinates in Equations 2 and 3.

Dashed black arrows represent a mantle gas trend and a crustal gas trend, both showing variable mixing with air. Plots created using R packages ggplot2 (Wickham et al., 2016a) and gridExtra (Auguie and Antonov, 2016). (A) Samples are color-coded by Well Type; (B) Samples are color-coded by Field or Volcano.

Blondes et al.

this is page 7

crustal input would be minor. The Colorado Plateau and Rocky Mountain Provinces have extensive mantle-derived regional volcanism and high heat flow, which is the dominant source for the CO₂ in these wells (e.g. Gilfillan et al., 2008; Shelton et al., 2016; Craddock et al., in revision). St John's Dome samples have had more interaction with air saturated water than McElmo Dome samples. Doe Canyon appears to have more of a crustal source than the other Colorado Plateau and Rocky Mountain Province wells. Ebels wells are the one instance of high CO₂ wells clearly on the crustal trend. Unlike the Colorado Plateau and Rocky Mountain Province reservoirs, these are from hydrocarbon reservoirs in the Pannonian Basin (Ballentine et al., 2001), and therefore the CO₂ is likely derived from burial heating of organic material in the crust.

The trends in Figure 4 rely heavily on our assumptions of the sources of different gases (Figure 3). However, different tectonic settings will likely have different diagnostic gas sources. Ideally, a future approach would be to develop gas composition and isotopic data across a variety of tectonic settings, including areas of Cenozoic magmatism, craton interiors, passive margins, etc. In each of these settings we would develop a comprehensive picture of gas sources (particularly non-hydrocarbon gases), and associated ilr coordinates like in Figure 4 to use for interpretation and to check for internal consistency of noble gas isotopes and gas compositions. This could also be expanded to include more sources than crust vs. mantle vs. gas, such as specific biologic processes that create geochemical and isotopic variability.

6 Conclusions

Noble gas isotopes and gas compositions can be combined using CoDa techniques to check for internal consistency of various traditional plots in studies of crustal gases. We show that it is possible to develop signature ilr coordinates that can differentiate between mantle, crust, and air sources, and show distinctive mantle and crustal trends.

Acknowledgements

We greatly appreciate funding from the U.S. Geological Survey Energy Resources Program, as well as reviews by Ricardo Olea and Matthew Merrill. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

- Aitchison, John. 1986. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.
- Aitchison, John, and Michael Greenacre. 2002. “Biplots of Compositional Data.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51 (4): 375–92. doi:10.1111/1467-9876.00275.
- Aka, F. T., M. Kusakabe, K. Nagao, and G. Tanyileke. 2001. “Noble Gas Isotopic Compositions and Water/Gas Chemistry of Soda Springs from the Islands of Bioko, São Tomé and Annobon, along with Cameroon Volcanic Line, West Africa.” *Applied Geochemistry* 16 (3): 323–338.
- Auguie, Baptiste, and Anton Antonov. 2016. *gridExtra: Miscellaneous Functions for “Grid” Graphics* (version 2.2.1). <https://cran.r-project.org/web/packages/gridExtra/index.html>.
- Ballentine, C. J., R. K. O’Nions, E. R. Oxburgh, F. Horvath, and J. Deak. 1991. “Rare Gas Constraints on Hydrocarbon Accumulation, Crustal Degassing and Groundwater Flow in the Pannonian Basin.” *Earth and Planetary Science Letters* 105 (1–3): 229–246.
- Ballentine, Chris J., Ray Burgess, and Bernard Marty. 2002. “Tracing Fluid Origin, Transport and

- Blondes et al. this is page 8
 Interaction in the Crust.” *Reviews in Mineralogy and Geochemistry* 47 (1): 539–614.
 doi:10.2138/rmg.2002.47.13.
- Ballentine, Chris J., and Pete G. Burnard. 2002. “Production, Release and Transport of Noble Gases in the Continental Crust.” *Reviews in Mineralogy and Geochemistry* 47 (1): 481–538.
 doi:10.2138/rmg.2002.47.12.
- Blondes, M. S., M. A. Engle, and N. J. Geboy. 2016. “A Practical Guide to the Use of Major Elements, Trace Elements, and Isotopes in Compositional Data Analysis: Applications for Deep Formation Brine Geochemistry.” In *Compositional Data Analysis: CoDaWork 2015.*, edited by Josep Antoni Martín-Fernández and Santiago Thió-Henestrosa, 187:13–29. Springer Proceedings in Mathematics & Statistics. http://link.springer.com/chapter/10.1007/978-3-319-44811-4_2.
- Burnard, Pete, ed. 2013. *The Noble Gases as Geochemical Tracers*. Advances in Isotope Geochemistry. Berlin Heidelberg: Springer-Verlag.
<http://link.springer.com/content/pdf/10.1007/978-3-642-28836-4.pdf>.
- Craddock, William H., Madalyn S. Blondes, Christina A. DeVera, and Andrew G. Hunt. in revision. “Mantle and Crustal Gases of the Colorado Plateau: Geochemistry, Sources, and Migration Pathways.” *Geochimica et Cosmochimica Acta*.
- Egozcue, J. J., and V. Pawlowsky-Glahn. 2005. “Groups of Parts and Their Balances in Compositional Data Analysis.” *Mathematical Geology* 37 (7): 795–828. doi:10.1007/s11004-005-7381-9.
- Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. 2003. “Isometric Logratio Transformations for Compositional Data Analysis.” *Mathematical Geology* 35 (3): 279–300. doi:10.1023/A:1023818214614.
- Elliot, Trevor, C. J. Ballentine, R. K. O’nions, and T. Ricchiuto. 1993. “Carbon, Helium, Neon and Argon Isotopes in a Po Basin (Northern Italy) Natural Gas Field.” *Chemical Geology* 106 (3–4): 429–440.
- Giggenbach, W. F, Y Sano, and H Wakita. 1993. “Isotopic Composition of Helium, and CO₂ and CH₄ Contents in Gases Produced along the New Zealand Part of a Convergent Plate Boundary.” *Geochimica et Cosmochimica Acta* 57 (14): 3427–55. doi:10.1016/0016-7037(93)90549-C.
- Gilfillan, Stuart MV, Chris J. Ballentine, Greg Holland, Dave Blagburn, Barbara Sherwood Lollar, Scott Stevens, Martin Schoell, and Martin Cassidy. 2008. “The Noble Gas Geochemistry of Natural CO₂ Gas Reservoirs from the Colorado Plateau and Rocky Mountain Provinces, USA.” *Geochimica et Cosmochimica Acta* 72 (4): 1174–1198.
- Graham, David W. 2002. “Noble Gas Isotope Geochemistry of Mid-Ocean Ridge and Ocean Island Basalts: Characterization of Mantle Source Reservoirs.” *Reviews in Mineralogy and Geochemistry* 47 (1): 247–317. doi:10.2138/rmg.2002.47.8.
- Hamilton, Nicholas. 2016. *ggtern: An Extension to “ggplot2”, for the Creation of Ternary Diagrams* (version 2.2.0). <https://cran.r-project.org/web/packages/ggtern/index.html>.
- Holland, Greg, and Stuart Gilfillan. 2013. “Application of Noble Gases to the Viability of CO₂ Storage.” In *The Noble Gases as Geochemical Tracers*, edited by P. Burnard, 177–223. Advances in Isotope Geochemistry. Berlin Heidelberg: Springer-Verlag.
http://link.springer.com/chapter/10.1007/978-3-642-28836-4_8.
- Jenden, P. D., I. R. Kaplan, R. Poreda, and H. Craig. 1988. “Origin of Nitrogen-Rich Natural Gases in the California Great Valley: Evidence from Helium, Carbon and Nitrogen Isotope Ratios.” *Geochimica et Cosmochimica Acta* 52 (4): 851–861.
- Kotarba, Maciej J., Keisuke Nagao, and Paweł H. Karnkowski. 2014. “Origin of Gaseous Hydrocarbons, Noble Gases, Carbon Dioxide and Nitrogen in Carboniferous and Permian Strata of the Distal Part of the Polish Basin: Geological and Isotopic Approach.” *Chemical Geology* 383: 164–179.

- Blondes et al. this is page 9
- Krooss, B. M., R. Littke, B. Müller, J. Frielingsdorf, K. Schwochau, and E. F. Idiz. 1995. “Generation of Nitrogen and Methane from Sedimentary Organic Matter: Implications on the Dynamics of Natural Gas Accumulations.” *Chemical Geology, Processes of Natural Gas Formation*, 126 (3): 291–318. doi:10.1016/0009-2541(95)00124-7.
- Kusakabe, Minoru, Keisuke Nagao, Takeshi Ohba, Jung Hun Seo, Sung-Hyun Park, Jong Ik Lee, and Byong-Kwon Park. 2009. “Noble Gas and Stable Isotope Geochemistry of Thermal Fluids from Deception Island, Antarctica.” *Antarctic Science* 21 (3): 255–267.
- Prinzhofer, Alain. 2013. “Noble Gases in Oil and Gas Accumulations.” In *The Noble Gases as Geochemical Tracers*, edited by P. Burnard, 225–47. Advances in Isotope Geochemistry. Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-28836-4_9.
- Puig, Roger, Raimon Tolosana-Delgado, Neus Otero, and Albert Folch. 2011. “Combining Isotopic and Compositional Data: A Discrimination of Regions Prone to Nitrate Pollution.” In *Compositional Data Analysis: Theory and Applications*, edited by V. Pawlowsky-Glahn and A. Buccianti, 302–317. John Wiley & Sons.
- R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- RStudio Team. 2016. *RStudio: Integrated Development Environment for R* (version 1.0.136). Boston, MA: RStudio, Inc. <http://www.rstudio.com>.
- Shelton, Jenna L., Jennifer C. McIntosh, Andrew G. Hunt, Thomas L. Beebe, Andrew D. Parker, Peter D. Warwick, Ronald M. Drake II, and John E. McCray. 2016. “Determining CO₂ Storage Potential during Miscible CO₂ Enhanced Oil Recovery: Noble Gas and Stable Isotope Tracers.” *International Journal of Greenhouse Gas Control* 51 (August): 239–53. doi:10.1016/j.ijggc.2016.05.008.
- Tolosana-Delgado, Raimon, N. Otero, and A. Soler. 2005. “A Compositional Approach to Stable Isotope Data Analysis.” In *Proceedings of CoDaWork '05*, edited by G. Mateu-Figueras and Barceló-Vidal. University of Girona, Spain.
- Wen, Tao, M. Clara Castro, Brian R. Ellis, Chris M. Hall, and Kyger C. Lohmann. 2015. “Assessing Compositional Variability and Migration of Natural Gas in the Antrim Shale in the Michigan Basin Using Noble Gas Geochemistry.” *Chemical Geology* 417: 356–70. doi:10.1016/j.chemgeo.2015.10.029.
- Wickham, Hadley, Winston Chang, and RStudio. 2016. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (version 2.2.1). <https://cran.r-project.org/web/packages/ggplot2/index.html>.
- Wickham, Hadley, Romain Francois, and RStudio. 2016. *dplyr: A Grammar of Data Manipulation* (version 0.5.0). <https://cran.r-project.org/web/packages/dplyr/index.html>.
- Wilke, Claus O., and Hadley Wickham. 2016. *cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2”* (version 0.7.0). <https://cran.r-project.org/web/packages/cowplot/index.html>.

Analysing activities in a classroom – Remembrances of Professor John Aitchison in Hong Kong with applications to a Service Provider

S. Y. Coleman

Industrial Statistics Research Unit, School of Mathematics and Statistics,
Newcastle University, Newcastle upon Tyne, UK; Shirley.coleman@ncl.ac.uk

Abstract

Compositional data analysis formed a main focus of statistical activities at Hong Kong University when Professor John Aitchison was head of the Statistics department. It was part of a new Master's degree in Statistics that he set up, and as this was the first such post graduate degree to be offered in Hong Kong, it attracted many gifted statisticians from the Government Statistical Service and other employments making it a very lively program.

Acknowledging the constrained nature of many types of data led to a new way of looking at proportions and percentages of components making up data items. Professor Aitchison's seminal book *The statistical analysis of compositional data* contains background theory and many examples of data arising from a wide variety of applications such as geology, economics and human behaviour. One example was an analysis of the daily activities of a statistician. This prompted an analysis of classroom activities in a range of classes and schools encountered during teacher training within the Professional Educational Studies department of Hong Kong University. It was found that the nature of the target class and the school level affected the pattern of lesson activities with more listening carried out in the higher target classes and higher level schools. More time was spent dealing with educational equipment in lower level schools.

Data analytics is increasingly popular in all walks of life and many small and medium enterprises are realising the benefits. Compositional data forms a large part of internal company operational data and its analysis can provide useful insight. For example, the changes in proportions of different activities undertaken over time are important information for a service provider. Using ternary diagrams to illustrate proportions is an informative way to share the findings with company staff.

Key words: business analytics, inspirational head, statistical consultancy.

1 Introduction

Staff and students alike were delighted to embrace Professor Aitchison's erudite and stately leadership of the Statistics department at Hong Kong University from the time he started work there in 1976 to his retirement in 1989. His seminal book *The statistical analysis of compositional data* contained background theory and many examples of data arising from applications including geology, economics and human behaviour encountered during his consulting projects with a wide range of colleagues.

Compositional data analysis formed a main focus of statistical activities at the University. It was part of a new Master's degree in Statistics that Professor Aitchison set up, and as this was the first such post graduate degree to be offered in Hong Kong, it attracted many gifted statisticians from the Government Statistical Service and other employments making it a very lively program.

Acknowledging the constrained nature of much commonly encountered data led to a new way of looking at proportions and percentages of components making up data items. One example in Professor Aitchison's book was an analysis of daily activities of statisticians. This prompted awareness of possible applications in many other activities.

This paper applauds Professor Aitchison's work and in the next section gives some examples of compositional data analysis and describes an analysis of classroom activities carried out in Hong Kong. Data analytics is increasingly popular in all walks of life and many small and medium enterprises are realising the benefits (Coleman, 2016). Compositional data forms a large part of internal company operational data and its analysis can provide useful insight. The third section gives an application in a healthcare setting. The final section is a conclusion and meditation on the benefits of working with charismatic people.

2 Compositional data analysis examples

Professor Aitchison wrote a suite of computer programs to accompany his book. The CODA programs and data used in the book enabled readers and students to replicate the analyses in the book and apply the techniques to any data that they might encounter from research and consulting projects. The analysis of the activity patterns of a statistician over 20 days was one such dataset. Figure 1 shows a hand drawing of the ternary plot of the activities summarised into work, sleep and other.

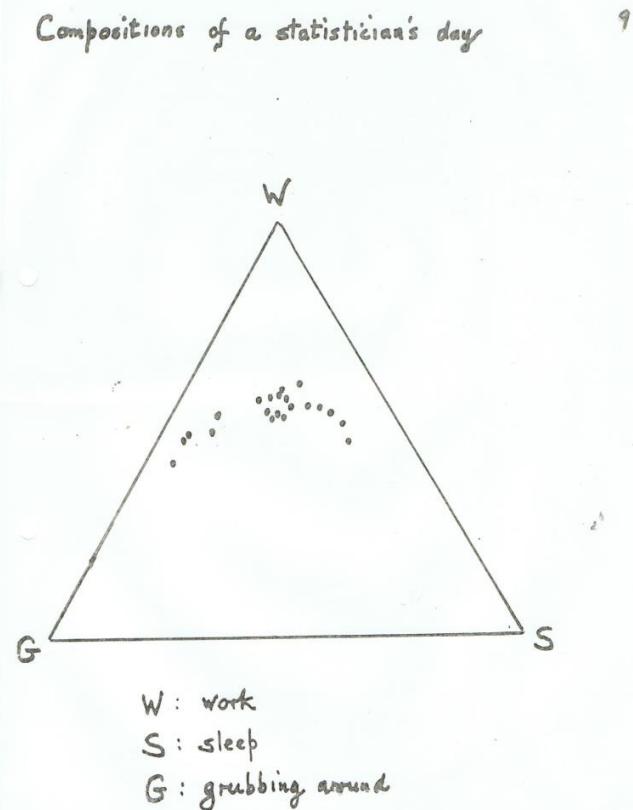


Figure 1 Ternary plot of activities of a statistician.

The ternary plot shows that a disagreeably large proportion of each day was spent working. This application prompted an analysis of the activities taking place in a classroom during a particular type of lesson.

Listening is an important part of teaching English language and was the focus of an analysis of classroom activities in a range of classes and schools encountered during teacher training within the Professional Educational Studies department of Hong Kong University (Coleman and Lee, 1988). As part of a project to study the effects of implementing loop systems for listening and the introduction of a listening section in the Hong Kong Certificate of Education English Language exam, a survey was carried out amongst student teachers of English in the Department of Professional Educational Studies and the Institute of Language in Education. The survey questionnaire contained general questions about the school, the listening equipment and the teachers' opinions on teaching listening. One item asked for the time spent on different activities and is thus a compositional data example. Responses were obtained from 99 student teachers.

The activities were summarised into 4 categories, being time spent:

- setting up and distributing, collecting and storing listening equipment
- preparing students for the listening activity, establishing and maintaining “class order”
- by students actually doing the listening activity
- checking student performance.

It was found that the nature of the target class and the school level affected the pattern of lesson activities with more listening carried out in the higher target classes and higher level schools. More time was spent dealing with educational equipment in lower level schools.

An interesting set of compositional data that Professor Aitchison used in his teaching was the proportions of different blood genotypes (MM, MN and NN) in 26 samples of people from different ethnic groups. The proportions are analysed by first calculating their geometric mean, g , then taking logs of the ratios of proportions to geometric mean giving a centred log ratio covariance matrix, then finding principal components (PCs). Dividing by the geometric mean introduces a sum to zero constraint on the log ratios but is different in nature to the original sum constraint on the proportions as it is the result of a construction that would induce dependence even if the original set of data points were not part of a zero sum constraint. The new constraint is reflected in the fact that the third PC has zero eigenvalue. Table 1 shows the analytical output.

Table 1 Principal components analysis of genotype proportions

Eigen analysis of the Covariance Matrix

Eigenvalue	3.1858	0.0447	0
Proportion	0.986	0.014	0
Cumulative	0.986	1	1

Variable	PC1	PC2
log(MN/g)	0.012	0.816
log(MM/g)	0.701	-0.418
log(NN/g)	-0.713	-0.398

The first PC indicates the variation and the second PC tells the equation of the curve along which there is variation. The expression for the second PC is approximately equal to a constant as its variation as indicated by its eigenvalue of 0.0447 is nearly zero.

PC2 is approximately, $2 \cdot \log(MN) - \log(MM) - \log(NN) = \text{constant}$

and this can be rearranged to $MN^2 / (MM \cdot NN) = \text{constant}$

The value of the constant can be approximated from the data and is just over 4.

Hence the equation is $MN^2 = 4 \cdot MM \cdot NN$ which is the Hardy-Weinberg equilibrium model that states that genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.

This is an interesting example showing the importance of looking at PCs with small eigenvalues especially if there is just one large PC and one small PC. Usually attention is only focused on PCs with large eigenvalues. So a very interesting result was obtainable from a simple set of sample data analysed taking account of the compositional nature of the data.

3 Compositional data in a healthcare setting

Compositional data forms a large part of internal company operational data and its analysis can provide useful insight. For example, the occurrence of failed and cancelled appointments on different days of the week, or changes in the proportions of different activities undertaken over time are important information for a service provider. Composite bar charts can be used to present the proportions but mask the fact that the proportions are constrained to add up to the whole.

Figure 2 shows a business oriented ternary diagram illustrating the change in proportions of activities carried out in a healthcare setting. The service provider is keen to move from routine activities to more advanced value added procedures and the diagram illustrates their success in moving closer towards the advanced practice vertex in quarter 2 compared to a predominance of routine examinations in quarter 1.

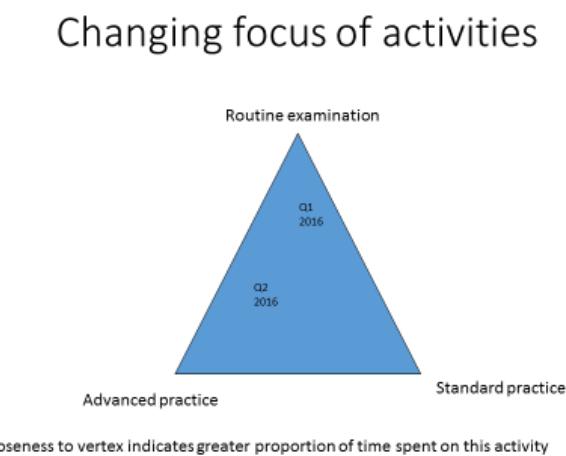


Figure 2 Ternary diagram illustrating changing practice.

Using ternary diagrams to illustrate proportions is an informative way to share data with company staff.

4 Conclusions

Learning about compositional data analysis was an enjoyable expansion of my knowledge of different types of statistical analysis. Studying the subject and then teaching it to students was a very satisfactory experience. The students grasped the ideas well and the applications are clearly wide and important.

The methodologies were an effective way to analyse the activities in the classrooms studied in our listening exercise. Using ternary diagrams is a good way to represent data in a healthcare setting.

Professor Aitchison's lectures were rich with applications and also theory, for which he consistently gave full and clear explanations and derivations. His style was affable and self-deprecating. He

S. Y. Coleman

page 6

described an incident in his inaugural lecture at Hong Kong University as a warning about the capricious nature of random variation. He was demonstrating the random nature of sampling by drawing coloured balls from a bag containing a mixture of different colours but as luck would have it he consistently drew exact proportions instead of the random variation he was intending to demonstrate.

Professor Aitchison was a charming man. He had a warm family life and dedicated his compositional data book in memorable style to his wife with the words:

“To M. the constant among many variables”

He conducted the Statistics Department at Hong Kong University in a very calm and inclusive manner encouraging high level research work and responsible, dedicated, professional focus on teaching the students.

It is a testament to his reach and influence that CODA is such a thriving community and that CODAWORKS conferences are so active and well attended.

References

- Aitchison, J.A. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.
- Coleman, S.Y. (2016). Data mining Opportunities for Small to Medium Enterprises from Official Statistics, *Journal of Official Statistics* 32(4), 849-866.
- Coleman, S.Y. and S. Lee (1988). Compositional data analysis of activities in a typical lesson' *Hong Kong Educational Research Journal* 3, 98-103.

Compositional data analysis of household waste recycling centres in Denmark

V.M.E. Edjabou¹, J. A. Martín-Fernández², A. Boldrin¹, and T.F. Astrup¹

¹Department of Environmental Engineering, Technical University of Denmark, UK;

vine@env.dtu.dk

² Dept. Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

Abstract

The Danish government has set a target of 50% recycling rates for household waste by 2022. To achieve this goal, the Danish municipalities should increase the source separation of household waste. While significant knowledge and experiences were locally gained, lessons learnt have not been extensively exploited country-wise, an important reason being that the influence of these changes has not been rigorously investigated and quantified, meaning that generalized conclusions could not be drawn so far. One of the reasons is that a consistent calculation method to assess and document the effect of these projects on the recycling rates does not exist. Thus, compositional data analysis technique was applied to analyze consistently waste data. Based on the waste composition obtained from a recycling center in Denmark, we analyzed the composition of waste treatment and disposal options. Zero and non-zero pattern was used to describe historical changes in the definition and components of waste fractions. Variation array was applied to determine the relationship between waste treatment and disposal options. As a result, compositional data analysis technique enables to analyze waste data regardless of the unit (mass or percentage).

Key words: geometric mean, recycling center, variation array, waste treatments.

1 Introduction

Over the last decade, about 80% of Danish waste is incinerated to produce 20% of all district heating and 5% of the electricity consumption (Fruergaard et al., 2010). In addition, the bottom ash from incineration plants are primarily used for construction purpose after recovery of scrap metal (Allegriini, 2014). However, mounting pressure on resource supply, to satisfy future societal needs, requires an effective use of available resources (European Commission, 2013).

To ensure a sustainable resource management, and the transition to the circular economy, the Danish Government, in 2013, launched its Resource Strategy Plan. This plan mandates that, by 2022, at least 50% of the following waste fractions: (1) paper, (2) board, (3) plastic, (4) metal, (5) wood, (6) glass and (7) food waste should be source-sorted and collected separately and recycled (Danish Government, 2013).

Generally, the Danish citizens increasingly dispose their waste at recycling centers (Toft et al., 2015). Recently, researchers showed that an increased amount of recyclable waste fractions (paper, board, plastic, metal, wood, glass and food waste) was misplaced in the containers intended for small combustible waste, which is currently incinerated (Edjabou et al., 2015). To motivate citizens to source separate their waste disposed at the recycling centers, numerous Danish municipalities and waste companies managing the recycling centers implemented various relevant legislations and initiatives to increase recycling rates.

While significant knowledge and experiences were locally gained, lessons learnt have not been extensively exploited country-wise, an important reason being that the influence of these changes has not been rigorously investigated and quantified, meaning that generalized conclusions could not be drawn so far. One of the reasons is that a consistent calculation method to assess and document the effect of these projects on the recycling rates does not exist.

Data handling is a particularly critical issue because using different metrics on the same data can provide contradicting results (Martín-Fernández et al., 2015). For example, when choosing the waste composition for traditional statistical analysis, using either a percentage composition or a mass composition is highly critical and may generate different and often contradictory results. Consequently, the interpretation of the results appears inconsistent, while comparison with other studies is not possible.

The overall objective of this study is to develop a comprehensive procedure for analysis of waste data set that reflect the inherent properties of the waste data set. This paper also aims at providing practitioners within solid waste management and planning with a systematic and easy way to analyze their data to effectively develop a consistent public awareness campaign and for future planning of recycling centers.

2 Methods and materials

2.1 Solid waste data

We analyzed yearly data for solid waste fractions from a recycling center in the suburb of Copenhagen in Denmark. The waste data were collected over the period 2010 – 2016.

In this recycling center, the solid waste was source-separated into 52 waste fractions.

For this study, we grouped the waste fractions into waste treatment and disposal options consisting of (1) incineration, (2) recycling, (3) landfill and (4) other treatments (Fischer, 2014). Here, other treatments are primary, special treatment of hazardous waste such as batteries, preservative treated wood, and polyvinylchloride (PVC).

2.2 Data analysis

We use zPatterns function from zCompositions package (Palarea-Albaladejo and Martín-Fernández, 2015) to describe the historical changes in the number, definition and components of waste fractions at this recycling center.

The geometric mean barplot (Martín-Fernández et al., 2015) was used to describe differences in waste

Edjabou, Martín-Fernández, Boldrin, and Astrup

3

treatment and disposal options between years from 2010 to 2016. A compositional variation array was used to explore the center and the variability of the waste data set (Pawlowsky-Glahn and Egozcue, 2011). Modelling and analysis of data were carried out using the open source R statistical programming language (R Core Team, 2017) and the freeware CoDaPack (Thió-Henestrosa and Comas-Cufí, 2011).

3 Historical changes in the number of waste fractions

Figure 1 presents the zero and non-zero of total mass per year of individual waste fractions recorded at the recycling center. We identified 14 zero patterns, and their percentage distribution is presented as horizontal bars. These zero patterns represent the number of combinations of zero values in each year and for each waste fractions. For this study, zero patterns is used to describe historical changes in the number of waste fractions at the recycling center.

The data presented in Figure 1 show that the mass of about 35% of the mass of waste fractions was continuously recorded over the period 2010-2016, suggesting that the definition and the components of these fractions remained unchanged during this period. These fractions include glass packaging, gardening waste, paper, board, metal, soil, combustible, impregnated wood, PVC, tires, hazardous waste and oil, large household appliances, asbestos, etc. However, the definition and components of 65% of waste fractions changed during the same period, of which of 29% of waste fractions was never disposed of at this recycling center.

The highest number of waste fractions at this recycling center was reported in 2010 and 2015, whereas the lowest was found in 2012.

These results suggest various waste sorting guidelines and container signage were implemented during this period.

Among the focus fractions defined by the Danish Resource Strategy Plan, the definition and components of wood and plastic waste varied during this period. In 2010, wood was segregated into two fractions consisting of cleaned wood, semi-cleaned wood. From 2011 to 2014, wood was collected as combustible and incinerated. However, in 2015 cleaned-wood waste fraction was established again at the recycling center.

Edjabou, Martín-Fernández, Boldrin, and Astrup

4

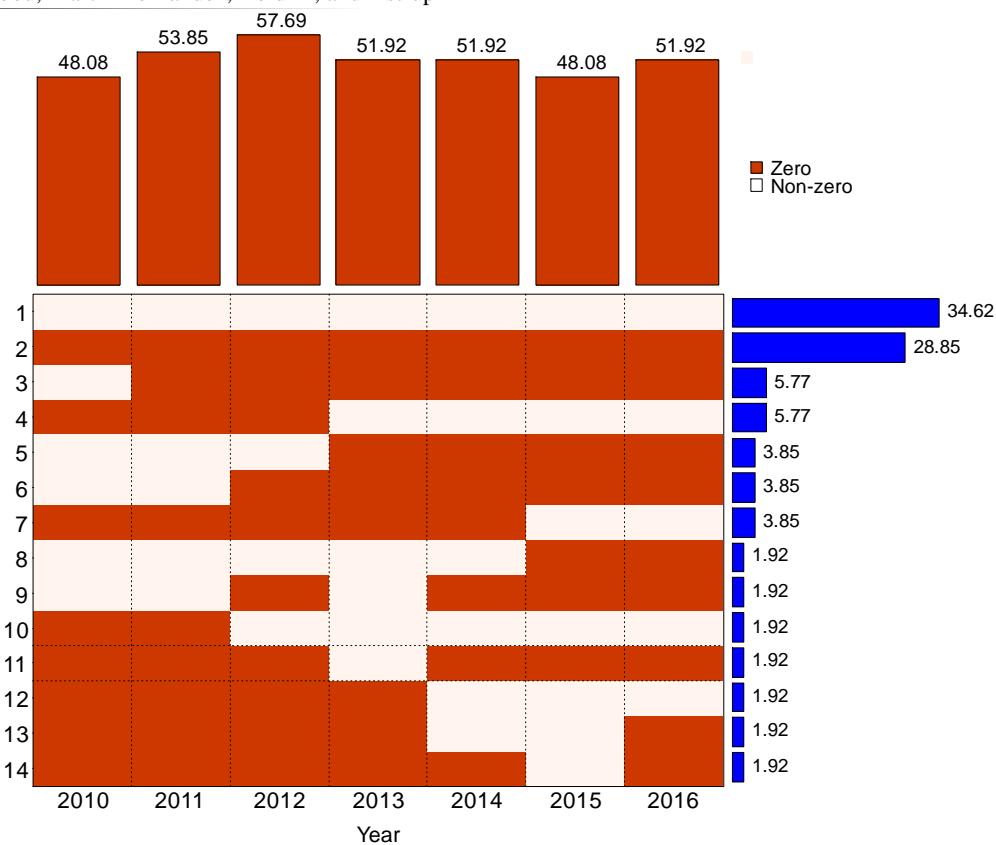


Figure 1: Patterns of zero and non-zero mass of individual waste fractions recorded at the recycling center in the suburb of Copenhagen in the period 2010-2016.

4 Compositional descriptive statistics

For this study, we considered waste treatment and disposal options. This means that 52 waste fractions were grouped according their treatment and disposal options. Thus, we analyzed the composition of waste treatment and disposal options consisting of landfill, other treatments, incineration and Recycling. We computed the center of the whole data set taking into consideration the Aitchison geometry (Martín-Fernández et al., 2015). For individual year, the composition was computed based on the total mass of the whole year.

The data in Table 1 show that recycling waste was the predominant waste fraction collected at the recycling center. On the other hand, solid waste intended for the other treatments was the smallest proportion of total waste disposed of at the study recycling center.

Table 1: Center in the whole period (2010-2016) and the composition for each year.

Year	Landfill	Other treatments	Incineration	Recycling
2010	6.6	1.5	26.3	62.4
2011	6.5	2.3	24.9	65.5
2012	5.6	1.4	24.5	67.4
2013	3.7	1.4	23.4	69.8
2014	4	0.8	25.4	68.9
2015	3.4	0.6	22.8	72.1
2016	3.8	0.7	22.3	72.7
2010-2016	4.6	1.1	24.2	68.3

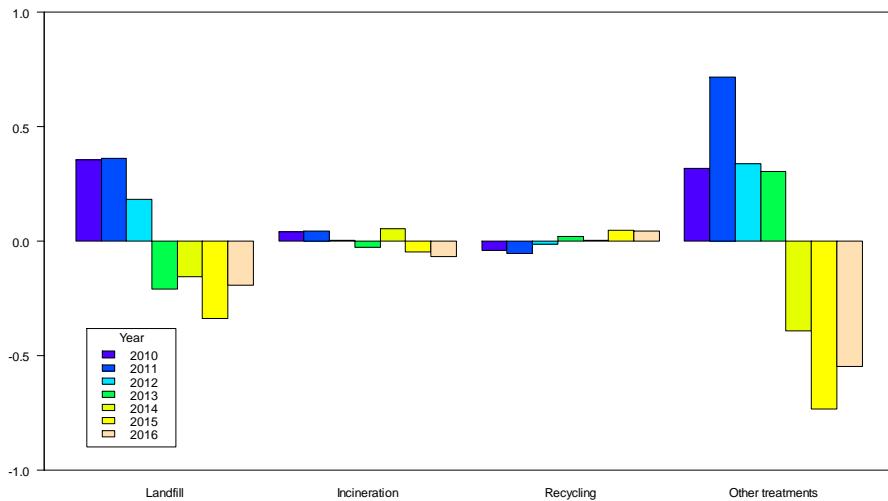


Figure 2: Geometric mean barplot showing the difference in the composition of waste treatment and disposal options over the period 2010-2016.

While data in Table 1 highlight considerably increased in the percentage of recyclables waste from 2010 to 2016, Figure 1 show that for incineration and recycling waste treatments, the differences between the groups (years) and the whole period are minor. However, the difference between groups (years) and the whole period are important for landfill and other treatments. Additionally, the proportion of these waste treatments options (landfill and special treatments) was lower than the whole period from 2013 to 2016 for landfill and from 2014 to 2016 for other treatments. The results may suggest that the changes in the number of waste fractions and the Danish Resource Strategy Plan launched in 2013 contributed to reduce the volume of misplaced waste in the containers intended for landfill and other treatments.

5 Variation array

The variation array is shown in Table 2 and it is divided into two rectangles. The upper rectangle presents ratio between waste treatment and disposal options as a pairwise log-ratio variances (variance $\ln(X_i/X_j)$).

The lower triangle shows the pairwise log-ratio means. Here, the numerator is by column (X_i) and the denominator (X_j) is by row. Furthermore, the sign (+ or -) of log-ratio mean values refer to the direction of the ratio between the two relevant waste treatment options. For example, the ratio between incineration and recycling can be calculated as: recycling/incineration = $\exp(1.12)$.

The percentages of centered log-ratio (clr) variances indicate that the largest single contributor to

total variation in the waste treatment options from 2010 to 2016 is other treatments, which amounted 55% of the total variation. Recycling was the second predominant contributor to the total variation (24%).

The data from Table 1 indicate a proportional relationship between recycling and incineration because their log-ratio variance is small and closed to zero (colored in blue). These results suggest that a reduction in the amount of misplaced materials in the containers intended for combustible may considerably increase the recycling rates of the study recycling center. In contrast, the results showed that there was no relationship between other treatments and both incineration and recycling.

Table 2: Variation array of the composition of waste fractions grouped per treatment options.

		Variance $\ln(X_i/X_j)$			
$X_i X_j$	Landfill	Incineration	Recycling	Other treatments	clr variances (%)
Landfill		0.06	0.09	0.10	8
Incineration	1.62		0.00	0.23	13
Recycling	2.74	1.12		0.29	24
Other treatments	-1.41	-3.03	-4.15		55
Mean $\ln(X_i/X_j)$					100

6 Conclusions

This study attempts to address the problem associated with analysis of data for of waste fractional composition. Based on zero pattern we found that the definition and components of 64% of waste fractions changed during the period from 2010 to 2016. The geometric mean barplot showed a considerable difference in the proportion of waste being landfilled and those treated and disposed of by means of other treatments options from 2010 to 2016. In contrast, a minor difference was observed in the proportion of incinerated and recycled waste during the same period. The variation array revealed a relationship between the proportions of waste incinerated and recycled.

Acknowledgements

The authors acknowledge the Danish Environmental Protection Agency (EPA), Kara/Novoren for providing data for this study.

References

- Allegrini, E. (2014). Resource recovery from waste incineration residues 69.
- Danish Government (2013). Denmark without waste: recycle more -incinerate less. Danish Ministry of the Environment, Copenhagen, Denmark.
- Edjabou, M.E., Jensen, M.B., Götze, R., Pivnenko, K., Petersen, C., Scheutz, C., Astrup, T.F. (2015). Municipal solid waste composition: Sampling methodology , statistical analyses , and case study evaluation. *Waste Management* 36, 12–23.
- European Commission (2013). SMES, Resource efficiency and green market.
- Toft, R., Fisher, C., Bøjesen, N.A., Kristensen, E. (2015) 2013. Affaldsstastistik (Waste statistics) 2013. Danish Environmental Agency (EPA) Copenhagen, Denmark
- Martín-Fernández, J., Daunis-i-Estadella, J., Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *Sort* 39, 1–22.
- Palarea-Albaladejo, J., Martín-Fernández, J.A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 143, 85–96.
- Pawlowsky-Glahn, V., Egozcue, J. (2011). Exploring Compositional Data with the CoDa-Dendrogram. *Austrian Journal of Statistics* 40, 103–113.
- R Core Team (2017). R: A Language and Environment for Statistical Computing.
- Thió-Henestrosa, S., Comas-Cufí, M. (2011). CoDaPack v2 USER 's GUIDE.

**Compositional data: simple questions, difficult answers;
or
implications of the sample space choice**

J. J. Egozcue¹, and V. Pawlowsky-Glahn²

¹Universidad Politécnica de Catalunya, Barcelona, Spain; *juan.jose.egozcue@upc.edu*

² Universitat de Girona, Spain

Abstract

Early definitions of compositional data were based on the constant sum of the components. In the eighties, John Aitchison complemented this definition with some properties and principles. However, a formal definition of compositional data and their different typologies is still pendent. Frequently, although not free of controversial opinions, compositional data are identified with those data that are analysed with the log-ratio approach. This implies that the attention is directed to the ratios between components. However, there are cases in which the parts do not have the adequate scale, e.g. the scale can be closer to the absolute scale more than to the ratio scale required for a log-ratio approach. Here two main points are addressed: (I) the scale of the data is frequently hidden by the constant sum constraint or other characteristics of the data, and (II) when data are claimed to be parts of a whole, there is no indication about whether these parts are overlapping or not.

Some simple questions may illustrate the lack of precision of the present definitions. Let us formulate one of them.

A Professor examined her students two times and their assessments consist of a score per exam. The scores are numbers from 0 to 10, both included. In order to study the relation between the two scores the Pearson correlation coefficient is computed. Is this correlation coefficient an appropriate measure of dependence? Is it spurious? Should a 4 score be considered as the pair (4, 6)? Is this a Likert scale? Are these data compositional? If yes, can they be treated using log-ratios, or may be should they be identified with a clr-transformed composition?

The first step of any statistical analysis should be to decide which is an appropriate sample space for the available data and which is its mathematical structure (operations, scales, distances, projections). This structure should allow answering the stated questions. However, a given data set may be viewed within different sample spaces with different structures, and each structure has different implications. The adequacy of the choice is referred to answers obtained for the stated questions. The definition of any type of compositional data requires a detailed description of the sample space and its structure.

Key words: scale, spurious correlation, overlapping categories, compositional equivalence, Aitchison geometry

1 Introduction

The attention to compositional data was motivated by the detection of spurious correlation (Pearson, 1897; Chayes, 1971). The notion of spurious correlation in the second reference was linked to the effect of normalizing compositions to a constant sum. This explains the fact that early definitions of Compositional Data (CoDa) were based on the Constant Sum Constraint (CSC). In the eighties, J. Aitchison introduced the log-ratio approach for the analysis of CoDa, which required additional assumptions. These assumptions can be summarized in two principles, the scale invariance and the subcompositional dominance of distances, which in turns assume that the compositional information is in the ratios between components (Aitchison, 1986, 1982). These principles or assumptions have been discussed and reformulated several times (Barceló-Vidal et al., 2001; Martín-Fernández et al., 2003; Aitchison and Egozcue, 2005; Egozcue, 2009; Egozcue and Pawlowsky-Glahn, 2011; Pawlowsky-Glahn et al., 2015; Barceló-Vidal and Martín-Fernández, 2016). However, these assumptions exclude some data types satisfying the CSC, but not fulfilling the principles of the log-ratio approach. A typical example of this is the presence of essential zeros. The log-ratio approaches place any composition containing a zero at infinity and this violates the assumption on ratio information. This means that hypothesis on how the scale of the data is perceived may be a difficulty to use CoDa-logratio techniques in their analysis. The hypothesis on the scale of the data and its transformation is one of the issues which are discussed here.

CoDa have been conceived as some positive components describing parts of a whole. However, there is no indication about whether these parts should be disjoint or may be overlapping in some way. The non-overlapping assumption is implicit in most of the CoDa applications, but it seems that there is no inconvenient to use log-ratio approaches when there is some kind of overlapping between different components of a composition whenever the assumptions on scale invariance and subcompositional coherence hold.

2 One question, several sample spaces

In order to informally discuss the importance of the data scale, the measures of difference, distances and, in general, the structure of a sample space, a simple question by example will be helpful. It is intended to enhance the fact that the choice of a sample space of a data set has an important subjective component, but it has also important implications. The choice of the sample space should be driven by the questions put forward to the data. The obtained answers can differ dramatically or not as they depend on the underlaying assumptions and the interpretation of the context. Next example provides a case where both real and compositional approaches can be adopted. The question follows.

A Professor examined her students two times and their assessments consist of a score per exam. The scores are numbers from 0 to 10, both included. The 0 score is the worst score and 10 score is the best one. In order to study the relation between the two exam scores, the Pearson correlation coefficient is computed. Is this correlation coefficient an appropriate measure of dependence or covariation? Is it spurious? Should a 4 score be considered as the pair (4, 6)? Is this a Likert scale? Are these data compositional? If yes, can they be treated using log-ratios, or may be should they be identified with a clr-transformed composition?

Several sample space approaches can be used to deal with this question and some of them are discussed in the following sections. In order to numerically illustrate the example, some scores from exams, X_0 and Y_0 , are used (see Appendix 3). In what follows, variances (var), covariances (cov) and correlations (cor) refer to sample variances, covariances and correlations. Some standard concepts of compositional data analysis are used without any introduction. They can be found, for instance, in Pawlowsky-Glahn et al. (2015).

Case A: Subjective scaling

A first step consists of considering the scale of the scores. One can realize that the obvious approach of directly computing the Pearson correlation coefficient between the two exam scores implicitly assumes that the scores are in an absolute scale, i.e. the distances between a 4 score and a 5 score is equal to the distance between 9 and 10 scores, and so on. In terms of sample space, this means that the two exam scores are assumed embedded in \mathbb{R}^2 and that its Euclidean geometry holds, restricted to the $[0, 10] \times [0, 10]$ square.

However, any teacher knows that the exam scores are strongly subjective relative to her/his way of evaluating. A weaker assumption is that this score scale varies from evaluator to evaluator, but remains approximately constant along score exams of the same subject and evaluator.

A way to correlate the two exam scores is to transform the subjective scale of the scores into an absolute scale in \mathbb{R}^2 using a scale function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ such that the distances between the re-scaled scores corresponds to the Lebesgue measure or absolute scale. These functions should be monotonous since the scores are assumed ordered. In practice, the function ϕ can be restricted to intervals containing the original scores and the re-scaled scores. A way to assess the adequate function ϕ in a particular case is to inquire the evaluator about her/his scale. Figure 2 shows different cases of ϕ , denoted ϕ_i , $i = 0, 1, \dots, 4$, two of them corresponding to the authors of this paper. The original data and their scaled versions are listed in the table in Appendix 3 under labels X_i , Y_i , $i = 0, 1, \dots, 4$.

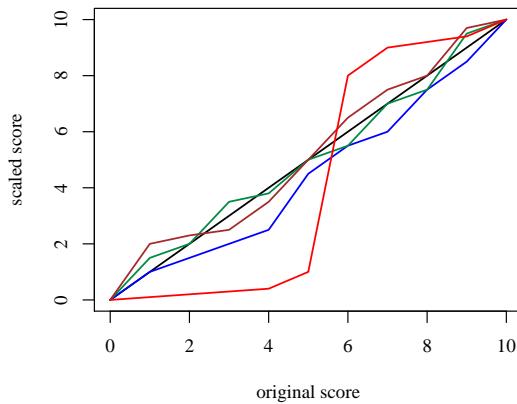


Figure 1: Five cases of scale function ϕ_i . The identity function ϕ_0 in black. Three different assessments in blue, green and brown. An extreme assessment ϕ_4 in red.

Table 1: Correlation between the scores of two exams (X_i, Y_i). Scores were scaled with the scale functions in Figure 2. Data presented in Appendix 3.

$\text{cor}(X_0, Y_0)$	$\text{cor}(X_1, Y_1)$	$\text{cor}(X_2, Y_2)$	$\text{cor}(X_3, Y_3)$	$\text{cor}(X_4, Y_4)$
0.5623	0.5370	0.5443	0.5444	0.4687

The Pearson correlation coefficient is the cosine of the angle subtended by the \mathbb{R}^n vectors which components are in the columns of the data matrix. It is computed as the ordinary inner product of the two normalized vectors. In order to support this interpretation the two data columns must be real since the inner product is taken in \mathbb{R}^n . As is well known, the Pearson correlation coefficient is

invariant under affine transformations (translations, multiplication by constants) of the compared scores. This important property allows to compare scores in modified scales. For instance, this allows to compare the marks of one exam within the interval $[0, 10]$ and the marks of another exam within the interval $[0, 100]$, or even within the interval $[-50, 50]$. Implicitly, we are assuming that the scores are equally informative when an affine transformation is applied to them.

From the results in Table 1, one can conclude that the transformation of the original scale changes the correlation and, therefore, it is a measure of dependence linked to the assumed scale. However, the changes in the correlations are not dramatic, even in the case of $\text{cor}(X_4, Y_4)$, the most extreme case of change of scale.

Case B: Scores as compositions

An alternative sample space for this problem is to conceive each score as a composition. For instance, a score of 4.5 can be equally represented as the two part composition $(4.5, 5.5)$ supplementing the score to 10. This approach does not allow any log-ratio approach as a zero score would be transformed into $(0, 10)$ and the ratio of the components is then not informative. A slight modification permits overcoming this difficulty: a score X is assigned a two part composition $(X - a, b - X)$ for some values $a < 0$ and $b > 10$, for instance $a = -1$ and $b = 11$, to be used in the sequel. This quite naïve approach assumes that the relevant information is contained in $(X - a)/(b - X)$. This is equivalent to represent the composition by the only coordinate $Z = \phi_{ab}(X) = \log((X - a)/(b - X))/\sqrt{2}$, where $\sqrt{2}$ is just a normalizing constant. The coordinate Z is assumed to be in an absolute scale. Then, these assumptions are a scaling of the initial scores by $\phi_{ab}(X)$. For the values $a = -1$ and $b = 11$, the scaling function is near to an affine scaling. Once the scores of the two exams are transformed, their correlation is 0.5455.

Correlation between compositions in two different simplices has been studied in different ways (Bergman and Holmquist, 2012; Székely et al., 2007). Here the distance correlation, also known as energy correlation, is used in different situations as it is a general correlation between samples in two metric sample spaces. This is the case for the two scores transformed by $\phi_{ab}(\cdot)$, and the computed distance correlation is 0.5719.

Case C: Scores as clr of a composition

Like in the previous approach (Case B), a score X can be considered as a pair $(X, 10 - X)$. Assuming that the scale of X is absolute, or has been transformed by means of a scale function, a (real) shift of the pair to $(X - 5, 5 - X)$ preserves the information. This is a common practice when dealing with Likert scales in surveys. Note that the 5 appearing in the shift is obtained as the mean value of X and $10 - X$ so that the sum of the shifted components is zero. Therefore, the pair $(X - 5, 5 - X)$ satisfies the conditions (absolute scale, zero sum) of a centered log-ratio transformation (clr) of a composition \mathbf{U} , which can be computed as

$$\mathbf{U} = \mathcal{C}(\exp(X - 5), \exp(5 - X)) .$$

This composition can be represented as an ilr (isometric log-ratio) coordinate $\phi_{\text{clr}}(X) = \text{ilr}(\mathbf{U}) = ((X - 5) - (5 - X))/\sqrt{2} = \sqrt{2}(X - 5)$. As in the previous case B, the Pearson and distance correlation between the coordinates for the scores from the two exams are computed, resulting in 0.5623 and 0.5851, respectively. It can be observed that the Pearson correlation, in this case, is equal to the left hand side correlation in Table 1, which correspond to the original data. This is not a surprise as $\phi_{\text{clr}}(X) = \sqrt{2}(X - 5)$ is an affine transformation of X (the same for Y).

Case D: Composition with a total

In previous cases, the two scores X and Y , transformed or not, are placed in different spaces which are finally put together for analysing relations between them. Another possibility is to consider the two scores as a couple (X, Y) which can be used to construct a composition. Although there are several possibilities, the following model is here considered. The pair (X, Y) is first centered,

$$\left(\frac{X - Y}{2}, \frac{Y - X}{2} \right) ,$$

and then identified with the clr of a composition \mathbf{V} , which expression and ilr coordinate are

$$\mathbf{V} = \mathcal{C} \left(\exp \left(\frac{X - Y}{2} \right), \exp \left(\frac{Y - X}{2} \right) \right) , \quad W = \text{ilr}(\mathbf{V}) = \frac{1}{\sqrt{2}}(X - Y) .$$

In this case, the difference $(X - Y)$ does not contain the complete information about the exams and it is preferable to keep the total score $T = X + Y$ in mind and treat it as a \mathcal{T} -space (Pawlowsky-Glahn et al., 2015). Since X and Y were considered real, so should be T . Under these assumptions, the sample space of $(V_1, V_2, T) = (c \exp[(X - Y)/2], c \exp[(Y - X)/2], X + Y)$ is contained in $\mathbb{S}^2 \times \mathbb{R}$. It can be represented by means of two real coordinates, namely (W, T) .

The question now is how the relationship between the two exam scores X and Y can be measured. The balance-association between two parts is defined as a constant balance between the two parts, or as the parts are proportional (Egozcue et al., 2013; Lovell et al., 2015; Egozcue et al., 2017). In this case, perfect association implies $W = (X - Y)/\sqrt{2}$ is constant. A measure of this balance-association is the sample variance of the balance; if it is small, it suggests a strong association; if it is large, then, the association is weak or does not exist. When the statistics $\text{var}(W) = \text{var}(X - Y)/2$ is small, association between the two scores can be considered. Note that the variance of the simple log-ratio is $\text{var}(\ln(V_1/V_2)) = 2\text{var}(W) = \text{var}(X - Y)$; the value was estimated as $\text{var}(X - Y) = 7.86$. A way of presenting the variation as a kind of determination coefficient can be $R_{\text{var}}^2 = 1/(1 + \text{var}(X - Y)) = 0.11$, which again suggests weak association between exam scores.

Apparently, the measure of association does not depend on the total $T = X + Y$, but it plays a role when the balance-association is not perfect. Measuring balance-association requires a variance reference to decide what can be considered as *small* or *large*. In this case a linear model helps understanding the role of the reference:

$$\ln \frac{V_1}{V_2} = \beta_0 + \beta_1 T + r ,$$

where r are the residuals. For non-significant $\beta_1 = 0$, the balance (or simple log-ratio) $\ln(V_1/V_2) = X - Y$ is not rejected to be constant and, consequently, balance association between $V_1 = c \exp[(X - Y)/2]$ and $V_2 = c \exp[(Y - X)/2]$ is also not rejected. In our example, the linear model is

$$(X - Y) = \beta_0 + \beta_1(X + Y) + r , \tag{1}$$

and the estimated values are $\beta_1 = -0.10499$ (p-value ??0.27), $R^2 = 0.038$. This indicates that non-association cannot be rejected. As the estimate of β_1 is proportional to

$$\text{cor}(X - Y, X + Y) = \frac{\text{var}(X) - \text{var}(Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} ,$$

which suggest balance-association for small values near to zero. The values obtained for the measures of association are $\text{var}(\ln \frac{V_1}{V_2}) = \text{var}(X - Y) = 7.86$ and $\text{cor}(X - Y, X + Y) = -0.20$, thus pointing out a weak relationship between the two exam scores.

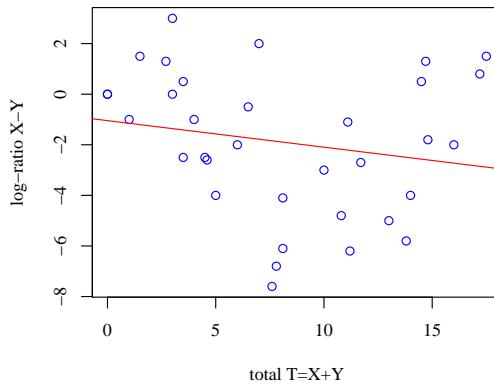


Figure 2: Scatterplot of $(X - Y, X + Y)$ and the regression line fitted for the model in Eq. 1. The slope is non-significant suggesting a weak balance association.

Case E: Composition including the sum of components

As in case B, the exam scores are assumed to be values in the interval $[a, b]$ ($a = -1$, $b = 11$). In this case the information is assumed to be contained in the ratio $(X - a)/(b - Y)$. Nevertheless, the information about the total score $X + Y$ can be relevant and, consequently, $X + Y$ is added as a component to the composition. The composition in \mathbb{S}^3 is

$$\mathbf{V} = \mathcal{C}(X - a, Y - a, X + Y).$$

In order to measure the relation between the two transformed scores $X - a$, $Y - a$, the variance of the log-ratio $\ln((X - a)/(Y - b))$ or its transformation $R_{\text{var}}^2 = (1/\ln((X - a)/(Y - b))) - 1$. A set of two balances has been selected for the representation of this composition in \mathbb{S}^3 . They are

$$W_1 = \sqrt{\frac{1}{2}} \ln \frac{X - a}{b - Y}, \quad W_2 = \sqrt{\frac{2}{3}} \ln \frac{\sqrt{(X - a)(b - Y)}}{X + Y},$$

the absolute value of the first one measuring association of the two exam scores and the second accounting for the relationship of the scores and their total. The value of the variation, as a measure of association between $X + 1$ and $Y + 1$, is

$$2\text{var}(W_1) = \text{var} \left(\ln \frac{X - a}{b - Y} \right) = 0.516, \quad R_{\text{var}}^2 = \frac{1}{1 + 2\text{var}(W_1)} = 0.66,$$

that estimates a slightly stronger relation between the exam scores.

3 Comparison of approaches

Although all approaches proposed suggest that the relation between the two exam scores is weak, one can ask for which one is the best or how different the approaches are. The first part of the question is easy: the best approach is that one which assumptions best fit the preferences of the analyst, that is the choice is subjective. The second part of the question can be approached using the expression in coordinates of the data represented in coordinates of each sample space. This comparison requires comparing real data in one or two dimensions. Distance correlation (Székely et al., 2007) allows to compare the interdistances between individuals for each approach. A large distance correlation, near to one, implies a large dependence between approaches; conversely

a small distance correlation points out almost independent variables representing the students. As the original data (X, Y) are the same for each approach, distance correlations between them visualize how close they are. The results are shown in Table 2. A graphical representation of

Table 2: Distance correlation matrix between different sample space cases.

	A	B	C	D	E
A	1.0000	0.9952	1.0000	0.9918	0.6073
B	0.9952	1.0000	0.9952	0.9871	0.6293
C	1.0000	0.9952	1.0000	0.9918	0.6073
D	0.9918	0.9871	0.9918	1.0000	0.5251
E	0.6073	0.6293	0.6073	0.5251	1.0000

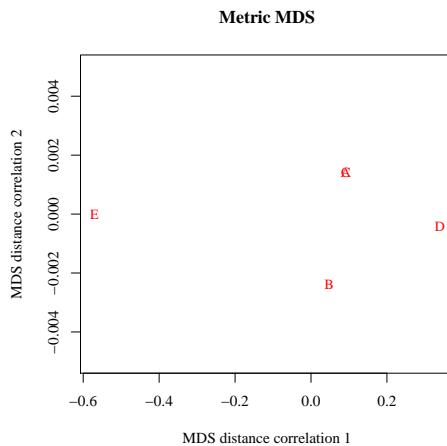


Figure 3: Multidimensional scaling of dissimilarities obtained from the distance correlations in Table 2. If $dcor$ is a distance correlation, dissimilarity is taken as $(1/dcor) - 1$. The symbols A and C are overlapped exactly. Note that the scale of the second axis is expanded for visualization.

the distance correlations in Table 2 can be obtained using multidimensional scaling. This requires transforming distance correlations into a distance or a dissimilarity. If $dcor$ is a distance correlation, the dissimilarity in MDS is $(1/dcor) - 1$. A multidimensional scaling of the dissimilarities is shown in Figure 3. The two dimensional projection is exact. Approaches A and C are exactly overlapping since their coordinates are proportional (see Table 4). Approaches A, B, C are very similar for this particular data set, and the main differences are between the approaches D, E which correspond to evaluating the relationship between scores with a log-ratio variance but a different scale of the variables. Nevertheless, these differences are not very important for the studied data as the conclusion is always: there is some relationship between the two exam scores but it is weak.

4 Conclusions

The main conclusion is that procedures for analyzing a given problem critically depend on the assumptions about the sample space and its structure. The best approaches correspond to those which assumptions better fit the stated problem. However, the choice of the adequate assumptions has a subjective character. Therefore, it is advisable to make clear which are the assumptions underlaying any statistical analysis.

Compositional data analysis is not an exception. The common log-ratio approach assumes that the parts of a composition have logistic scale, that relevant information resides in the ratios between components, and that compositions can be represented in the simplex endowed with the Aitchison geometry. This implies that perturbation is the natural operation between compositions; that analysis in subcompositions give coherent results with analysis in a larger composition. However, alternative assumptions are possible. The example proposed admits several sample space approaches despite its initial obvious non-compositional character. The assumptions proposed were mainly related to questions of scale and were reduced to selection of measures of association between variables. In real cases Pearson correlation was reported; in compositional cases variation (variance of a log-ratio) was used.

A comparison between different approaches has been conducted using distance correlation. It is based on the fact that each approach assumes a distance between individuals (students in this particular case), that is, the sample space provides different distances in each approach and they are compared with the distance correlation.

Acknowledgements

The authors acknowledge financial support by the Spanish Ministry of Education and Science under project ‘CODA-RETOS’ (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)) and by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project ‘COSDA’ (Ref. 2014SGR551).

References

- Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 44(2), 139–177.
- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London (UK): Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
- Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology* 37(7), 829–850.
- Barceló-Vidal, C. and J.-A. Martín-Fernández (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics* 45, 57–71.
- Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG’01 – The VII Annual Conference of the International Association for Mathematical Geology*, Cancun (Mex), pp. 20 p.
- Bergman, J. and B. Holmquist (2012). A measure of dependence between two compositions. *Australian & New Zealand Journal of Statistics* 54(4), 451–461.
- Chayes, F. (1971). *Ratio Correlation*. University of Chicago Press, Chicago, IL (USA). 99 p.
- Egozcue, J. J. (2009). Reply to “On the Harker variation diagrams;...” by J. A. Cortés. *Mathematical Geosciences* 41(7), 829–834.
- Egozcue, J. J., D. Lovell, and V. Pawlowsky-Glahn (2013). Testing compositional association. In P. F. K. Hron and M. Templ (Eds.), *Proceedings of the 5th Workshop on Compositional Data Analysis – CoDaWork 2013*. ISBN: 978-3-200-03103-6, <http://coda.data-analysis.at/>.

- Egozcue, J. J. and V. Pawlowsky-Glahn (2011). Basic concepts and procedures. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 12–28. John Wiley & Sons. 378 p.
- Egozcue, J. J., V. Pawlowsky-Glahn, and G. B. Gloor (2017). Linear association in compositional data analysis. *Austrian Journal of Statistics*. submitted.
- Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler (2015, 03). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput Biol* 11(3), e1004075.
- Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35(3), 253–278.
- Pawlowsky-Glahn, V., J. J. Egozcue, and D. Lovell (2015). Tools for compositional data with a total. *Statistical Modelling* 15(2), 175–190.
- Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. Statistics in practice. John Wiley & Sons, Chichester UK. 272 pp.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London LX*, 489–502.
- Székely, G. J., M. L. Rizzo, and N. K. Barikov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6), 2769–2794.

A Data

Table 3: The exam scores used to illustrate the question stated are in the columns X_0, Y_0 . They correspond to 34 students. The scores correspond to a pair of exams held in the Universitat Politècnica de Catalunya, but the course, professor, subject, and students are hidden for the sake of confidentiality. Next columns $X_i, Y_i, i = 1, 2, 3, 4$ are the same data after scaling them by the scale functions $\phi_i, i = 1, 2, 3, 4$. The scale function ϕ_4 corresponds to the more extreme assessment.

	X_0	Y_0	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	1.50	1.50	1.25	1.25	1.75	1.75	2.15	2.15	0.15	0.15
2	2.00	6.10	1.50	5.55	2.00	5.65	2.30	6.60	0.20	8.10
3	4.00	9.00	2.50	8.50	3.80	9.50	3.50	9.70	0.40	9.40
4	2.00	4.00	1.50	2.50	2.00	3.80	2.30	3.50	0.20	0.40
5	3.00	0.00	2.00	0.00	3.50	0.00	2.50	0.00	0.30	0.00
6	8.00	6.70	7.50	5.85	7.50	6.55	8.00	7.20	9.20	8.70
7	0.50	7.30	0.50	6.45	0.75	7.15	1.00	7.65	0.05	9.06
8	0.50	4.50	0.50	3.50	0.75	4.40	1.00	4.25	0.05	0.70
9	1.00	7.10	1.00	6.15	1.50	7.05	2.00	7.55	0.10	9.02
10	0.00	7.60	0.00	6.90	0.00	7.30	0.00	7.80	0.00	9.12
11	1.00	3.50	1.00	2.25	1.50	3.65	2.00	3.00	0.10	0.35
12	2.00	0.70	1.50	0.70	2.00	1.05	2.30	1.40	0.20	0.07
13	6.50	8.30	5.75	7.80	6.25	8.10	7.00	8.51	8.50	9.26
14	9.50	8.00	9.25	7.50	9.75	7.50	9.85	8.00	9.70	9.20
15	9.00	8.20	8.50	7.70	9.50	7.90	9.70	8.34	9.40	9.24
16	5.00	6.10	4.50	5.55	5.00	5.65	5.00	6.60	1.00	8.10
17	5.00	9.00	4.50	8.50	5.00	9.50	5.00	9.70	1.00	9.40
18	3.00	7.80	2.00	7.20	3.50	7.40	2.50	7.90	0.30	9.16
19	1.50	0.00	1.25	0.00	1.75	0.00	2.15	0.00	0.15	0.00
20	1.00	3.60	1.00	2.30	1.50	3.68	2.00	3.10	0.10	0.36
21	2.00	1.50	1.50	1.25	2.00	1.75	2.30	2.15	0.20	0.15
22	4.00	9.80	2.50	9.70	3.80	9.90	3.50	9.94	0.40	9.88
23	7.00	9.00	6.00	8.50	7.00	9.50	7.50	9.70	9.00	9.40
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.50	3.00	0.50	2.00	0.75	3.50	1.00	2.50	0.05	0.30
26	4.50	2.50	3.50	1.75	4.40	2.75	4.25	2.40	0.70	0.25
27	3.50	6.50	2.25	5.75	3.65	6.25	3.00	7.00	0.35	8.50
28	0.00	1.00	0.00	1.00	0.00	1.50	0.00	2.00	0.00	0.10
29	7.50	7.00	6.75	6.00	7.25	7.00	7.75	7.50	9.10	9.00
30	3.00	3.50	2.00	2.25	3.50	3.65	2.50	3.00	0.30	0.35
31	4.50	7.20	3.50	6.30	4.40	7.10	4.25	7.60	0.70	9.04
32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	1.50	2.50	1.25	1.75	1.75	2.75	2.15	2.40	0.15	0.25
34	2.50	8.70	1.75	8.20	2.75	8.90	2.40	9.19	0.25	9.34

B Summary table

The following table summarizes assumptions, expressions of variables and coordinates in the presented cases of sample spaces.

Table 4: Summary of assumptions for the studying relation between two exam scores (X, Y) in different approaches. Under *variables* simple transforms of (X, Y) into both real or compositional variables and its sample space (SS) as they are in the data matrix; row scale reports two types of scale for reference variables. When data are grouped in columns of the data matrix they are also in a sample space (SS). Label *var.* *assoc.* indicates which measure of association is used to deal with the problem of relations between X and Y . Under *coordinates* there are the expression of the real coordinates representing the data in \mathbb{R} or \mathbb{R}^2 .

case	variables	row SS	row scale	columns SS	var. assoc.	coordinates
A	(X, Y)	$[0, 10] \times [0, 10]$ $\subset \mathbb{R} \times \mathbb{R}$	absolute	$([0, 10]^n)^2$ $\subset \mathbb{R}^n \times \mathbb{R}^n$	correlation on coordinates	(X, Y)
B	$(X - a, b - X)$ $(Y - a, b - Y)$	$[a, b]^2 \times [a, b]^2$ $\subset \mathbb{S}^2 \times \mathbb{S}^2$	logistic	$\subset \mathbb{S}^n \times \mathbb{S}^n$	correlation on coordinates	$\ln((X - a)/(b - X))/\sqrt{2}$ $\ln((Y - a)/(b - Y))/\sqrt{2}$
C	$\mathcal{C} \exp(X - 5, 5 - X)$ $\mathcal{C} \exp(Y - 5, 5 - Y)$	$\subset \mathbb{S}^2 \times \mathbb{S}^2$	logistic	$\subset (\mathbb{R}^2)^n \times (\mathbb{R}^2)^n$	correlation on coordinates	$\sqrt{2}((X - 5), (Y - 5))$
D	$V_1 = c \exp[(X - Y)/2]$ $V_2 = c \exp[(Y - X)/2]$ $T = X + Y$	$\subset \mathbb{S}^2 \times \mathbb{R}$	V logistic T absolute	$\subset (\mathbb{S}^2 \times \mathbb{R})^n$	log-ratio variance	$W = (X - Y)/\sqrt{2}$ $T = X + Y$
E	$V_1 = c(X - a)$ $V_2 = c(Y - a)$ $V_3 = c(X + Y)$	$\subset \mathbb{S}^3$	logistic	$\subset (\mathbb{S}^3)^n$	log-ratio variance	$W_1 = 1/\sqrt{2} \ln((X - a)/(Y - a))$ $W_2 = \sqrt{2/3} \ln(\sqrt{(X - a)(Y - a)}/(X + Y))$

Merging key concepts in the chemistry of natural waters with compositional data analysis: Updates to basic water quality plots

**M.A. Engle^{1, 2}, A. Buccianti³, R.A. Olea^{1, 4}, and
M.S. Blondes¹**

¹U.S. Geological Survey University, Reston, Virginia, USA;

²Dept. Of Geological Sciences, University of Texas at El Paso,
El Paso, Texas, USA

³Dept. of Earth Sciences, University of Florence, Firenze, Italy

⁴Email: *rolea@usgs.gov*

Abstract

In the last decade, substantial efforts have been made in understanding how to apply compositional data analysis (CoDa) to the interpretation of water quality data. Presently, there is an improved understanding of the relationship between stoichiometry and chemical equilibrium and to log-ratio interpretative methods. This paper provides an update to CoDa-based graphical methods and their applications to water quality data, and offers examples of log-ratio versions of scatterplots, Stiff diagrams, Durov plots, and trilinear diagrams. Our intention is that this work can be used as a quick tutorial on current knowledge and ideas, and to increase the application of CoDa methods in routine interpretation of water chemistry data.

Key words: scatterplot; trilinear diagram; Stiff diagram; Durov plot; isometric log-ratio

1 Introduction

Like many types of compositional data, water chemistry is a unique subject with important distinctions that require further clarification and study. In 2005, three significant papers introduced important concepts and approaches for applications of compositional data analysis (CoDa) to water quality data. Buccianti and Pawlowsky-Glahn (2005) demonstrated the role of units (e.g., mg/L vs. mol/kg or eq/L) and their impact on the analyses, formulating log-ratio based principal component analysis (PCA) and frequency distributions of PCA scores as important interpretive tools, and utilized simplified log-contrasts to identify processes. The work of Otero et al. (2005) explained the relative nature of log-ratios, demonstrated the use of links in log-ratio based PCA to identify potential sources, and showed the importance of water concentration as a parameter in CoDa. Finally, Tolosana-Delgado et al. (2005) showed how stable isotope ratios can be incorporated into CoDa for applications to water quality data. Since the publication of these landmark papers, many other papers utilizing CoDa methods to interpret water quality data have been published, and multiple new approaches have been realized and presented (Buccianti and Pawlowsky-Glahn, 2005; Buccianti, 2011; Buccianti and Magli, 2011; Gallo and Puccianti, 2013; Engle and Rowan, 2013; Buccianti and others, 2014; Engle and Blondes, 2014; Engle and others, 2014; Buccianti, 2015; Blake and others, 2016; Buccianti and Zuo, 2016; Engle and others, 2016).

The purpose of this paper is to provide updates on some of these new approaches. Specifically, many commonly used water quality plots (i.e., scatterplots, trilinear diagrams, and Stiff diagrams) provide rapid, visual information about large datasets but may contain spurious correlations and other problems associated with improper treatment of compositional data (Buccianti and Magli, 2011; Engle and Rowan, 2013). The dataset applied for examination here consists of chemical data for 3,066 samples of brackish and saline groundwater from the Dockum aquifer in west Texas, USA. The Dockum is an important potential source of non-freshwater for use in hydraulic fracturing in underlying hydrocarbon reservoirs. The data were downloaded from the Texas Water Development Board and cleaned using methods described in Reyes (2014). Data were converted to units of mmol/L, unless otherwise specified.

2 Basic water quality plots

Water quality plots are important tools for the rapid screening and interpretation of water quality data, and allow visual identification of patterns and differences between samples (Hem, 1985). Common examples of water quality plots include Piper plots, Durov diagrams, Stiff diagrams, scatter plots, and trilinear diagrams. This section presents some suggested log-ratio based alternatives to these standard methods. One trait of basic data analysis is that often the work is circular; initial findings may lead to reorganizing the data and re-assessing the results. To that end, a simple robust centered log-ratio based on principal components using the subcomposition [Ca, Mg, Na, Cl] shows that the data set can be naturally broken up into categories (Group I and Group II; Figure 1). The Group I samples are elevated in Na and Cl relative to Ca and Mg, and Group II waters exhibit the opposite trend. Additional examination of these same 2 groups by Reyes (2014) showed they are isotopically distinct and likely from different sources. Our approach will adopt this basic interpretation in the subsequent analyses.

2.1 Scatterplots

One of the primary tools for basic water interpretation is scatterplots, wherein individual parameters or ions are plotted against one another. While such plots are commonly applied, they have potential to provide spurious or incorrect results (Filzmoser and others, 2010; Engle and Rowan 2013). An ideal tool to replace scatterplots is conversion of the ions to an isometric log-ratio (ilr). However, because a D -part composition is reduced to $D-1$ ilr coordinates, a comparison of two parameters of

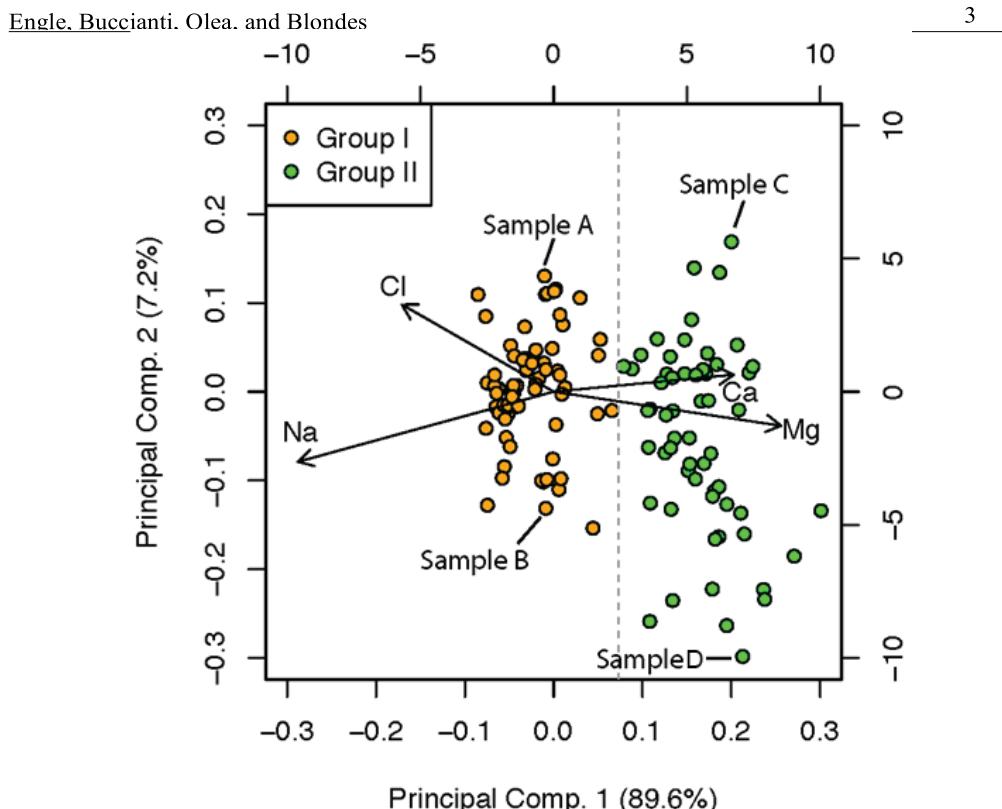


Figure 1: Robust PCA biplot of the subcomposition [Ca, Cl, Na, Mg]. The data were visually parsed into the groups, using the vertical dashed gray line as the divider.

interest is replaced by a single value, which may not be particularly insightful. One approach around this problem is to consider the concentration of water (calculated as density – total dissolved solids) in the samples as the 3rd part (Otero and others, 2005), allowing for the creation of two ilr coordinates (z_1 and z_2) from two ions:

$$\text{ilr}(\mathbf{x}_3) = (z_1, z_2) = \left(\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}, \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{\sqrt{x_1 x_2}}{\text{H}_2\text{O}} \right),$$

where \mathbf{x}_3 is the 3 part composition [x_1 , x_2 , H₂O], and x_1 and x_2 are compositional parameters identified in both water samples.

Using Ca and SO₄ and x_1 and x_2 , respectively, one can quickly gather that the origin and reaction pathway of the two groups of water are dramatically different (Figure 2). The Group I data consistently have Ca/SO₄ molar ratios <1, but approach a value of zero with increasing salinity. This is consistent with a fluid which may have originally been very sodic (Figure 1, Group 1), but increased its salinity through the dissolution of evaporite minerals, including anhydrite (which has a Ca/SO₄ ratio of 1). Also added to the plot are geochemically modeled anhydrite solubility thresholds for pure water at 25 °C and a 200 g/L NaCl solution at 50 °C (see Engle and Blondes, 2014 for details). The Group I waters, when plotting closest to 0 on the y-axis, plot beyond the anhydrite solubility threshold for pure water, suggesting that these correspond to high salinities where anhydrite solubility is enhanced. Conversely, the data for Group II samples (high Ca & Mg, low Na & Cl concentrations) exhibit a nearly random distribution on the plot suggesting a more varied origin and geochemical evolution. Similar types of plots can be made for any pair of solutes and allow for basic interpretation and addition of thermodynamics or other critical measurements.

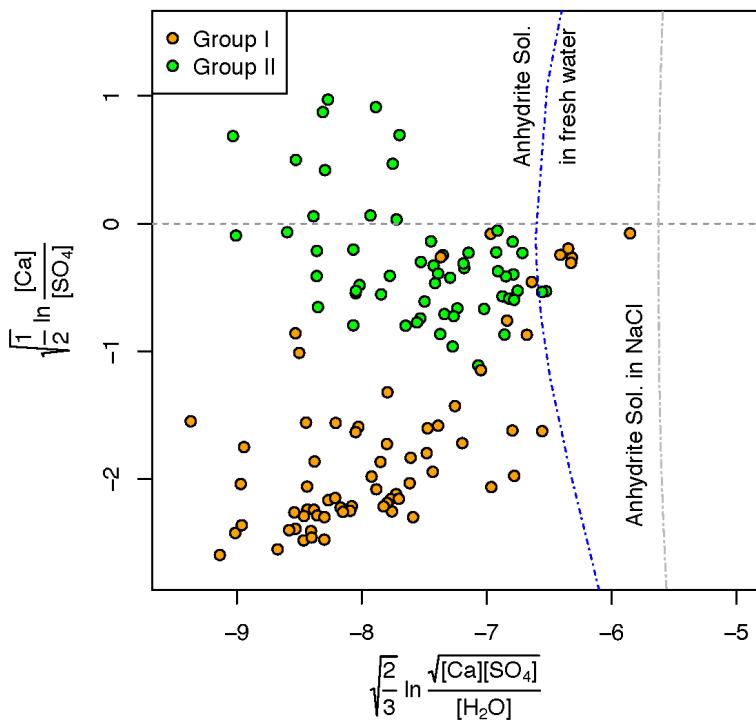


Figure 2: Isometric log-ratio converted data for the subcomposition $[Ca, SO_4, H_2O]$. Solubility increases to the right.

2.2 Trilinear diagrams

Trilinear plots are very common in water quality data analysis, either as stand-alone plots, or as parts of other plots (e.g., Piper plots and Durov diagrams). Conversion of data used in trilinear diagrams to ilr-based scatterplots is trivial and simple, given that ilr transformation of a 3-part composition provides 2-ilr coordinates. As is typical for a Piper or Durov plot, we examine subcompositions of major cations $[Ca, Mg, Na]$ and anions $[Cl, HCO_3, SO_4]$. To adequately deal with missing or zero values in the carbonate species, CO_3 and HCO_3 were converted to HCO_3 based on total inorganic carbon content and sample pH. Alternatively, a geochemical model could have been applied to estimate concentrations of both CO_3 and HCO_3 which were originally reported as zero. These relatively simple approaches provide additional methods to handle censored or below detection limit data beyond some of the more common imputation methods utilized in CoDa (Palarea-Albaladejo and others, 2014).

Comparison of the cations highlights the fact that data at the edges of trilinear diagrams become compressed and trends can be difficult to identify (Figure 3A). For instance, while both plots show clear separation of the two groups of samples, the trilinear diagram suggests that the Ca/Mg of the Group I sample is nearly constant, and much less variable than that of the Group II samples (Figure 3). However, the scatterplot of the same ilr-transformed variables shows that while the variance of the Ca/Mg is generally lower for the Group I samples; 5 or more of the Group I samples have unusually low ratios which may necessitate further investigation. Examination of the anion data in the trilinear diagram shows few obvious trends. However the positive slope of the data in the corresponding scatterplot (particularly for the Group I samples in Figure 2) indicates that there is a geochemical evolution of the samples and as the Cl/SO₄ ratio increases, there is a corresponding increase in HCO₃ relative to Cl and SO₄. One possible reason for this trend is that carbonate solubility increases as a function of NaCl salinity, suggesting that high TDS conditions may induce carbonate dissolution. Such results suggest that

Engle, Buccianti, Olea, and Blondes

5

the anions are linked in a way not identified in the trilinear diagrams. Beyond these individual figures, the scatterplots can be combined and extended to mimic the structure of Piper plots and Durov diagrams to allow for examining the corresponding changes between anions and cations (Engle and Rowan, 2014).

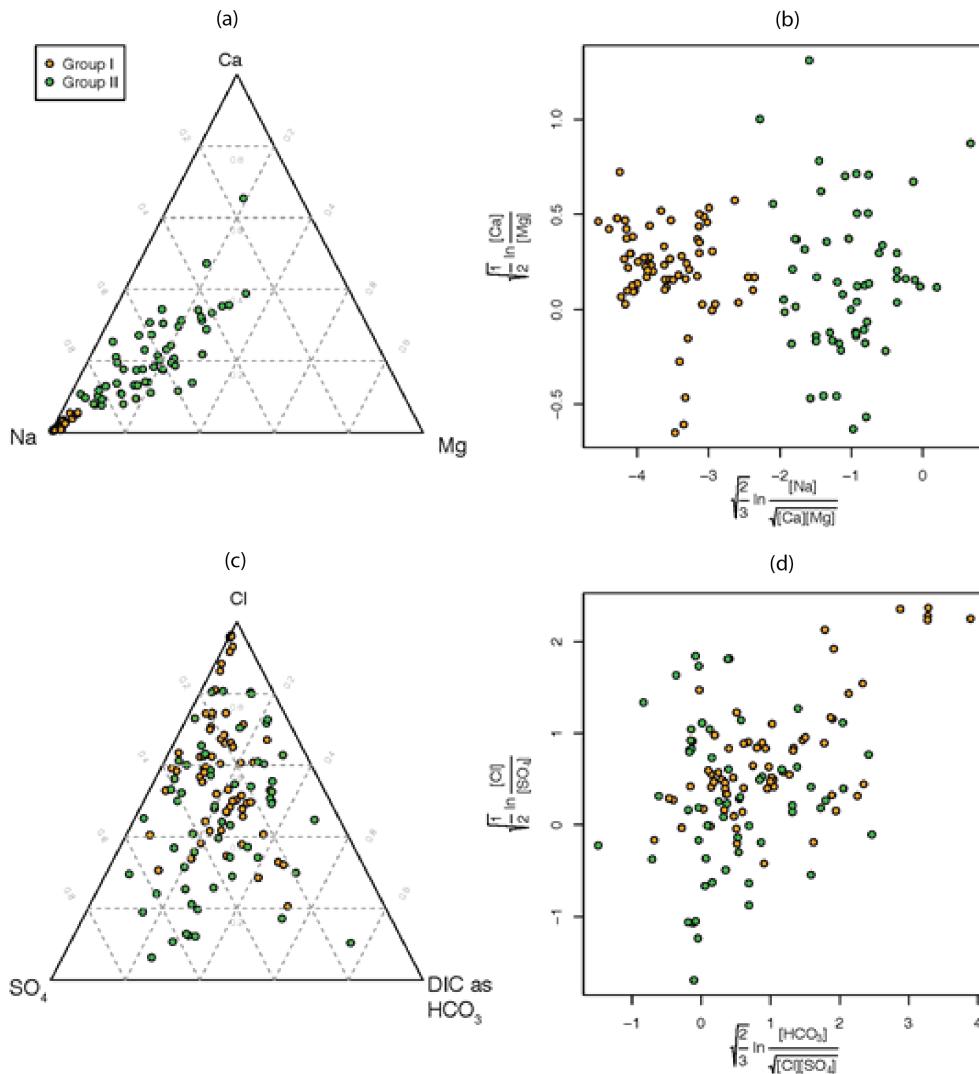


Figure 3: Comparison of trilinear diagrams with scatterplots of the same ilr-transformed variables for major cations (top) and major anions (bottom).

2.3 Stiff diagrams

Stiff diagrams create polygons for each sample that capture the major ion chemistry and can quickly allow for identification of water types (Hem 1985) based on the absolute equivalents of individual ions (Figure 4). Buccianti (2013) proposed a similar type of plot, wherein the absolute equivalents of the ions are replaced by ilr coordinates of three different subcompositions. In this case, we modified their approach slightly by using the following coordinates:

$$z_1 = \frac{1}{\sqrt{2}} \ln \frac{\text{Cl}}{\text{Na}},$$

$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{\text{HCO}_3}{\text{Ca}},$$

and $z_3 = \frac{1}{\sqrt{2}} \ln \frac{\text{SO}_4}{\text{Mg}}$.

The three different coordinates are spread out along the y-axis to create a polygon (Figure 5) that is consistent with the sequence of the pairs in the Stiff diagram (i.e., z_1 on top, z_2 in the middle and z_3 in the bottom). Using this modified version of the ilr balances, negative values are produced when the cation concentration exceeds the anion concentration in milliequivalents, causing the polygons to protrude to the left. This mimics the nature of the Stiff diagrams, where cations plot to the left of zero. For this exercise, data for 4 water samples (samples A, B, C, and D) that represent end-members of the two groups were chosen from Figure 1. A basic Stiff diagram quickly shows that the major ion composition and salinity of the four samples are substantially different. However, reliance upon absolute values makes relative differences less obvious, especially for the lower salinity samples (e.g., Sample C). Examination the ilr version of the plot shows differences between the samples but also some additional insights. For instance, the Cl/Na ratio of Sample A is close to 1 (zero on the axis, due to the natural log), suggesting input from halite dissolution. Similarly, sample D has a value near zero for z_2 , suggesting that calcite dissolution may be its primary source of Ca and HCO_3 despite the concentrations of both being relatively low. Also of note, the absolute concentrations do not affect the scale of the polygons, thus lower salinity samples are as easy to examine as higher salinity samples. Moreover, the ilr version of the plots can be collocated with sample locations on maps to allow for simple spatial analysis and identification of water quality types, as is typically done for Stiff diagrams. This allows for additional uses of graphical methods.

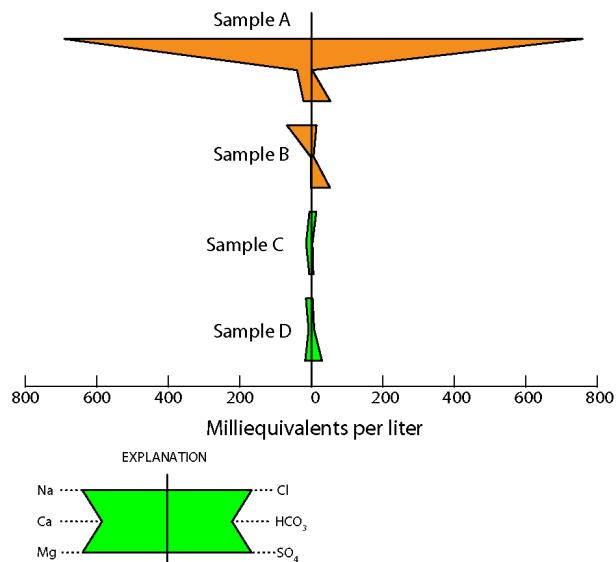


Figure 4: Stiff diagram showing milliequivalents per liter of major anions and cations for 4 individual water samples.

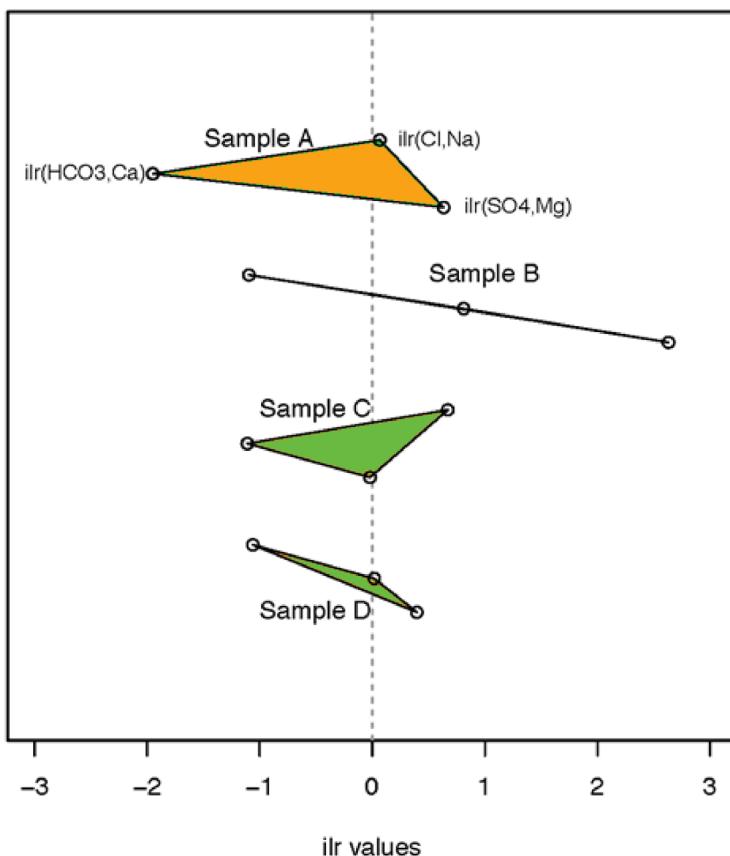


Figure 5: Isometric log-ratio alternative to a Stiff diagram (Buccianti 2013).

3 Conclusions

Since 2005, several new developments in the application of CoDa to water quality data have been published. Here we provided some suggested updates to three conventional water quality plots (i.e., scatterplots, trilinear diagrams, and Stiff diagrams) which utilize the ilr transformation and provide equal, if not more insightful, information than the original figures. Ability to create custom ilr coordinates has allowed for linkages to stoichiometry, saturation products, and solute sources/end-members. As CoDa techniques for additional parameters are developed it will only continue to add to these available tools. While many advances in understanding and process identification have been made using CoDa techniques, simple, intuitive approaches still need to be developed and conveyed.

Acknowledgements

This research was funded by the U.S. Geological Survey Energy Resources Program (Walter Guidroz, Program Coordinator).

References

- Blake, S., T. Henry, J. Murray, R. Flood, M. R. Muller, A. G. Jones, and V. Rath. 2016. "Compositional Multivariate Statistical Analysis of Thermal Groundwater Provenance: a Hydrogeochemical Case Study From Ireland." *Applied Geochemistry* 75 (C). Elsevier Ltd: 171–88. doi:10.1016/j.apgeochem.2016.05.008.
- Blondes, M. S., M. A. Engle, and N. J. Geboy. 2016. "A Practical Guide to the Use of Major Elements, Trace Elements, and Isotopes in Compositional Data Analysis: Applications for Deep Formation Brine Geochemistry ." In *Compositional Data Analysis*, edited by J A Martín-Fernández and S Thió-Henestrosa, 13–29. doi:10.1007/978-3-319-44811-4_2.
- Buccianti, A.. 2011. "Isometric Log-Ratio Co-ordinates and Their Simple Use in Water Geochemistry." *Boletín Geológico Y Minero* 122 (4): 453–58.
- Buccianti, A. 2013. "Is Compositional Data Analysis a Way to See Beyond the Illusion?." *Computers & Geosciences* 50 (C). Elsevier: 165–73. doi:10.1016/j.cageo.2012.06.012.
- Buccianti, A. 2015. "The FOREGS Repository: Modelling Variability in Stream Water on a Continental Scale Revising Classical Diagrams From CoDa (Compositional Data Analysis) Perspective." *Journal of Geochemical Exploration* 154 IS - (July): 94–104.
- Buccianti, A., and R Magli. 2011. "Metric Concepts and Implications in Describing Compositional Changes for World River's Water Chemistry." *Computers & Geosciences* 37 (5): 670–76. doi:doi: DOI: 10.1016/j.cageo.2010.04.017.
- Buccianti, A., and V Pawlowsky-Glahn. 2005. "New Perspectives on Water Chemistry and Compositional Data Analysis." *Mathematical Geology* 37 (7): 703–27. doi:10.1007/s11004-005-7376-6.
- Buccianti, A., and R Zuo. 2016. "Weathering Reactions and Isometric Log-Ratio Coordinates: Do They Speak to Each Other?." *Applied Geochemistry* 75: 189–99. doi:10.1016/j.apgeochem.2016.08.007.
- Buccianti, A., B. Nisi, J. A. Martín-Fernández, and J. Palarea-Albaladejo. 2014. "Methods to Investigate the Geochemistry of Groundwaters with Values for Nitrogen Compounds Below the Detection Limit." *Journal of Geochemical Exploration* 141 (C). Elsevier B.V.: 78–88. doi:10.1016/j.gexplo.2014.01.014.
- Engle, M. A., and M. S. Blondes. 2014. "Linking Compositional Data Analysis with Thermodynamic Geochemical Modeling: Oilfield Brines From the Permian Basin, USA." *Journal of Geochemical Exploration* 141. Elsevier B.V.: 61–70. doi:10.1016/j.gexplo.2014.02.025.
- Engle, M. A., and E. L. Rowan. 2013. "Interpretation of Na-Cl-Br Systematics in Sedimentary Basin Brines: Comparison of Concentration, Element Ratio, and Isometric Log-Ratio Approaches." *Mathematical Geosciences* 45: 87–101. doi:10.1007/s11004-012-9436-z.
- Engle, M. A., and E L Rowan. 2014. "Geochemical Evolution of Produced Waters From Hydraulic Fracturing of the Marcellus Shale, Northern Appalachian Basin: a Multivariate Compositional Data Analysis Approach." *International Journal of Coal Geology* 126: 45–56. doi:10.1016/j.coal.2013.11.010.
- Engle, M. A., F. R. Reyes, M. S. Varonka, W. H. Orem, L. Ma, A. J. Ianno, T. M. Schell, P. Xu, and K. C. Carroll. 2016. "Geochemistry of Formation Waters From the Wolfcamp and 'Cline' Shales: Insights Into Brine Origin, Reservoir Connectivity, and Fluid Flow in the Permian Basin, USA." *Chemical Geology* 425 (May): 76–92. doi:10.1016/j.chemgeo.2016.01.025.
- Engle, M. A., M. Gallo, K. T. Schroeder, N. J. Geboy, and J. W. Zupancic. 2014. "Three-Way Compositional Analysis of Water Quality Monitoring Data." *Environmental and Ecological Statistics* 21 (January): 565–81. doi:10.1007/s10651-013-0268-x.
- Filzmoser, P., K. Hron, and C. Reimann. 2010. "The Bivariate Statistical Analysis of Environmental (Compositional) Data." *Science of the Total Environment* 408 (19): 4230–38. doi:10.1016/j.scitotenv.2010.05.011.
- Gallo, M and A. Buccianti. 2013. "Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy)." *Environmetrics* 24: 269–277.
- Hem, John David. 1985. "Study and Interpretation of the Chemical Characteristics of Natural Water.". U. S. Geological Survey Water-Supply Paper 2254.
- Otero, N., R. Tolosana-Delgado, A. Soler, V. Pawlowsky-Glahn, and A. Canals. 2005. "Relative vs. Absolute

- Engle, Buccianti, Olea, and Blondes 9
Statistical Analysis of Compositions: a Comparative Study of Surface Waters of a Mediterranean River.”
Water Research 39 (7): 1404–14. doi:10.1016/j.watres.2005.01.012.
- Palarea-Albaladejo, J., J. A. Martín-Fernández, and A. Buccianti. 2014. “Compositional Methods for Estimating Elemental Concentrations Below the Limit of Detection in Practice Using R.” *Journal of Geochemical Exploration* 141 (C). Elsevier B.V.: 71–77. doi:10.1016/j.gexplo.2013.09.003.
- Reyes, F R. 2014. “Exploring the Hydrogeologic Controls on Brackish Water and Its Suitability for Use in Hydraulic Fracturing: the Dockum Aquifer, Midland Basin, Texas.” M.S. Thesis, University of Texas at El Paso.
- Tolosana-Delgado, R, N Otero, and A Soler. 2005. “A Compositional Approach to Stable Isotope Data Analysis.” In,CoDaWork 2005, 1–11.

Differential proportionality – a normalization-free approach to differential gene expression

I. Erb^{1,2,*}, T. Quinn³, D. Lovell⁴, and C. Notredame^{1,2}

¹Centre for Genomic Regulation (CRG),

The Barcelona Institute for Science and Technology, Barcelona, Spain;

² Universitat Pompeu Fabra (UPF), Barcelona, Spain; *ionas.erb@crg.eu

³Bioinformatics Core Research Group, Deakin University, Geelong, Victoria, Australia

⁴Queensland University of Technology, Brisbane, Queensland, Australia

Abstract

Gene expression data, such as those generated by next generation sequencing technologies (RNA-seq), are of an inherently relative nature: the total number of sequenced reads has no biological meaning. This issue is most often addressed with various normalization techniques which all face the same problem: once information about the total mRNA content of the origin cells is lost, it cannot be recovered by mere technical means. Additional knowledge, in the form of an unchanged reference, is necessary; however, this reference can usually only be estimated. Here we propose a novel method where sample normalization is unnecessary, but important insights can be obtained nevertheless. Instead of trying to recover absolute abundances, our method is entirely based on ratios, so normalization factors cancel by default. Although the differential expression of individual genes cannot be recovered this way, the ratios themselves can be differentially expressed (even when their constituents are not). Yet, most current analyses are blind to these cases, while our approach reveals them directly. Specifically, we show how the differential expression of gene ratios can be formalized by decomposing log-ratio variance (LRV) and deriving intuitive statistics from it. Although small LRVs have been used to detect proportional genes in gene expression data before, we focus here on the change in proportionality factors between groups of samples (e.g. tissue-specific proportionality). For this, we propose a statistic that is equivalent to the squared *t*-statistic of one-way ANOVA, but for gene ratios. In doing so, we show how precision weights can be incorporated to account for the peculiarities of count data, and, moreover, how a moderated statistic can be derived in the same way as the one following from a hierarchical model for individual genes. We also discuss approaches to deal with zero counts, deriving an expression of our statistic that is able to incorporate them. In providing a detailed analysis of the connections between the differential expression of genes and the differential proportionality of pairs, we facilitate a clear interpretation of new concepts. The proposed framework is applied to a data set from GTEx consisting of 98 samples from the cerebellum and cortex, with selected examples shown. An R package containing a computationally efficient implementation of the approach is in preparation and will be released shortly as an addendum to the propr package.

Key words: Differential gene expression, sample normalization, proportionality, count ratios, moderated statistics, covariance regularization, count zeros.

1 Introduction

Normalization techniques for transcriptome sequencing data continues to be of high interest to the data analysis community (e.g. see (Dillies et al., 2013) for a review and (Lun, A. et al., 2016) for a recent example in single-cell RNA-seq). For sample normalization between entirely different conditions, however, ever more sophisticated techniques cannot close the knowledge gap that is of a principal nature: the total mRNA content of the cells of origin is unknown and can only be obtained with an appropriate ‘absolute’ technique.

It has been argued that normalizations can be avoided by performing a log-ratio transformation of the data (Fernandes et al., 2013; Lovell et al., 2015). Such data transformations, however, depend on the reference that is used. The danger here is that the resulting transformed data is ultimately interpreted in a gene-wise fashion. Interpreting log-ratio transformed expression data as referring to gene abundances (instead of ratios with respect to a given reference) runs into the exact same problems as using normalizations (Erb and Notredame, 2016). It effectively means that the log-ratio transformation is seen as a normalization (that has, as it were, an additional aura of technical sophistication). The only way out of this dilemma seems to be to let go of the gene-wise perspective entirely and instead consider ratios as the basic objects of interest. Although some information will remain hidden this way (such as the true differential gene expression between absolute abundances), the remaining signal will be inherently unbiased.

Here we propose a formal framework for understanding *differential ratio expression*, a change in the ratio of abundances between experimental groups. In doing so, we show that techniques developed for the analysis of the differential expression of genes (e.g. methods known from the limma/voom approach (Smyth, 2004; Smyth, 2005; Law et al., 2014) apply to the analysis of differential ratios as well. This seems intuitive when considering gene ratios as depicted in Figure 1D: an identical picture could be obtained using read counts of a differentially expressed gene instead of gene ratios as shown. However, the interpretation of differential ratios differs considerably.

First, we must consider what it means for a gene ratio to remain unchanged across all sample data. The answer is that the two genes change in the same way (or otherwise remain both unchanged). Figure 1A shows this case in a scatter plot of the read counts for two genes (a splicing factor and a polymerase subunit). Note that although the gene ratio may remain the same, the genes themselves could have joint differential expression. Such gene-wise differential expression is not detected by the ratio approach: although the two genes appear differentially expressed between the tissues, their approximately constant ratio, as shown in Figure 1B, does not reveal this. However, without knowing absolute mRNA abundances, genes may appear differentially expressed only as an artifact of their relative nature.

Second, we must consider what it means for a gene ratio to differ between experimental groups. Figures 1C and 1D shows an example of tissue-specific gene ratios. Here, the two genes (the same splicing factor as before and a kinase) are correlated in both tissues (with a similar strength of correlation), but with different slopes. This means their proportionality factor is tissue-specific (i.e. they have *differential proportionality*). In terms of biochemistry, this could indicate a change in the stoichiometry of the protein products resulting from these mRNAs. Preliminary GO-category enrichment analyses support this view, showing that differentially proportional pairs often contain genes that form protein complexes like those involved in transcription or ribosomal activity.

Current standard methods are not tailored to infer differentially proportional pairs (c.f., Figure 3), although a special class of them, involving receptor subunits in the human brain, has been found by considering time-dependent correlations (Bar-Shira et al., 2015). One method, differential correlation (Tesson et al., 2010), is concerned with differential correlation coefficients, but not with the differential slopes of linear relationships. Importantly, current methods always include a normalization step that—in the best case scenario—introduces extra noise, thus reducing efficacy compared with a method that picks up such signals directly.

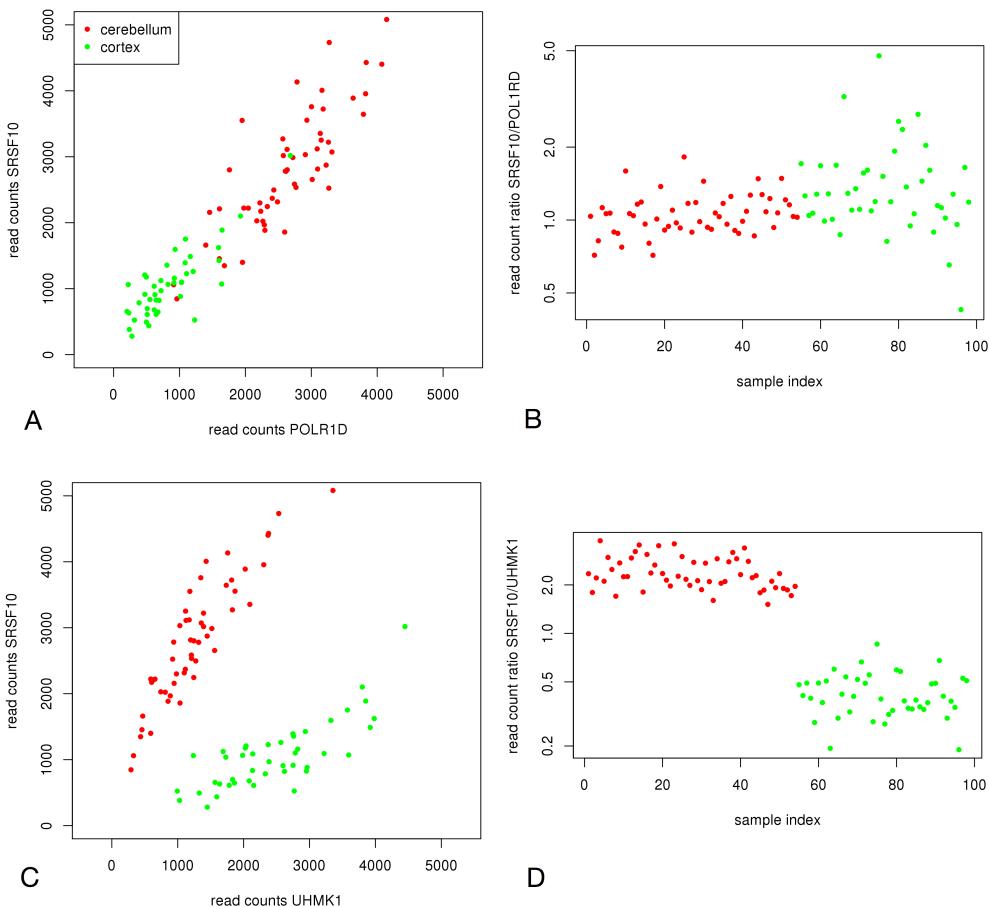


Figure 1: Constant and changing ratios across 98 samples from two tissues: (A) Scatter plot of two genes having an approximately constant read count ratio across all samples (i.e. proportional genes). (B) Ratio plot of the same two genes as in panel A. Although panel A suggests their differential expression, ratios are unable to reveal it. (C) Example of *differentially proportional* genes. Their correlation appears to be about equally strong in both tissues, but the slope of their linear relationship changes between the tissues. (D) Ratio plot of the same two genes as in panel C. The tissue-specific proportionality factors can be detected clearly, and the picture suggests that conventional methods of differential gene expression can be applied to ratios as well.

2 Methods and Results

2.1 Simple statistics for differential proportionality

We start by introducing a short-hand notation which allows us to denote projections of the log-ratios of two vectors \mathbf{x}, \mathbf{y} having n components (e.g. a gene or transcript pair) onto a subset of size k :

$$\mathbf{L}_{1,\dots,k}^{\mathbf{x},\mathbf{y}} := \left(\log \frac{x_1}{y_1}, \dots, \log \frac{x_k}{y_k} \right). \quad (1)$$

Equivalently, the log-ratio mean (LRM) and variance (LRV) evaluated on this subset are denoted by $E(\mathbf{L}_{1,\dots,k}^{\mathbf{x},\mathbf{y}})$ and $\text{var}(\mathbf{L}_{1,\dots,k}^{\mathbf{x},\mathbf{y}})$ respectively. Let us now assume we have a natural partition of our n samples into two subsets (conditions, or tissues) of experimental replicates of sizes k and $n - k$. To avoid clutter, we drop \mathbf{x}, \mathbf{y} from the notation in the following equation. It is well known that variance evaluates to

$$\begin{aligned} \text{var}(\mathbf{L}_{1,\dots,n}) &= E(\mathbf{L}_{1,\dots,n}^2) - E^2(\mathbf{L}_{1,\dots,n}) \\ &= \frac{kE(\mathbf{L}_{1,\dots,k}^2) + (n - k)E(\mathbf{L}_{k+1,\dots,n}^2)}{n} - \frac{(kE(\mathbf{L}_{1,\dots,k}) + (n - k)E(\mathbf{L}_{k+1,\dots,n}))^2}{n^2} \\ &= \frac{kE^2(\mathbf{L}_{1,\dots,k}) + (n - k)E^2(\mathbf{L}_{k+1,\dots,n})}{n} + \frac{k\text{var}(\mathbf{L}_{1,\dots,k}) + (n - k)\text{var}(\mathbf{L}_{k+1,\dots,n})}{n} \\ &\quad - \frac{(kE(\mathbf{L}_{1,\dots,k}) + (n - k)E(\mathbf{L}_{k+1,\dots,n}))^2}{n^2} \\ &= \frac{k(n - k)}{n^2} (E(\mathbf{L}_{1,\dots,k}) - E(\mathbf{L}_{k+1,\dots,n}))^2 + \frac{k\text{var}(\mathbf{L}_{1,\dots,k}) + (n - k)\text{var}(\mathbf{L}_{k+1,\dots,n})}{n}. \end{aligned} \quad (2)$$

This is the well-known decomposition into between-group variance (first term) and within-group variance (second term) known from analysis of variance (ANOVA). Note that all variances throughout the text are defined as the biased estimators (so the sum of squares are divided by k rather than $k - 1$, with k the number of summands). As will be seen from the discussion below, differential proportionality can be studied relative to LRV and there is no need for evaluation of the total size of LRV (which is a problem when studying proportionality across all the samples). If we divide (2) by $\text{var}(\mathbf{L}_{1,\dots,n})$, we obtain as summands the various proportions of (weighted) group variances and of the between-group variance to the overall variance. For illustration, this is visualized as a ternary diagram in Figure 2A. The proportion of within-group variance with respect to overall variance is thus a function of the three LRVs:

$$\vartheta(\mathbf{x}, \mathbf{y}) = \frac{k\text{var}(\mathbf{L}_{1,\dots,k}^{\mathbf{x},\mathbf{y}}) + (n - k)\text{var}(\mathbf{L}_{k+1,\dots,n}^{\mathbf{x},\mathbf{y}})}{n\text{var}(\mathbf{L}_{1,\dots,n}^{\mathbf{x},\mathbf{y}})}. \quad (3)$$

Conveniently, ϑ is a number between zero and one. When approaching zero it indicates that the total LRV is explained by the squared difference in group LRMs (Fig. 2B). A large enough difference means that scatter plots of \mathbf{y} vs. \mathbf{x} will have different slopes depending on the condition the samples come from. This case is thus characterized by tissue-specific proportionality factors (or group LRMs). We call this type of differential proportionality *disjointed* proportionality here.

We can use ϑ for testing this property on our vector pairs and evaluate its significance using a simple permutation test for an estimate of the false discovery rate (FDR). Alternatively, a classical test-statistic known from one-way ANOVA with two groups is the squared t -statistic F . It is related to ϑ by

$$F = (n - 2) \frac{(1 - \vartheta)}{\vartheta}. \quad (4)$$

This statistic can be used to do a classical F -test of the null hypothesis of equal group (population) LRMs under standard ANOVA assumptions. Note that regardless of the statistic used, multiple testing corrections are especially important in the ratio context due to the large number of gene

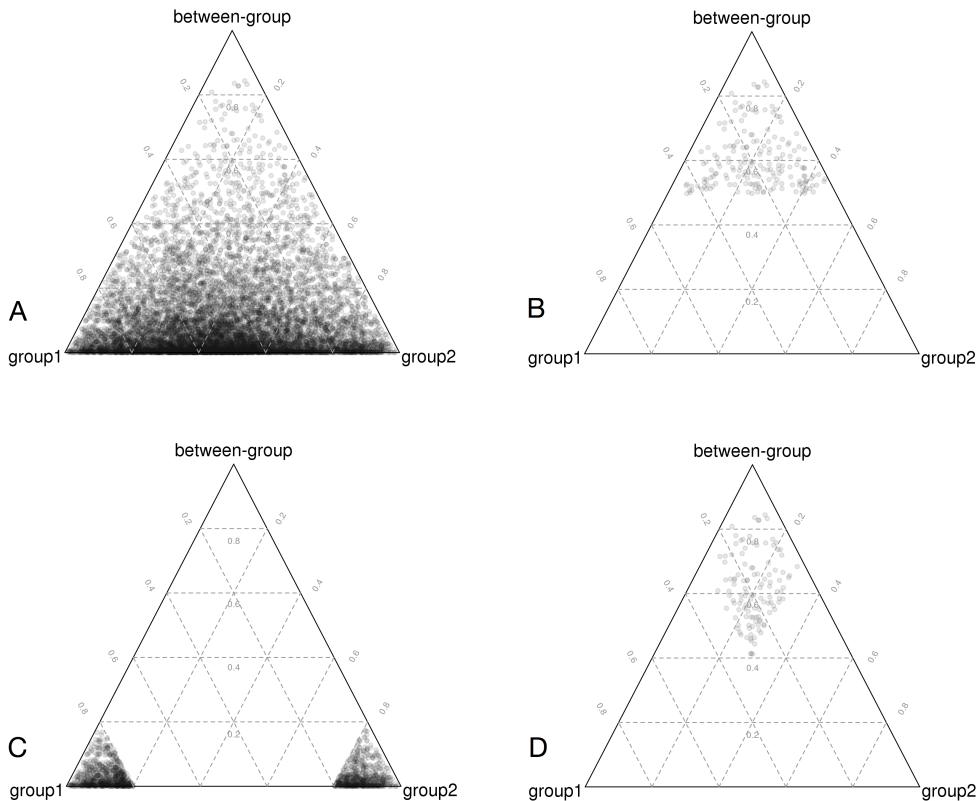


Figure 2: Decomposition of log-ratio variance into (weighted) group variances and between-group variance shown in ternary diagrams. Data from our example from GTEx (group 1 cerebellum, group 2 cortex) are shown. For better visibility, a subset of 10,000 randomly sampled gene pairs were selected. (A): The 10,000 dots corresponding to LRVs of each gene pair. (B): Gene pairs fulfilling $\vartheta < 0.5$ (disjointed proportionality). (C): Gene pairs fulfilling $\vartheta_e < 0.2$ (emergent proportionality). (D): Gene pairs fulfilling $\vartheta_e > 0.7$. Such cut-offs from below induce a cut-off on ϑ and an additional restriction on the difference between weighted group variances.

pairs that get tested. These can be efficiently obtained by estimating the FDR, such as by using the plug-in estimate from a permutation procedure (see e.g. Hastie et al. 2013).

We have seen that disjointed proportionality describes pairs where between-group variance constitutes the major part of their LRV. Another type of differential proportionality can be defined for those pairs where one of the group LRVs dominates the total LRV. A scatter of \mathbf{y} vs. \mathbf{x} will then show proportionality for samples in one condition but no correlation for the other condition. We will call this type of proportionality *emergent* to distinguish it from disjointed proportionality. In complete analogy to the definition of ϑ , from (2) we get

$$\vartheta_1(\mathbf{x}, \mathbf{y}) = \frac{n\text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{y}} - k\text{var } L_{1,\dots,k}^{\mathbf{x},\mathbf{y}}}{n\text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{y}}}, \quad (5)$$

as the proportion of the sum of between-group variance and the LRV of group 2 to the total LRV. Small values of ϑ_1 indicate that the LRV of group 1 constitutes the major part of the total LRV, which is our defining feature of emergent proportionality in group 2. A convenient measure for detecting emergent proportionality regardless of group can be defined as

$$\vartheta_e(\mathbf{x}, \mathbf{y}) = 1 - \frac{\max(k\text{var } L_{1,\dots,k}^{\mathbf{x},\mathbf{y}}, (n-k)\text{var } L_{k+1,\dots,n}^{\mathbf{x},\mathbf{y}})}{n\text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{y}}}, \quad (6)$$

of which a cut-off from above will give us the a set of pairs that are proportional in just one of the two conditions (Fig. 2C). Let us now look at the relationship between ϑ_e and ϑ . Note that we have

$$\vartheta_e = 1 - \vartheta + \frac{\min(k\text{var } L_{1,\dots,k}^{\mathbf{x},\mathbf{y}}, (n-k)\text{var } L_{k+1,\dots,n}^{\mathbf{x},\mathbf{y}})}{n\text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{y}}}. \quad (7)$$

It follows that

$$1 - \vartheta \leq \vartheta_e \leq 1 - \vartheta/2, \quad (8)$$

with the equality $1 - \vartheta = \vartheta_e$ holding if one of the group LRVs vanishes and $\vartheta_e = 1 - \vartheta/2$ in the case of equality of weighted group LRVs $k\text{var } L_{1,\dots,k}^{\mathbf{x},\mathbf{y}} = (n-k)\text{var } L_{k+1,\dots,n}^{\mathbf{x},\mathbf{y}}$. It transpires that ϑ_e can be used to study both types of differential proportionality since large values of it enforce small ϑ . For this, a second cut-off on ϑ_e , this time from below, needs to be determined. However, note that a cut-off $\vartheta_e > C$ would enforce a somewhat stricter definition on disjointed proportionality, where the induced cut-off $\vartheta < 2(1-C)$ can only be attained for equality of weighted group LRVs, a condition that is relaxed when going further down with ϑ . In fact, cut-offs from below on ϑ_e cut the upper corner of the ternary diagram with two lines that yield a diamond shape as opposed to the triangle that results from a cut-off on ϑ (Fig. 2D). Thus ϑ_e allows for better control of the correlation within the groups. This can be useful when filtering out those differentially proportional pairs that consist of genes having differential expression but which are not proportional within the groups. This case will be discussed in section 2.4.

2.2 Introducing precision weights

RNA-seq data show a pronounced mean-variance relationship that leads to biases when linear models are fit to them. However, log-ratios do not show the mean-variance relationship of the counts directly. The problem here is rather that we should have less confidence in ratios when they involve low counts, as their precision will be lower due to the mean-variance relationship. It has been suggested that an incorporation of the mean-variance relationship via precision weights makes count data accessible for linear modelling (Law et al., 2014) and weighting in general leads to better benchmark performance (Liu et al., 2015). Here we need weights for log-ratios rather than log counts. We can combine the weights $\omega(x_i)$ for read counts of gene \mathbf{x} in condition i into a ratio weight by simply multiplying the weights of both genes involved. Let us denote these weights by

$$\omega_i^{\mathbf{x},\mathbf{y}} = \omega(x_i)\omega(y_i). \quad (9)$$

The overall weight of a given ratio for the set of samples $1, \dots, k_1$ from condition 1 is then

$$\Omega_1^{\mathbf{x},\mathbf{y}} = \sum_{i=1}^{k_1} \omega_i^{\mathbf{x},\mathbf{y}}. \quad (10)$$

Let us now drop the upper indices for the gene pair. The weighted log-ratio means and variances for a given gene pair in condition 1 will then be

$$E_\omega(L_{1,\dots,k_1}) = \frac{1}{\Omega_1} \sum_{i=1}^{k_1} \omega_i \log \frac{x_i}{y_i}, \quad (11)$$

$$\text{var}_\omega(L_{1,\dots,k_1}) = \frac{1}{\Omega_1} \sum_{i=1}^{k_1} \omega_i \left(\log \frac{x_i}{y_i} - E_\omega(L_{1,\dots,k_1}) \right)^2. \quad (12)$$

The decomposition of weighted log-ratio variance goes through as before, and a weighted statistic

$$\vartheta_\omega = \frac{\Omega_1 \text{var}_\omega L_{1,\dots,k} + \Omega_2 \text{var}_\omega L_{k+1,\dots,k_1+k_2}}{(\Omega_1 + \Omega_2) \text{var}_\omega L_{1,\dots,k_1+k_2}} \quad (13)$$

can be defined in analogy to (3). Here we were just interested in the sums, not the actual variances. Note that we can define a unbiased weighted variance estimator specifically for reliability weights. For this, the prefactor in (12) changes from $1/\Omega_1$ to $1/(\Omega_1 - \sum \omega_i^2 / \Omega_1)$.

2.3 A moderated statistic for ratios

It has been shown that similarities in expression between the genes can be exploited by assuming an underlying prior distributions of within-group variances and log-fold changes in a gene-expression matrix (Lönnstedt and Speed, 2001; Smyth, 2004). The resulting hierarchical model can be used to derive a moderated t -statistic whose parameters can be estimated from the data in empirical-Bayes fashion. The moderated statistic has been shown to be much more powerful than the classical t -statistic in simulation-based benchmarks, see (McCarthy and Smyth, 2009). The moderation effectively adds a small amount to the within-group variance of a gene and could thus be understood as the regularization of a covariance matrix, see e.g. (Witten and Tibshirani, 2009). Here we show how a moderated statistic can be derived for ratios starting from the hierarchical model of their constituent genes.

First note that the information contained in all pairwise ratios is highly redundant, and only the ratios with respect to a given reference (in form of one specific gene or the geometric mean of all the genes) are necessary to recover all the variation in the data set (Aitchison, 2003). If we denote this reference by \mathbf{z} , the following equation shows how an arbitrary log-ratio variance can be written in terms of the covariance matrix of ratios with respect to this reference:

$$\text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{y}} = \text{var } L_{1,\dots,n}^{\mathbf{x},\mathbf{z}} + \text{var } L_{1,\dots,n}^{\mathbf{y},\mathbf{z}} - 2\text{cov}(L_{1,\dots,n}^{\mathbf{x},\mathbf{z}}, L_{1,\dots,n}^{\mathbf{y},\mathbf{z}}). \quad (14)$$

Retaining the subset of ratios with respect to a reference is known as log-ratio transformation (of the alr type in case \mathbf{z} is a gene, of the clr type in case \mathbf{z} is the geometric mean of the genes). If our reference is unchanged across samples, the transformation results in a normalization of the gene expressions, and the resulting ratios are proportional to absolute expressions. For these kind of data the hierarchical model was derived in (Lönnstedt and Speed, 2001; Smyth, 2004). Note, however, that here we do not require any particular properties of the reference. Let us denote the pooled within-group variance of the log-ratios with reference \mathbf{z} by

$$s_{\mathbf{x},\mathbf{z}}^2 = \frac{k \text{var } L_{1,\dots,k}^{\mathbf{x},\mathbf{z}} + (n-k) \text{var } L_{k+1,\dots,n}^{\mathbf{x},\mathbf{z}}}{n}. \quad (15)$$

Given the hierarchical model, it was shown that the posterior mean of the inverse population variance $\sigma_{\mathbf{x}, \mathbf{z}}^{-2}$, given the sample variance (15), has the form

$$\tilde{s}_{\mathbf{x}, \mathbf{z}}^{-2} = \frac{d_{\mathbf{z}} + n}{d_{\mathbf{z}} s_{\mathbf{z}}^2 + n s_{\mathbf{x}, \mathbf{z}}^2}, \quad (16)$$

where $d_{\mathbf{z}}$ and $s_{\mathbf{z}}^2$ are the parameters of the Gamma distribution serving as a prior for the variance (15). We will not go into more detail of the underlying Bayesian model here but just mention that a moderated t -statistic can be obtained by replacing $s_{\mathbf{x}, \mathbf{z}}^2$ in the original t -statistic by $\tilde{s}_{\mathbf{x}, \mathbf{z}}^2$. In the following we use (16) as a justification for moderating (adding a small amount to) the within-group variances. This can also be seen as a kind of regularization of the covariance matrix of the log-ratios that have \mathbf{z} as a reference. From (16) we can now derive moderated versions of F and ϑ for all the gene ratios.

Let us denote by F' the ratio of between-group over within-group LRV for a given gene pair. F' is the same as F in Equation (4) without the factor $(n - 2)$. We have

$$F'(\mathbf{x}, \mathbf{y}) = \frac{k(n - k)}{n^2} \frac{\left(E(L_{1, \dots, k}^{\mathbf{x}, \mathbf{y}}) - E(L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{y}}) \right)^2}{s_{\mathbf{x}, \mathbf{y}}^2}. \quad (17)$$

Applying the relationship (14) on both groups of samples separately, the pooled within-group variance can be written as

$$\begin{aligned} s_{\mathbf{x}, \mathbf{y}}^2 &= \frac{k}{n} \left(\text{var } L_{1, \dots, k}^{\mathbf{x}, \mathbf{z}} + \text{var } L_{1, \dots, k}^{\mathbf{y}, \mathbf{z}} - 2\text{cov}(L_{1, \dots, k}^{\mathbf{x}, \mathbf{z}}, L_{1, \dots, k}^{\mathbf{y}, \mathbf{z}}) \right) + \\ &\quad \frac{n - k}{n} \left(\text{var } L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{z}} + \text{var } L_{k+1, \dots, n}^{\mathbf{y}, \mathbf{z}} - 2\text{cov}(L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{z}}, L_{k+1, \dots, n}^{\mathbf{y}, \mathbf{z}}) \right) \\ &= s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2 - 2c_{\mathbf{x}, \mathbf{y}}^{\mathbf{z}}, \end{aligned} \quad (18)$$

where the within-group covariance $c_{\mathbf{x}, \mathbf{y}}^{\mathbf{z}}$, with respect to the reference \mathbf{z} , is defined by

$$c_{\mathbf{x}, \mathbf{y}}^{\mathbf{z}} = \frac{k}{n} \text{cov}(L_{1, \dots, k}^{\mathbf{x}, \mathbf{z}}, L_{1, \dots, k}^{\mathbf{y}, \mathbf{z}}) + \frac{n - k}{n} \text{cov}(L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{z}}, L_{k+1, \dots, n}^{\mathbf{y}, \mathbf{z}}). \quad (19)$$

Returning to (17), we thus have

$$F'(\mathbf{x}, \mathbf{y}) = K \frac{\left(E(L_{1, \dots, k}^{\mathbf{x}, \mathbf{y}}) - E(L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{y}}) \right)^2}{s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2 - 2c_{\mathbf{x}, \mathbf{y}}^{\mathbf{z}}}, \quad (20)$$

where we also used the short-hand expression

$$K = \frac{k(n - k)}{n^2}. \quad (21)$$

The idea is now to replace the terms $s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2$ by their moderated versions derived from (16). We find

$$\tilde{s}_{\mathbf{x}, \mathbf{z}}^2 + \tilde{s}_{\mathbf{y}, \mathbf{z}}^2 = \frac{2d_{\mathbf{z}} s_{\mathbf{z}}^2 + n(s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2)}{d_{\mathbf{z}} + n}. \quad (22)$$

Inserting this into (20) yields a moderated F' :

$$\tilde{F}'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \frac{K \left(E(L_{1, \dots, k}^{\mathbf{x}, \mathbf{y}}) - E(L_{k+1, \dots, n}^{\mathbf{x}, \mathbf{y}}) \right)^2}{\frac{2d_{\mathbf{z}} s_{\mathbf{z}}^2 + n(s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2)}{d_{\mathbf{z}} + n} - 2c_{\mathbf{x}, \mathbf{y}}^{\mathbf{z}}}. \quad (23)$$

The parameters $d_{\mathbf{z}}$ and $s_{\mathbf{z}}^2$ can be determined, e.g. using the limma package (Smyth, 2005). Whether the dependence on the choice of \mathbf{z} is of any practical importance needs to be investigated empirically.

From \tilde{F}' we get immediately the corresponding expressions for \tilde{F} and $\tilde{\vartheta}$ by applying (4):

$$\tilde{F} = \tilde{F}'(n - 2), \quad (24)$$

$$\tilde{\vartheta} = \frac{1}{1 + \tilde{F}}. \quad (25)$$

Although we did not use the weighted variances here for clarity and to ease the notational burden, it is straightforward to derive weighted versions of the moderated statistics applying the precision weights described in the previous section.

2.4 Relation to differential expression

If we assume that we know the identity of an unchanged reference \mathbf{z} , it provides us with an ideal normalization (as mentioned in the previous section). The statistic $\vartheta(\mathbf{x}, \mathbf{z})$ could then be used as a measure for the amount of differential expression of gene \mathbf{x} , whose log-fold change would be

$$b_{\mathbf{x}} = E(L_{1,\dots,k}^{\mathbf{x}, \mathbf{z}}) - E(L_{k+1,\dots,n}^{\mathbf{x}, \mathbf{z}}). \quad (26)$$

We will now show that if we have two sufficiently strong differentially expressed genes whose log-fold changes have opposite signs, then they will form a differentially proportional pair. Hence, no within-group correlations of the genes are required in this case for their ϑ to be small¹. More formally, we assume

$$\vartheta(\mathbf{x}, \mathbf{z}) \leq c, \quad (27)$$

$$\vartheta(\mathbf{y}, \mathbf{z}) \leq c, \quad (28)$$

$$b_{\mathbf{x}} b_{\mathbf{y}} < 0. \quad (29)$$

The log-ratio change of the gene pair \mathbf{x}, \mathbf{y} is

$$\begin{aligned} E(L_{1,\dots,k}^{\mathbf{x}, \mathbf{y}}) - E(L_{k+1,\dots,n}^{\mathbf{x}, \mathbf{y}}) \\ = E(L_{1,\dots,k}^{\mathbf{x}, \mathbf{z}}) - E(L_{1,\dots,k}^{\mathbf{y}, \mathbf{z}}) - E(L_{k+1,\dots,n}^{\mathbf{x}, \mathbf{z}}) + E(L_{k+1,\dots,n}^{\mathbf{y}, \mathbf{z}}) = b_{\mathbf{x}} - b_{\mathbf{y}}. \end{aligned} \quad (30)$$

Using this and (20), we obtain

$$F'(\mathbf{x}, \mathbf{y}) = \frac{K(b_{\mathbf{x}} - b_{\mathbf{y}})^2}{s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2 - 2c_{\mathbf{x}, \mathbf{y}}^2}. \quad (31)$$

Since correlation coefficients have absolute values below one and the arithmetic mean bounds the geometric mean, we have

$$s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2 - 2c_{\mathbf{x}, \mathbf{y}}^2 \leq 2(s_{\mathbf{x}, \mathbf{z}}^2 + s_{\mathbf{y}, \mathbf{z}}^2). \quad (32)$$

Now (17) implies that

$$s_{\mathbf{x}, \mathbf{z}}^2 = \frac{Kb_{\mathbf{x}}^2}{F'(\mathbf{x}, \mathbf{z})} = \frac{Kb_{\mathbf{x}}^2 \vartheta(\mathbf{x}, \mathbf{z})}{1 - \vartheta(\mathbf{x}, \mathbf{z})} \leq \frac{Kb_{\mathbf{x}}^2 c}{1 - c}, \quad (33)$$

with the bound following from our condition (27), and for $s_{\mathbf{y}, \mathbf{z}}^2$ from (28). We can now return to (31) to bound

$$F'(\mathbf{x}, \mathbf{y}) \geq \frac{K(b_{\mathbf{x}} - b_{\mathbf{y}})^2}{2K \frac{c}{1-c} (b_{\mathbf{x}}^2 + b_{\mathbf{y}}^2)} = \frac{b_{\mathbf{x}}^2 + b_{\mathbf{y}}^2 - 2b_{\mathbf{x}} b_{\mathbf{y}}}{2 \frac{c}{1-c} (b_{\mathbf{x}}^2 + b_{\mathbf{y}}^2)} \geq \frac{1 - c}{2c}, \quad (34)$$

with the last bound following from (29). We thus find that (27)-(29) imply differential proportionality in the sense that

$$\vartheta(\mathbf{x}, \mathbf{y}) \leq \frac{2}{1 + 1/c}. \quad (35)$$

¹This means there are at least two kinds of pairs with small ϑ : the ones where genes are proportional within the two groups of samples, and those where both genes are unrelated but differentially expressed individually. The latter have a larger within-group LRV and thus need to compensate with a larger overall LRV.

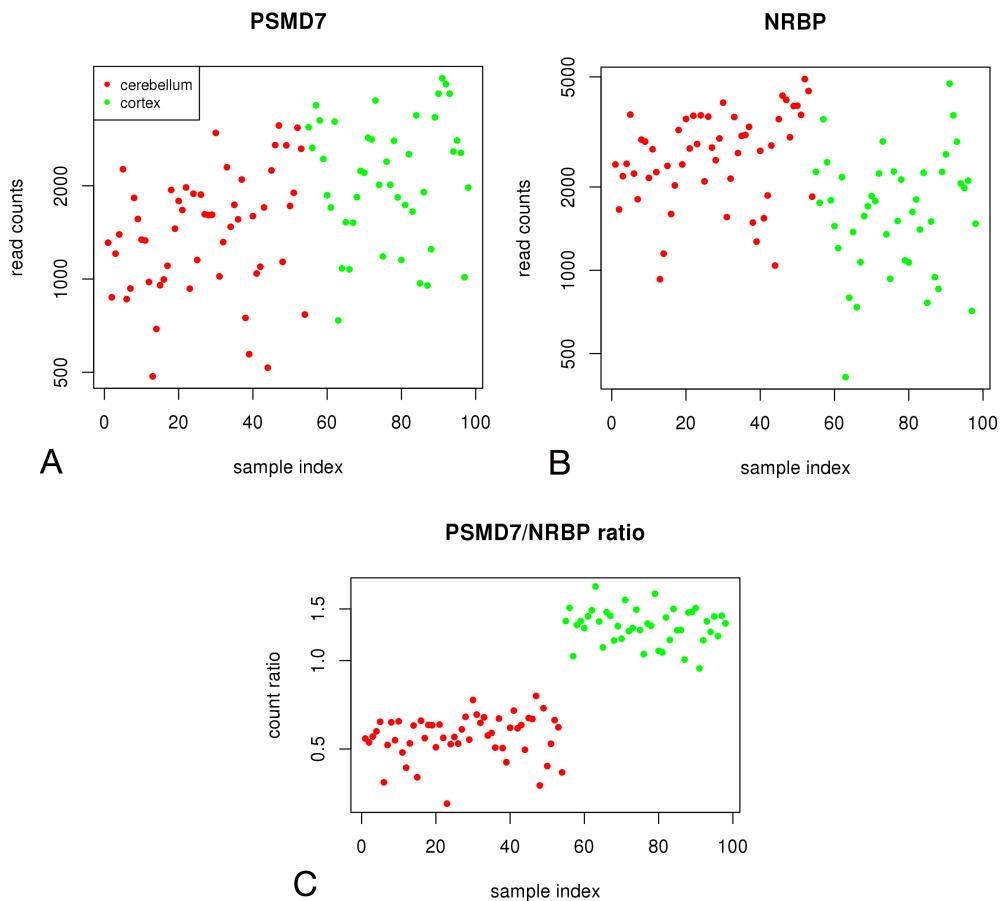


Figure 3: Differential expression of individual genes is not necessary for the pair to be differentially proportional: (A) Read counts plotted against the sample index for the gene PSMD7 (a proteasome subunit). Read counts do not indicate any apparent differences between tissues. (B) A similar situation as in panel A, but for a nuclear receptor binding protein. (C) The ratio plot of the genes from panels A and B. There is a clear difference in the gene ratios, although the individual read counts show no apparent differential expression.

In a similar fashion, more complicated relationships could be derived where the conditions (28) and (29) get relaxed. Instead, we will now look at the reversed question: What can we know about differential expression of the individual genes when the pair is differentially proportional? The only assumption we make is

$$\vartheta(\mathbf{x}, \mathbf{y}) \leq C. \quad (36)$$

Starting from (31), we have

$$\frac{1-C}{C} \leq F'(\mathbf{x}, \mathbf{y}) = \frac{K(b_{\mathbf{x}} - b_{\mathbf{y}})^2}{s_{\mathbf{x}, \mathbf{y}}^2} = \frac{\left(\sqrt{s_{\mathbf{x}, \mathbf{z}}^2 \frac{1-\vartheta(\mathbf{x}, \mathbf{z})}{\vartheta(\mathbf{x}, \mathbf{z})}} - \sqrt{s_{\mathbf{y}, \mathbf{z}}^2 \frac{1-\vartheta(\mathbf{y}, \mathbf{z})}{\vartheta(\mathbf{y}, \mathbf{z})}} \right)^2}{s_{\mathbf{x}, \mathbf{y}}^2}, \quad (37)$$

where the last equality was obtained rewriting the second equality in (33). The $\vartheta(\mathbf{x}, \mathbf{z})$ for which we get the smallest value of F' permitted by C (i.e. where the equality holds) is obtained by solving the quadratic equation. We get

$$\sqrt{\frac{1-\vartheta(\mathbf{x}, \mathbf{z})}{\vartheta(\mathbf{x}, \mathbf{z})}} = \sqrt{\frac{s_{\mathbf{y}, \mathbf{z}}^2 (1-\vartheta(\mathbf{y}, \mathbf{z}))}{s_{\mathbf{x}, \mathbf{z}}^2 \vartheta(\mathbf{y}, \mathbf{z})}} \pm \sqrt{\frac{s_{\mathbf{x}, \mathbf{y}}^2 (1-C)}{s_{\mathbf{x}, \mathbf{z}}^2 C}} \quad (38)$$

Values of the left-hand side leading to bigger F' are obtained below the “–” and above the “+” solution. We are in the latter regime if $s_{\mathbf{x}, \mathbf{z}}^2 \frac{1-\vartheta(\mathbf{x}, \mathbf{z})}{\vartheta(\mathbf{x}, \mathbf{z})} \geq s_{\mathbf{y}, \mathbf{z}}^2 \frac{1-\vartheta(\mathbf{y}, \mathbf{z})}{\vartheta(\mathbf{y}, \mathbf{z})}$. We can assume this to be fulfilled (because \mathbf{x} and \mathbf{y} indices can just be swapped in case it is not). Thus choosing the more convenient of the two ϑ , we obtain

$$\frac{1-\vartheta(\mathbf{x}, \mathbf{z})}{\vartheta(\mathbf{x}, \mathbf{z})} \geq \frac{\left(\sqrt{s_{\mathbf{y}, \mathbf{z}}^2 \frac{1-\vartheta(\mathbf{y}, \mathbf{z})}{\vartheta(\mathbf{y}, \mathbf{z})}} + \sqrt{\frac{s_{\mathbf{x}, \mathbf{y}}^2 (1-C)}{s_{\mathbf{x}, \mathbf{z}}^2 C}} \right)^2}{s_{\mathbf{x}, \mathbf{z}}^2} \geq \frac{s_{\mathbf{x}, \mathbf{y}}^2}{s_{\mathbf{x}, \mathbf{z}}^2} \frac{(1-C)}{C}. \quad (39)$$

We have thus found the following bound for one of the genes in the gene pair:

$$\vartheta(\mathbf{x}, \mathbf{z}) \leq \frac{1}{1 + \frac{s_{\mathbf{x}, \mathbf{y}}^2 (1-C)}{s_{\mathbf{x}, \mathbf{z}}^2 C}}. \quad (40)$$

Intuitively this makes sense: when the genes are correlated within the groups, the within-group LRV of the pair $s_{\mathbf{x}, \mathbf{y}}^2$ can be small compared to $s_{\mathbf{x}, \mathbf{z}}^2$, and then C may not be sufficiently small for differential expression of \mathbf{x} (see Figure 3 for an example). For differential expression we thus require a minimum within-group LRV of the differentially proportional pair. Note, however, that although we can control for both $s_{\mathbf{x}, \mathbf{y}}^2$ and C , the within-group variance of the gene $s_{\mathbf{x}, \mathbf{z}}^2$ remains inaccessible to us from a strict ratio point of view because it would require our knowledge of the reference \mathbf{z} leading to the correct normalization. Although for this reason we cannot precisely quantify how small C needs to be, the obtained bound on $\vartheta(\mathbf{x}, \mathbf{z})$ shows qualitatively that differentially proportional pairs with sufficiently high within-group variance will contain at least one differentially expressed gene.

2.5 Handling zeros

As reviewed in (Martín-Fernández et al., 2011), zeros resulting from undersampling (known as count zeros, and a major source of zeros in RNA-seq data) can best be dealt with assuming a Dirichlet prior leading to posterior counts where pseudocounts are added to the original counts. Along the same lines, one can also choose a resampling strategy, where repeated drawings from the posterior distribution lead to a kind of pseudo-replicates that do not contain zeros, which will represent variation expected from the original counts (Fernandes et al., 2013; Tarazona et al., 2015). Since an additive modification does not preserve ratios, a kind of multiplicative modification of a given count

$$\tilde{x}_{k,i} = \begin{cases} c & \text{if } x_{k,i} = 0, \\ (1 - c \cdot |\{j : x_{k,j} = 0\}|) \cdot x_{k,i} & \text{otherwise,} \end{cases} \quad (41)$$

was suggested (Martín-Fernández et al., 2011). Here the column indices i go over the genes in the given condition k , and the $\tilde{x}_{k,i}$ are the counts modified by the pseudocount c (which, for simplicity, we assume to be independent of the samples here). The fact that ratios are not preserved when simply adding the pseudocount, however, is felt strongest in the case of low counts, where ratios should not be trusted anyway. To alleviate the problem, it thus seems essential to use the precision weights of section 2.2 when calculating the relevant statistics.

While pseudocounts need an associated distributional theory to estimate them, a well-founded heuristic that has been used widely in data analysis are power transformations of the Box-Cox type. In the limiting case of a power tending to zero, these return the logarithm:

$$\log(x) = \lim_{\alpha \rightarrow 0} \frac{x^\alpha - 1}{\alpha}. \quad (42)$$

It has been shown by (Greenacre, 2009) that this transformation establishes a connection between Correspondence Analysis (CA) of the transformed data and log-ratio analysis, which is obtained as a limiting case of CA when letting α tend toward zero. This is interesting because CA handles zeros naturally. We will briefly describe this replacement strategy here. As shown in (Greenacre, 2011), from re-writing LRV in the form

$$\text{var}(\mathbf{L}_{1,\dots,n}^{\mathbf{x},\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^n \left(\log \frac{x_i}{(\prod_{j=1}^n x_j)^{\frac{1}{n}}} - \log \frac{y_i}{(\prod_{j=1}^n y_j)^{\frac{1}{n}}} \right), \quad (43)$$

a similarity with the (squared) χ^2 distance used in CA becomes evident. Here we show this distance for data raised to the power of α and with rows summing to one:

$$d_\alpha(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i^\alpha}{\frac{1}{n} \sum_{j=1}^n x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{n} \sum_{j=1}^n y_j^\alpha} \right)^2. \quad (44)$$

We can obtain (44) directly from (43) by applying (42) for nonzero α and replacing geometric by arithmetic means (which is justified in the limit $\alpha \rightarrow 0$).

A precision-weighted ϑ like in (13) that can also handle zeros can thus be defined by

$$\begin{aligned} \vartheta_{\alpha\omega}(\mathbf{x}, \mathbf{y}) = & \\ & \sum_{i=1}^{k_1} \omega_i \left(\frac{x_i^\alpha}{\frac{1}{\Omega_1} \sum_{j=1}^{k_1} \omega_j x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{\Omega_1} \sum_{j=1}^{k_1} \omega_j y_j^\alpha} \right)^2 + \sum_{i=k_1+1}^{k_1+k_2} \omega_i \left(\frac{x_i^\alpha}{\frac{1}{\Omega_2} \sum_{j=k_1+1}^{k_1+k_2} \omega_j x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{\Omega_2} \sum_{j=k_1+1}^{k_1+k_2} \omega_j y_j^\alpha} \right)^2 \\ & \sum_{i=1}^n \omega_i \left(\frac{x_i^\alpha}{\frac{1}{\Omega_1+\Omega_2} \sum_{j=1}^{k_1+k_2} \omega_j x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{\Omega_1+\Omega_2} \sum_{j=1}^{k_1+k_2} \omega_j y_j^\alpha} \right)^2 \end{aligned} \quad (45)$$

Note that the weighting scheme differs from the one used in CA where weights are determined from row and column sums and low counts get upweighted. The choice of α needs to trade off closeness to the original LRV values (for gene pairs not containing zero counts small α are more accurate) with the amount by which zeros should get punished (pairs containing zeros can have lower ϑ if α is larger).

2.6 GTEx data

For the practical examples shown here, we used data from the Genotype Tissue Expression (GTEx) project (Lonsdale et al., 2013). Reads were mapped using TopHat2 (Kim et al., 2013) and gene counts were obtained from the Flux Capacitor (Montgomery, 2010). 10,842 genes with nonzero counts throughout 7867 samples from 40 tissues were used, then samples were additionally filtered

for low ischemic times. Finally, only samples from two approximately balanced brain tissues (54 cerebellum and 44 cortex samples) were retained to match the use case discussed in this article. At an FDR of 5% (estimated by permutation tests) we find a cut-off $\vartheta < 0.94$ covering 26.6 million gene pairs (45% of all pairs). At $\vartheta < 0.69$ (4.56 million pairs) no false positives were detectable anymore. For high confidence disjointedly proportional pairs with clear within-tissue correlations, we settled for a much stricter cut-off of $\vartheta \leq 0.2$ (chosen subjectively by visual inspection of scatter plots) comprising 13,000 pairs. Conventional differential expression analysis using edgeR (Robinson et al., 2010) and DeSeq2 (Love et al., 2014) find about half of all considered genes differentially expressed at an FDR of 5%.

3 Outlook

While here we have presented how differential expression of ratios can be formalized, a practical proof of concept needs more in-depth analysis of relevant biological data sets. Preliminary results show that the approach holds great promise since the phenomenon of stoichiometry switches appears to be wide-spread both between tissues and between developmental stages when using data from BrainSpan (<http://developinghumanbrain.org>). These results will be reported elsewhere. The principle is not limited to providing a list of interesting gene pairs. Differential proportionality induces a distance measure between genes (e.g. in the form of ϑ) that can be used in a network analysis that is independent of normalization. Our R implementation, available soon as an addendum to the propr package (Quinn et al., 2017), will provide an entry point to relevant graph-based analyses.

Acknowledgements

I.E. thanks Christian Stenvang for checking differential expression using the edgeR and DeSeq2 packages. T.Q. thanks Tamsyn Crowley and Mark Richardson for their advice and expertise on next generation sequencing. I.E. and C.N. were supported by CRG internal funds provided by the Catalan Government.

References

- Dillies, M.A. et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14(6), pp. 671–683.
- Lun, A. et al. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* 17(1), pp. 75.
- Fernandes, A. et al. (2013). ANOVA-Like Differential Gene Expression Analysis of Single-Organism and Meta-RNA-Seq. *PLoS one* 8(7), e67019.
- Lovell, D. et al. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comp Biol* 11, e1004075.
- Erb, I. and Notredame, C (2016). How should we measure proportionality on relative gene expression data? *Theory Biosci* 135(1-2), pp. 21–36.
- Smyth, G.K. (2004). Linear models and empirical Bayes Methods for assessing differential expression in microarray experiments. *Stat Appl Genet* 3(1), Art. 3.
- Smyth, G.K. (2005). Limma: linear models for microarray data. R. Gentleman, V. Carey,

- S. Dudoit, R. Irizarry (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420. New York: Springer.
- Law, C.W. et al. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts *Genome Biology* 15, R29.
- Bar-Shira, O. et al. (2015). Gene Expression Switching of Receptor Subunits in Human Brain Development. *PLoS Comp Biol* 8(7), e67019.
- Tesson, B.M. et al. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* 11, pp. 497.
- Hastie, T. et al. (2013). *The Elements of Statistical Learning*. New York: Springer.
- Liu, R. et al. (2015). Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res* 43(15), e97.
- Lönnstedt, I and Speed, T (2002). Replicated Microarray data. *Statistica Sinica* 12(1), pp. 31–46.
- McCarthy, D.J. and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25(6), pp. 765–771.
- Witten, D. and Tibshirani, R. (2009) Covariance-Regularized Regression and Classification for High Dimensional Problems. *J. Royal. Stat. Soc. B* 71(3), pp. 615–636.
- Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell, NJ: The Blackburn press.
- Martín-Fernández, J.A. et al. (2011). Dealing with zeros. V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis*, pp. 43–58. Chichester, U.K.: Wiley.
- Tarazona, S. et al. (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package *Nucleic Acids Res* 43(21), e140.
- Greenacre, M. (2009). Power transformations in correspondence analysis. *Comput Statist Data Anal* 53, pp. 3107–3116.
- Greenacre, M. (2011). Measuring subcompositional incoherence. *Math Geosci* 43, pp. 681–693.
- Lonsdale, J. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat Genet* 45, 580585.
- Kim, D., et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14, R36.
- Montgomery, S.B. et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, pp. 773–777.
- Robinson, M.D. et al. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data *Bioinformatics* 26(1), pp. 139–140.
- Love, M.I., et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biology* 15(12), pp. 550.
- BrainSpan: Atlas of the Developing Human Brain. Available from: <http://developinghumanbrain.org>
- Quinn, T., et al. (2017) propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis *bioRxiv* 104935 doi: <https://doi.org/10.1101/104935>

A distribution on the simplex of the Generalized Beta type

Monique Graf

Institut de Statistique, Université de Neuchâtel

Elpacos Statistics, la Neuveville, Switzerland; *monique.p.n.graf@bluewin.ch*

Abstract

Consider a random vector with positive components following a compound distribution where the compounding parameter multiplies fixed scale parameters. The closed random vector is the vector divided by the sum of its components. We explicit on what conditions the distribution of the closed random vector does not depend on the mixing distribution. When the original vector has independent generalized Gamma components, it is shown that the unrelatedness of the distribution of the closed random vector to the compounding distribution depends on the parameters of the generalized Gamma. This fact is exemplified with the multivariate Generalized Beta distribution of the second kind (MGB2) in which the compounding parameter follows an inverse Gamma distribution. We call the most general distribution of the closed random vector, for which the compounding parameter has no influence, the simplicial Generalized Beta (SGB). Some properties and moments of the SGB are derived. Conditional moments given a sub-composition give a way to impute missing parts when knowing a sub-composition only. Maximum likelihood estimators of the parameters are obtained. The method is applied to the well known Arctic lake example.

Key words: Dirichlet distribution; Generalized Beta distribution of the second kind; simplicial Generalized Beta; maximum likelihood estimation; imputation.

1 Introduction

Let \mathbf{Y} be a random vector with independent components following Gamma distributions. Then the closed random vector $\mathcal{C}(\mathbf{Y})$ follows the Dirichlet distribution. As Aitchison (1986) says

... the popular Dirichlet class on the simplex has so many drawbacks that it has virtually no role to play in simplicial inference. For example, it has no simple perturbation or power transformation properties and so is ill-suited to the basic operations of the simplex. Moreover, it has so many inbuilt independence properties that, apart from being a model of extreme independence, it has almost no role to play in the investigation of the nature of the dependence structure of compositional variability. (...) Many statisticians are attempting to extend the Dirichlet class of distributions on the simplex in the hope that greater generality will bring greater realism than the simple Dirichlet class. Unfortunately I think they are likely to fail, since even the simple Dirichlet class with all its elegant mathematical properties does not have any exact perturbation properties.

In this paper, a generalization of the Dirichlet family that possess perturbation and power transformation properties is proposed. We call it the simplicial generalized Beta (SGB). It is along the lines of Craiu and Craiu (1969) who obtained a similar distribution by closing a random vector with independent generalized Gamma components, see Kotz et al. (2000, p. 490). Unfortunately, their paper is difficult to find.

The search of extensions of the Dirichlet distribution is an old problem. Only a few will be recalled here. Kotz et al. (2000, Chapter 49) describe several of them. If $\mathbf{U} = (U_1, \dots, U_D)$ follows a Dirichlet distribution with parameters p_1, \dots, p_D , then $U_j / \sum_{i=j}^D U_i$ are independent Beta random variables with parameters p_j and $\sum_{i=j}^D p_i$. When these parameters take arbitrary (positive) values, the generalized Dirichlet distribution is obtained (Connor and Mosimann, 1969). These authors define the concept of complete neutrality, i.e. independence of $C(U_{j+1}, \dots, U_D)$ and (U_1, \dots, U_j) , for $1 \leq j \leq D$. Complete neutrality involves the order of the components, see also James (1975). The Dirichlet distribution is completely neutral and remains so under all permutations of components. The SGB distribution is not completely neutral. Another generalization is done by Rayens and Srinivasan (1994) who define generalized Liouville distributions that also embed the Dirichlet. It is shown that the SGB is not a member of this class. To address the criticism of extreme independence, Ongaro et al. (2008) define a flexible Dirichlet model. Let $(Y_i, W_i, X, i = 1, \dots, D)$ be independent Gamma variables with different parameters. Consider the closed vector $\mathcal{C}(\mathbf{Y} + X\mathbf{W}) = (\mathbf{Y} + X\mathbf{W}) / (\sum Y_i + XW_i)$. Then $\mathcal{C}(\mathbf{Y} + X\mathbf{W})$ follows a flexible Dirichlet distribution. Ongaro et al. (2008) show that the flexible Dirichlet has a representation as a finite mixture of Dirichlet distributions.

Mateu-Figueras et al. (2003) make an important contribution by showing that the log-ratio approach advocated by Aitchison (1986) can be interpreted as a change of measure within the simplex. Monti et al. (2015, 2016) apply the principle to the scaled Dirichlet distribution. They prefer to call this distribution the shifted Dirichlet distribution, when the Aitchison's measure (Pawlowsky-Glahn et al., 2007) is used rather than the Lebesgue measure.

The paper is organized as follows: Let \mathbf{Y} be a random vector following a compound distribution, where the compounding variable multiplies the scale parameters. The influence of the compounding variable on the compositional distribution of $\mathcal{C}(\mathbf{Y})$ is examined in Section 2. Then the simplicial generalized Beta distribution (SGB) distribution is defined. In Section 3, some properties of the SGB are given. An application is developed in Section 4. A discussion is found in Section 5.

2 Influence of mixture on a compositional distribution

A mean to generate multivariate distributions is to consider a random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_D)$ of components that are independent given a random parameter, and then obtain the expected distribution of \mathbf{Y} with respect to the distribution of the random parameter. When all the variables are positive and continuous, the resulting multivariate density is

$$f(\mathbf{y}) = f(y_1, y_2, \dots, y_D) = \int_0^\infty \prod_{k=1}^D f_k(y_k | \theta) g(\theta) d\theta,$$

where $g(\cdot)$ is the density of the compounding parameter Θ and $f_k(y_k | \theta)$ and $g(\theta)$ can depend on other fixed parameters.

To generate a distribution on the simplex from the distribution of the vector \mathbf{Y} of positive random variables, we can compute the distribution of the closed random vector $\mathcal{C}(\mathbf{Y}) = \mathbf{Y} / \sum_{k=1}^D Y_k$. A question of interest is to evaluate the influence of the mixing scheme on the distribution of $\mathcal{C}(\mathbf{Y})$.

Let us introduce the following change of variables,

$$\begin{cases} t &= \sum_{j=1}^D y_j, \\ u_k &= y_k/t, \quad k = 1, \dots, D-1. \end{cases} \quad (1)$$

The Jacobian of the transformation is t^{D-1} . Let $\mathbf{u} = (u_1, \dots, u_{D-1})$ and $u_D = 1 - \sum_{k=1}^D u_k$. We

have

$$\begin{aligned} f(\mathbf{u} | \theta) &= \int_0^\infty f(\mathbf{y}(\mathbf{u}, t) | \theta) t^{D-1} dt \\ f(\mathbf{u}) &= \int_0^\infty f(\mathbf{u} | \theta) g(\theta) d\theta. \end{aligned}$$

It is clear that if $f(\mathbf{u} | \theta)$ does not depend on θ , then the mixing distribution acts on $T = \sum Y_k$ only and has no effect on the distribution of the composition \mathbf{U} .

2.1 Dirichlet family of distributions

Consider first the case where no mixing distribution is used.

Let the density of the random vector \mathbf{Y} be given by the product of independent two-parameters Gamma variables, with scale parameters $b_k > 0$ and shape parameters $p_k > 0$, $k = 1, \dots, D$,

$$f(\mathbf{y}) = \prod_{k=1}^D \frac{1}{\Gamma(p_k)b_k} (y_k/b_k)^{p_k-1} \exp(-y_k/b_k), \quad y_k \geq 0.$$

Then $\mathcal{C}(\mathbf{Y})$ follows a scaled Dirichlet distribution (Monti et al., 2015, 2016).

When $b_k = 1$ ($k = 1, \dots, D$), the distribution of $\mathcal{C}(\mathbf{Y})$ is the ordinary Dirichlet distribution.

It is possible to add some flexibility by considering the generalized Gamma distribution for Y_k , with density

$$f_k(y_k; a_k, b_k, p_k) = \frac{a_k}{\Gamma(p_k)b_k} (y_k/b_k)^{a_k p_k - 1} \exp(-(y_k/b_k)^{a_k}),$$

which depends on still another set of parameters $a_k > 0$, $k = 1, \dots, D$.

Let us introduce a compounding parameter, acting on the scales b_k in the following way,

$$\begin{aligned} f(\mathbf{y} | \theta) &= \prod_{k=1}^D f_k(y_k; a_k, \theta^{1/a_k} b_k, p_k) \\ &= \prod_{k=1}^D \frac{a_k}{\Gamma(p_k)\theta^{1/a_k} b_k} \left(\frac{y_k}{\theta^{1/a_k} b_k} \right)^{a_k p_k - 1} \exp \left(- \left(\frac{y_k}{\theta^{1/a_k} b_k} \right)^{a_k} \right), \end{aligned} \quad (2)$$

then

$$f(\mathbf{y}) = \int_0^\infty f(\mathbf{y} | \theta) g(\theta) d\theta \quad (3)$$

for some mixing distribution defined by the density g .

Examples

1. When $\Theta = 1$ with probability 1, then $f(\mathbf{y}) = f(\mathbf{y} | \theta)$. In this case, $f(\mathbf{u}) = f(\mathbf{u} | \theta)$ trivially.
2. When Θ follows an $InvG(1, q)$ (inverse Gamma distribution), then \mathbf{Y} follows an $MGB2(\{a_k, b_k, p_k, k = 1, \dots, D\}, q)$ distribution, (Yang et al., 2011).

The next theorem gives the necessary and sufficient conditions for \mathbf{U} to depend on the mixing scheme when the conditional distribution is given by Equation (2).

Theorem 1. *The distribution of $\mathbf{U} = \mathcal{C}(\mathbf{Y})$ stemming from the distribution for \mathbf{Y} in Equation (3), with the conditional density given by Equation (2), does not depend on the mixing distribution if and only if $a_k = a$ for all $k = 1, \dots, D$.*

Proof

1. $\{a_k, k = 1, \dots, D\}$ not constant implies dependence on θ .

Making the change of variables defined in Equation (1) and setting $u_D = 1 - \sum_{j=1}^{D-1} u_j$, we obtain

$$\begin{aligned} f(\mathbf{u}, t|\theta) &= \prod_{k=1}^D \left[\frac{a_k}{\Gamma(p_k)\theta^{1/a_k} b_k} \left(\frac{tu_k}{\theta^{1/a_k} b_k} \right)^{a_k p_k - 1} \exp \left(- \left(\frac{tu_k}{\theta^{1/a_k} b_k} \right)^{a_k} \right) \right] t^{D-1} \\ &= \left[\prod_{k=1}^D \frac{a_k}{\Gamma(p_k)\theta^{1/a_k} b_k} \left(\frac{u_k}{b_k} \right)^{a_k p_k - 1} \right] \exp \left[- \sum_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \frac{u_k}{b_k} \right)^{a_k} \right] t^{D-1} \prod_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \right)^{(a_k p_k - 1)} \\ &= \left[\prod_{k=1}^D \frac{a_k}{\Gamma(p_k)b_k} \left(\frac{u_k}{b_k} \right)^{a_k p_k - 1} \right] \exp \left[- \sum_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \frac{u_k}{b_k} \right)^{a_k} \right] \prod_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \right)^{a_k p_k} \frac{1}{t} \\ &= f(\mathbf{u}|\theta)f(t|\mathbf{u}, \theta). \end{aligned}$$

We want to find the constant of integration C , such that

$$\begin{aligned} C \int_0^\infty f(t|\mathbf{u}, \theta) dt &= \int_0^\infty \exp \left[- \sum_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \frac{u_k}{b_k} \right)^{a_k} \right] \prod_{k=1}^D \left(\frac{t}{\theta^{1/a_k}} \right)^{a_k p_k} \frac{1}{t} dt \\ &= \int_0^\infty \exp \left[-\theta^{-1} \sum_{k=1}^D \left(\frac{t u_k}{b_k} \right)^{a_k} \right] \theta^{-P} \prod_{k=1}^D t^{a_k p_k} \frac{1}{t} dt. \end{aligned}$$

Unfortunately, this expression is difficult to integrate analytically, but would be easy to obtain numerically, if the parameters were known. It is clear nevertheless that, if the parameters a_k are not constant, the result still depends on θ . This implies that in this case the distribution of the composition depends on the mixing scheme.

2. $\{a_k, k = 1, \dots, D\}$ constant implies independence on θ .

If $a_k = a$ for all $k = 1, \dots, D$, $f(t|\mathbf{u}, \theta)$ is easily integrated.

Setting

$$c_k = (u_k/b_k)^a, \quad v = \left(\sum_{k=1}^D c_k \right) \frac{t^a}{\theta} \text{ and } dv = a \left(\sum_{k=1}^D c_k \right) \frac{t^a}{\theta} \frac{1}{t} dt,$$

we have

$$\begin{aligned} C \int_0^\infty f(t|\mathbf{u}, \theta) dt &= \int_0^\infty \exp \left\{ - \left(\sum_{k=1}^D c_k \right) \frac{t^a}{\theta} \right\} \left(\frac{t^a}{\theta} \right)^P \frac{1}{t} dt \\ &= \frac{1}{a \left(\sum_{k=1}^D c_k \right)^P} \int_0^\infty \exp(-v) v^{P-1} dv = \frac{\Gamma(P)}{a \left(\sum_{k=1}^D c_k \right)^P} \\ &= \frac{\Gamma(P)}{a \left(\sum_{k=1}^D (u_k/b_k)^a \right)^P}. \end{aligned} \tag{4}$$

Thus the constant in Equation (4) does not depend on θ . The density of the compositional vector \mathbf{U} is obtained by integrating t out,

$$f(\mathbf{u}|\theta) = \int_0^\infty f(\mathbf{u}, t|\theta) dt = \left[\prod_{k=1}^D \frac{a}{\Gamma(p_k)b_k} \left(\frac{u_k}{b_k} \right)^{ap_k - 1} \right] \frac{\Gamma(P)}{a \left(\sum_{k=1}^D (u_k/b_k)^a \right)^P} = f(\mathbf{u}). \tag{5}$$

Thus this distribution does not depend on θ . \square

2.2 Simplicial generalized beta distribution

In the sequel, we suppose that $a_k = a$, $k = 1, \dots, D$.

Substituting $1 - \sum_{j=1}^{D-1} u_j$ to u_D into Equation (5), we obtain a generalization of the scaled Dirichlet distribution.

Definition 1. *The joint density of the random composition \mathbf{U} given by*

$$f_{\mathbf{U}}(\mathbf{u}_{-D}) = \frac{\Gamma(P)a^{D-1}}{\prod_{j=1}^D \{\Gamma(p_j)b_j\}} \frac{\prod_{k=1}^{D-1} (u_k/b_k)^{ap_k-1} \left[(1 - \sum_{j=1}^{D-1} u_j)/b_D \right]^{ap_D-1}}{\left[\sum_{k=1}^{D-1} (u_k/b_k)^a + \left((1 - \sum_{j=1}^{D-1} u_j)/b_D \right)^a \right]^P} \quad (6)$$

$$0 < u_j < 1, \quad j = 1, \dots, D-1, \quad \sum_{j=1}^{D-1} u_j < 1,$$

$$a, b_j > 0, j = 1, 2, \dots, D.$$

is called the simplicial generalized beta distribution and denoted by

$$SGB(a, \{b_j, p_j, j = 1, \dots, D\}).$$

Kotz et al. (2000) mention a similar distribution obtained by closing a random vector with independent generalized Gamma components (Craiu and Craiu, 1969). Unfortunately, their paper is difficult to find.

Let us define the L_a -norm of the vector

$$\left(u_1/b_1, u_2/b_2, \dots, u_{D-1}/b_{D-1}, (1 - \sum_{j=1}^{D-1} u_j)/b_D \right),$$

by

$$\|\mathbf{u}/\mathbf{b}\|_a = \left[\sum_{k=1}^{D-1} (u_k/b_k)^a + \left((1 - \sum_{j=1}^{D-1} u_j)/b_D \right)^a \right]^{1/a}.$$

Please note that in the present notations, $\|\cdot\|_a$ does not represent the Aitchison's norm, see e.g. Mateu-Figueras et al. (2003) or Pawlowsky-Glahn et al. (2007), but the L_a -norm.

We obtain the density in Equation (6) written as

$$f_{\mathbf{U}}(\mathbf{u}_{-D}) = \frac{\Gamma(P)a^{D-1}}{\prod_{j=1}^D \Gamma(p_j)} \times \\ \prod_{k=1}^{D-1} \left\{ \frac{u_k/b_k}{\|\mathbf{u}/\mathbf{b}\|_a} \right\}^{ap_k} \left\{ \frac{(1 - \sum_{j=1}^{D-1} u_j)/b_D}{\|\mathbf{u}/\mathbf{b}\|_a} \right\}^{ap_D} \frac{1}{\prod_{k=1}^{D-1} u_k (1 - \sum_{j=1}^{D-1} u_j)}. \quad (7)$$

The vector \mathbf{b} is defined up to a multiplicative constant, because the fractions in curled brackets are scale-free. Thus \mathbf{b} can be seen as a composition. The last fraction in Equation (7) appears, because the density $f_{\mathbf{U}}$ is expressed with respect to the Lebesgue measure, see Monti et al. (2015) for more details.

Setting

$$\mathbf{Z}(\mathbf{U}) = \frac{(\mathbf{U}/\mathbf{b})^a}{(\|\mathbf{U}/\mathbf{b}\|_a)^a} = a \odot (\mathbf{U} \ominus \mathbf{b}), \quad (8)$$

where $\mathbf{Z}(\mathbf{U})$ is a composition, we can also express the density in Equation (7) in the following way,

$$f_{\mathbf{U}}(\mathbf{u}_{-D}) = \frac{\Gamma(P)}{\prod_{j=1}^D \Gamma(p_j)} \prod_{k=1}^D z_k(\mathbf{u})^{p_k} \frac{1}{\prod_{k=1}^{D-1} u_k \left(1 - \sum_{j=1}^{D-1} u_j\right)}. \quad (9)$$

It is clear from Equation (9) that the mixed moments of $\mathbf{Z}(\mathbf{U})$ are given by

$$\mathbb{E} \left(\prod_{k=1}^D Z_k(\mathbf{U})^{\alpha_k} \right) = \frac{\Gamma(P)}{\Gamma(P + \sum_{k=1}^D \alpha_k)} \prod_{k=1}^D \frac{\Gamma(p_k + \alpha_k)}{\Gamma(p_k)}.$$

In particular,

$$\begin{aligned} \mathbb{E}(Z_k(\mathbf{U})) &= \frac{p_k}{P}, & k = 1, \dots, D \\ \mathbb{E}(Z_k(\mathbf{U})^{1/a}) &= \frac{\Gamma(P)}{\Gamma(P + 1/a)} \frac{\Gamma(p_k + 1/a)}{\Gamma(p_k)} & k = 1, \dots, D. \end{aligned}$$

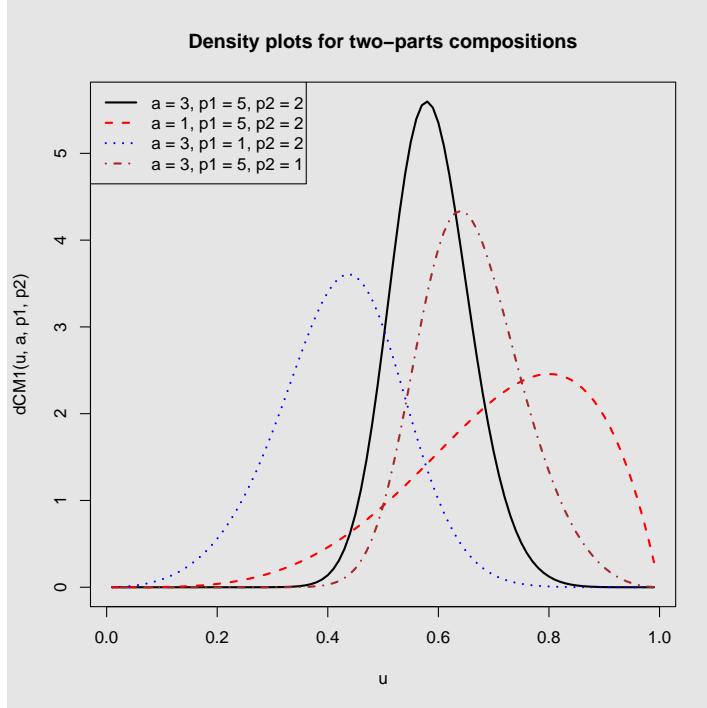


Figure 1: Densities for varying parameters and $b_1 = b_2 = 1/2$

The introduction of a gives a lot of flexibility, see Figure 1, which shows different densities for the two-parts compositions case.

Monti et al. (2011, p.130, Proposition 10.2.1) obtain the density of the Dirichlet distribution in Aitchison's geometry, that is on making a change of variables to a set of $D - 1$ orthogonal contrasts at the log scale, processing the analysis in this space and taking the inverse image in the simplex.

For more explanations on the change of measure principle, see Pawlowsky-Glahn et al. (2007). Likewise, the density of \mathbf{U} with respect to Aitchison's measure is given by

$$\tilde{f}_{\mathbf{U}}(\mathbf{u}) = \frac{\sqrt{D} \Gamma(P)}{\prod_{j=1}^D \Gamma(p_j)} \prod_{k=1}^D z_k(\mathbf{u})^{p_k},$$

In this distribution, all the parameters have a compositional meaning. The composition \mathbf{b} plays the role of location and the parameter a plays the role of scale, as it can be seen from the simplicial relationship in Equation (8). The parameters p_k are shape parameters inherited from the Dirichlet distribution followed by the composition $\mathbf{Z}(\mathbf{U})$.

Denoting by $E_A(\cdot)$ the expectation operator in Aitchison's geometry, we have, similarly to (Monti et al., 2011, p.135 for the Dirichlet case),

$$E_A[\mathbf{Z}(\mathbf{U})] = \mathcal{C}\left(e^{\psi(p_1)}, \dots, e^{\psi(p_D)}\right),$$

where ψ is the digamma function. See also Section 3.2 for a direct justification.

Thus in Aitchison's geometry, the expectation $E_A(\mathbf{U})$ of the composition \mathbf{U} is simply given by

$$E_A(\mathbf{U}) = \mathbf{b} \oplus E_A[\mathbf{Z}(\mathbf{U})] \odot (1/a),$$

that is

$$E_A(U_k) = \frac{b_k \exp\{\psi(p_k)/a\}}{\sum_{j=1}^D b_j \exp\{\psi(p_j)/a\}} \quad k = 1, \dots, D. \quad (10)$$

By contrast, the expectation of \mathbf{U} is difficult to obtain analytically under the Lebesgue measure. Because $\mathbf{Z}(\mathbf{U})^{1/a}$, defined at Equation (8), is proportional to \mathbf{U} , we can estimate the expected composition by

$$E(\mathbf{U}) \approx \mathcal{C}\left(\frac{b_1 \Gamma(p_1 + 1/a)}{\Gamma(p_1)}, \dots, \frac{b_D \Gamma(p_D + 1/a)}{\Gamma(p_D)}\right). \quad (11)$$

This expression is not exact, because it ignores the fact that \mathbf{U}/\mathbf{b} is multiplied by $1/\|\mathbf{U}/\mathbf{b}\|_a$ which is a random variable.

3 Properties of the simplicial generalized Beta distribution

1. The case $a = 1$ is the scaled Dirichlet distribution (Monti et al., 2015). We obtain the ordinary Dirichlet distribution when $a = 1$ and \mathbf{b} is such that $b_j = 1, j = 1, \dots, D$ (recall that \mathbf{b} is defined up to an arbitrary multiplicative constant).
2. Given $2 \leq C \leq D$, Aitchison (1986) defines a general partition of the composition $\mathbf{U} = (U_1, \dots, U_D) = (U_1, \dots, U_{D_1}, |, U_{D_1+1}, \dots, U_{D_1+D_2}|, \dots, U_{D_1+D_2+\dots+D_{C-1}+1}, \dots, U_D)$ as the set of compositions

$$\mathcal{P} = \{S^{(c)}, c = 1, \dots, C; T\},$$

where $S^{(c)} = \mathcal{C}(U_{D_1+D_2+\dots+D_{c-1}+1}, \dots, U_{D_1+D_2+\dots+D_c})$ are sub-compositions, and $T = (T_1, \dots, T_C)$ is the composition of the corresponding amalgamations. The random composition \mathbf{U} admits complete sub-compositional independence if \mathcal{P} forms an independent set, for all C and every ordering of the parts. The Dirichlet distribution possesses this property which is an extreme case of compositional independence. It is in general not the case for the SGB.

3. Rayens and Srinivasan (1994) introduce a new family - the Liouville type distributions - that embeds the Dirichlet and admits more general dependence properties.

Definition 2. Suppose that a distribution on the simplex can be expressed as

$$f(\mathbf{u}) = K(\mathbf{u}_{-D}) \prod_{k=1}^{D-1} u_k^{p_k-1}.$$

If the kernel $K(\mathbf{u}_{-D})$ can be put into the form

$$K(\mathbf{u}_{-D}) = h \left(\sum_{j=1}^{D-1} (u_j/b_j)^{\beta_j} \right), \quad (12)$$

where h is a continuous function of $\mathbb{R}_+^1 \rightarrow \mathbb{R}_+^1$, then the distribution is a generalized Liouville distribution.

Theorem 2. The $SGB(a, \{b_j, p_j, j = 1, \dots, D\})$ distribution is of the generalized Liouville type if and only if it is a Dirichlet distribution.

Proof Taking the density $f_{\mathbf{U}}(\mathbf{u}_{-D})$ as expressed in Equation (6), we see that the kernel is, up to a constant factor,

$$K(\mathbf{u}_{-D}) \propto \frac{\left[(1 - \sum_{j=1}^{D-1} u_j)/b_D \right]^{ap_D-1}}{\left\{ \sum_{j=1}^{D-1} (u_j/b_j)^a + (1 - \sum_{j=1}^{D-1} u_j)/b_D \right\}^P},$$

and cannot be put into the form given in Equation (12), except if $a = b_j = 1$, in which case it reduces to

$$K(\mathbf{u}_{-D}; a = 1, b_j = 1, j = 1, \dots, D) \propto \left[(1 - \sum_{j=1}^{D-1} u_j) \right]^{p_D-1}.$$

Thus $h(x; a = 1, b_j = 1, j = 1, \dots, D) = (1 - x)^{p_D-1}$. \square

The SGB is thus a type of generalization different from the Liouville family. In Theorem 3 below, it is shown that if \mathbf{U} is SGB, two sub-compositions involving different parts are independent, but that they are not independent of the corresponding amalgamation, except if the SGB is a Dirichlet distribution.

3.1 Mixed moments of ratios of parts

It is equivalent to compute moment of ratios and log-ratios of parts from the distribution of the composition \mathbf{U} or from the initial vector \mathbf{Y} , because

$$U_k/U_j = Y_k/Y_j \quad \text{for all } j, k = 1, \dots, D.$$

The generating function of the mixed moments of the random vector \mathbf{Y} is given by

$$M_{\mathbf{Y}}(t_1, \dots, t_D) = E(Y_1^{t_1} \dots Y_D^{t_D}).$$

Ratios and any other scale-free transformations of \mathbf{Y} are characterized by exponents t_k that sum to zero, thus for the random composition $\mathbf{U} = \mathcal{C}(\mathbf{Y})$,

$$M_{\mathbf{U}}(t_1, \dots, t_{D-1}) = M_{\mathbf{Y}}(t_1, \dots, t_D), \text{ if } t_1 + \dots + t_D = 0.$$

Set $t_+ = \sum_{j=1}^{D-1} t_j$. Then the mixed moment ratios of the random composition following a $SGB(a, \{b_j, p_j\}, j = 1, \dots, D)$ distribution are given by the corresponding moment of a product of generalized Gamma random variables, namely,

$$\begin{aligned} M_{\mathbf{U}}(t_1, \dots, t_{D-1}) &= M_{\mathbf{Y}}(t_1, \dots, t_{D-1}, -t_+) \\ &= E \left[\left(\frac{U_1}{U_D} \right)^{t_1} \cdots \left(\frac{U_{D-1}}{U_D} \right)^{t_{D-1}} \right] \\ &= \frac{\prod_{k=1}^{D-1} (b_k)^{t_k} \left\{ \prod_{k=1}^{D-1} \Gamma(p_k + t_k/a) \right\} \Gamma(p_D - t_+/a)}{(b_D)^{t_+} \prod_{j=1}^D \Gamma(p_j)} \\ &\quad -ap_k < t_k, k = 1, \dots, D-1; t_+ < ap_D. \end{aligned} \quad (13)$$

Let \mathbf{e}_j be the vector of length $D-1$ with 1 in the j -th position and 0 otherwise. We have the following expectations,

$$E \left(\frac{U_j}{U_D} \right) = M_{\mathbf{U}}(\mathbf{e}_j) = \frac{b_j}{b_D} \frac{\Gamma(p_j + 1/a)\Gamma(p_D - 1/a)}{\Gamma(p_j)\Gamma(p_D)}, \quad j = 1, \dots, D-1 \quad (14)$$

$$E \left(\frac{U_i}{U_j} \right) = M_{\mathbf{U}}(\mathbf{e}_i - \mathbf{e}_j) = \frac{b_i}{b_j} \frac{\Gamma(p_i + 1/a)\Gamma(p_j - 1/a)}{\Gamma(p_i)\Gamma(p_j)}, \quad i \neq j \quad (15)$$

$$E \left(\left[\frac{U_i}{U_j} \right]^2 \right) = M_{\mathbf{U}}(2\mathbf{e}_i - 2\mathbf{e}_j) = \left[\frac{b_i}{b_j} \right]^2 \frac{\Gamma(p_i + 2/a)\Gamma(p_j - 2/a)}{\Gamma(p_i)\Gamma(p_j)}, \quad (16)$$

$$E \left(\frac{U_i U_k}{U_j^2} \right) = M_{\mathbf{U}}(\mathbf{e}_i + \mathbf{e}_k - 2\mathbf{e}_j) = \left[\frac{b_i b_k}{b_j^2} \right] \frac{\Gamma(p_i + 1/a)\Gamma(p_k + 1/a)\Gamma(p_j - 2/a)}{\Gamma(p_i)\Gamma(p_k)\Gamma(p_j)},$$

$$\begin{aligned} E \left(\frac{U_i}{U_j} \frac{U_k}{U_\ell} \right) &= M_{\mathbf{U}}(\mathbf{e}_i - \mathbf{e}_j + \mathbf{e}_k - \mathbf{e}_\ell) \\ &= \frac{b_i b_k}{b_j b_\ell} \frac{\Gamma(p_i + 1/a)\Gamma(p_j - 1/a)\Gamma(p_k + 1/a)\Gamma(p_\ell - 1/a)}{\Gamma(p_i)\Gamma(p_j)\Gamma(p_k)\Gamma(p_\ell)} \\ &= E \left(\frac{U_i}{U_j} \right) E \left(\frac{U_k}{U_\ell} \right), \quad i, j, k, \ell \text{ all distinct.} \end{aligned} \quad (17)$$

Thus distinct pairs of ratios of parts are uncorrelated.

Summing Equation (15) over $i, i \neq j$, we obtain the expectation of $(1 - U_j)/U_j$. Then it is easy to deduce the expectation of $1/U_j$,

$$E \left(\frac{1}{U_j} \right) = 1 + \frac{\Gamma(p_j - 1/a)}{b_j \Gamma(p_j)} \sum_{i \neq j} \frac{b_i \Gamma(p_i + 1/a)}{\Gamma(p_i)}. \quad (18)$$

The expression for $E(1/U_j)^2$ and higher moments are obtained by the same principle.

3.2 Mixed moments of log-ratios of parts

All contrasts at the log scale can be expressed as linear combination of log-ratios. The function $M_{\mathbf{U}}$ in Equation (13) can be seen as the moment generating function of the log-ratios of parts.

Setting $\mathbf{t} = (t_1, \dots, t_{D-1})$, we have

$$\begin{aligned} \mathbb{E} \log \left(\frac{U_i}{U_D} \right) &= \frac{d}{dt} \Big|_{t=0} M_{\mathbf{U}}(t\mathbf{e}_i) = \log \left(\frac{b_i}{b_D} \right) + \frac{1}{a} (\psi(p_i) - \psi(p_D)) \quad (19) \\ \text{Cov} \left(\log \frac{U_i}{U_D}, \log \frac{U_j}{U_D} \right) &= \frac{1}{a^2} \psi^{(1)}(p_D) > 0 \\ \text{Var} \left(\log \frac{U_i}{U_j} \right) &= \text{Var} \left(\log \frac{U_i}{U_D} - \log \frac{U_j}{U_D} \right) \\ &= \frac{1}{a^2} (\psi^{(1)}(p_i) + \psi^{(1)}(p_j)), \quad i \neq j, \\ \text{Cov} \left(\log \frac{U_i}{U_k}, \log \frac{U_j}{U_\ell} \right) &= \text{Cov} \left(\log \frac{U_i}{U_D} - \log \frac{U_k}{U_D}, \log \frac{U_j}{U_D} - \log \frac{U_\ell}{U_D} \right) = 0, \\ &\quad i, j, k, \ell \text{ all distinct.} \end{aligned}$$

Thus distinct pairs of log-ratios of parts are uncorrelated. We also see that the choice of the D -th component doesn't matter. The technique can be readily applied to ilr transforms of any kind. Notice that from Equation (19), we recover Equation (10),

$$\mathbb{E}_A(\mathbf{U}) = \mathcal{C} \left[\exp \mathbb{E} \log \left(\frac{\mathbf{U}}{U_D} \right) \right].$$

3.3 Marginal and conditional distributions

The next theorem gives the distribution of a general partition with two sub-compositions. The generalization to $C > 2$ sub-compositions is immediate. The theorem shows that, except in the Dirichlet case, the SGB distributions do not suffer from complete sub-compositional independence. The distributional properties of the general partition is given. A consequence of the theorem is that it is possible to impute missing parts using the available information. An example will be given in Section 4.2.

Theorem 3. Let \mathbf{U} be a random composition following a $SGB(a, \{b_j, p_j\}, j = 1, \dots, D)$ distribution. Consider a subset J of $\{1, 2, \dots, D\}$. Let us denote the corresponding sub-vectors of \mathbf{U} and \mathbf{b} by \mathbf{U}_J and \mathbf{b}_J respectively, and their complements by \mathbf{U}_{D-J} and \mathbf{b}_{D-J} . Let $\mathbf{V} = \mathcal{C}(\mathbf{U}_J)$ and $\mathbf{W} = \mathcal{C}(\mathbf{U}_{D-J})$ the two sub-compositions, and $(X, 1-X)$ the amalgamation, where $X = \sum_{j \in J} U_j$.

1. The sub-compositions \mathbf{V} and \mathbf{W} are independently distributed as $SGB(a, \{(b_j^*, p_j), j \in J\})$ and $SGB(a, \{(b_j^*, p_j), j \in D - J\})$ respectively, where $\mathbf{b}_J^* = \mathcal{C}(\mathbf{b}_J)$ and $\mathbf{b}_{D-J}^* = \mathcal{C}(\mathbf{b}_{D-J})$.
2. The conditional distribution of $(X|\mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w})$ is

$$SGB \left(a, \{(\|\mathbf{v}/\mathbf{b}_1\|_a^{-1}, P_1), (\|\mathbf{w}/\mathbf{b}_2\|_a^{-1}, P_2)\} \right),$$

where $P_1 = \sum_{j=1}^r p_j$ and $P_2 = \sum_{j=r+1}^D p_j$.

3. The best predictors of $X/(1-X)$ and of $X^a/(1-X)^a$ are given by

$$\begin{aligned} \mathbb{E} \left(\frac{X}{1-X} \middle| \mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w} \right) &= \frac{\|\mathbf{w}/\mathbf{b}_2\|_a \Gamma(P_1 + 1/a) \Gamma(P_2 - 1/a)}{\|\mathbf{v}/\mathbf{b}_1\|_a \Gamma(P_1) \Gamma(P_2)}. \quad (20) \\ \mathbb{E} \left(\frac{X^a}{(1-X)^a} \middle| \mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w} \right) &= \frac{(\|\mathbf{w}/\mathbf{b}_2\|_a)^a P_1}{(\|\mathbf{v}/\mathbf{b}_1\|_a)^a P_2} \end{aligned}$$

4. The best predictor of $\log[X/(1-X)]$ is given by

$$\mathbb{E} \left(\log \frac{X}{1-X} \middle| \mathbf{V} = \mathbf{v}, \mathbf{W} = \mathbf{w} \right) = \log \frac{\|\mathbf{w}/\mathbf{b}_2\|_a}{\|\mathbf{v}/\mathbf{b}_1\|_a} + \frac{1}{a} [\psi(P_1) - \psi(P_2)].$$

5. The conditional expectation of X under Aitchison's measure is

$$E_A(X) = \frac{\|\mathbf{v}/\mathbf{b}_1\|_a^{-1} \exp(\psi(P_1)/a)}{\|\mathbf{v}/\mathbf{b}_1\|_a^{-1} \exp(\psi(P_1)/a) + \|\mathbf{w}/\mathbf{b}_2\|_a^{-1} \exp(\psi(P_2)/a)}$$

3.4 Partial derivatives of the log-likelihood

The most convenient form of the joint density of the random composition \mathbf{U} is given at Equation (9). Now, consider N independent compositions \mathbf{U}_i , $i = 1, \dots, N$. The log-likelihood is given by

$$\begin{aligned} & \ell(a, (b_1, p_1), \dots, (b_D, p_D) | \mathbf{u}_{i,-D}, i = 1, \dots, N) \\ &= N \left[(D-1) \log(a) + \log \Gamma(P) - \sum_{k=1}^D \log \Gamma(p_k) \right] + \sum_{i=1}^N \sum_{k=1}^D p_k \log z_k(\mathbf{u}_i) \\ & \quad - \text{terms not depending on parameters.} \end{aligned}$$

with $z(\mathbf{u}_i) = (z_1(\mathbf{u}_i), \dots, z_D(\mathbf{u}_i))$ given at Equation (8).

Set

$$K_{bi} = \sum_{j=1}^D z_j(\mathbf{u}_i) \log \left(\frac{u_{ij}}{b_{ij}} \right).$$

Then the partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial a} &= \frac{N(D-1)}{a} + \sum_{i=1}^N \sum_{k=1}^D p_k \left[\log \left(\frac{u_{ik}}{b_{ik}} \right) - K_{bi} \right] \quad (21) \\ \frac{\partial \ell}{\partial b_{ik}} &= \sum_{j=1}^D p_j \frac{\partial \log z_j(\mathbf{u}_i)}{\partial b_{ik}} = \frac{a}{b_{ik}} (P z_k(\mathbf{u}_i) - p_k) \quad k = 1, \dots, D, \quad i = 1, \dots, N \end{aligned}$$

$$\frac{\partial \ell}{\partial p_k} = N(\psi(P) - \psi(p_k)) + \sum_{i=1}^N \log z_k(\mathbf{u}_i) \quad k = 1, \dots, D. \quad (22)$$

In the following section, a model is introduced by letting the scale parameters b_{ik} be functions of covariates.

4 Example

Consider the "Arctic Lake" example of Coakley and Rust (1968) and utilized by Aitchison (1986) for simplicial regression, who says

In sedimentology, specimens of sediments are traditionally separated into three mutually exclusive and exhaustive constituents -sand, silt and clay- and the proportions of these parts by weight are quoted as (sand, silt, clay) compositions. The (sand, silt, clay) compositions of 39 sediment samples at different water depths in an Arctic lake were recorded. Again we recognize substantial variability between compositions. Questions of obvious interest here are the following. Is sediment composition dependent on water depth? If so, how can we quantify the extent of the dependence? If we regard sedimentation as a process, do these data provide any information on the nature of the process? Even at this stage of investigation we can see that this may be a question of compositional regression.

The dataset can be found e.g. in van den Boogaart et al. (2014). This example was also used by Monti et al. (2015) within the scaled Dirichlet context. The goal is to model the effect of the depth covariate on the sediment composition. In their analysis, Monti et al. (2015) model the shape parameters p_k as functions of the depth. Here, the scale parameters are made dependent on the depth, which seems more natural.

4.1 Model and estimation

In the present context, the relationship between the sediment composition and the depth is modeled through the scale parameters.

With $d_i = \log(\text{depth})$ for composition $i, i = 1, \dots, 39$, the ratios of scales (i.e. the first two ilr components of the scale parameters $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3}), i = 1, \dots, 39$) depend on four parameters $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}$ in the following way

$$\begin{array}{lll} \text{silt/sand:} & b_{i2}/b_{i1} & = g_{i1} = \exp(\beta_{10} + \beta_{11}d_i) \\ \text{clay}/\sqrt{\text{silt*sand}}: & b_{i3}/\sqrt{b_{i1}b_{i2}} & = g_{i2} = \exp(\beta_{20} + \beta_{21}d_i) \\ \text{Constraint:} & b_{i1} + b_{i2} + b_{i3} & = 1. \end{array} \quad (23)$$

Thus the scale parameters b_{i1}, b_{i2}, b_{i3} are computed from Equation (23),

$$b_{i1} = \frac{1}{1 + g_{i1} + \sqrt{g_{i1}g_{i2}}} \quad b_{i2} = \frac{g_{i1}}{1 + g_{i1} + \sqrt{g_{i1}g_{i2}}} \quad b_{i3} = \frac{\sqrt{g_{i1}}g_{i2}}{1 + g_{i1} + \sqrt{g_{i1}g_{i2}}}. \quad (24)$$

Then the partial derivatives of the scales parameters with respect to the model parameters $\beta_{r\alpha}, r = 1, 2; \alpha = 0, 1$ are

$$\frac{\partial b_{ik}}{\partial \beta_{1\alpha}} = \frac{\partial b_{ik}}{\partial g_{i1}} \frac{\partial g_{i1}}{\partial \beta_{1\alpha}} \quad \frac{\partial b_{ik}}{\partial \beta_{2\alpha}} = \frac{\partial b_{ik}}{\partial g_{i2}} \frac{\partial g_{i2}}{\partial \beta_{2\alpha}} \quad \alpha = 0, 1, \quad k = 1, 2, 3.$$

The other derivatives are zero.

We have

$$\begin{array}{ll} \frac{\partial b_{i1}}{\partial g_{i1}} = -b_{i1}^2 \left[1 + \frac{g_{i2}}{2\sqrt{g_{i1}}} \right] & \frac{\partial b_{i1}}{\partial g_{i2}} = -b_{i1}^2 \sqrt{g_{i1}} \\ \frac{\partial b_{i2}}{\partial g_{i1}} = -b_{i1}^2 g_{i1} \left[1 + \frac{g_{i2}}{2\sqrt{g_{i1}}} \right] + b_{i1} & \frac{\partial b_{i2}}{\partial g_{i2}} = -b_{i1}^2 g_{i1}^{3/2} \\ \frac{\partial b_{i3}}{\partial g_{i1}} = -\frac{\partial b_{i1}}{\partial g_{i1}} - \frac{\partial b_{i2}}{\partial g_{i1}} & \frac{\partial b_{i3}}{\partial g_{i2}} = -\frac{\partial b_{i2}}{\partial g_{i2}} - \frac{\partial b_{i3}}{\partial g_{i2}}. \end{array}$$

Using these relationships, we find

$$\frac{\partial \ell}{\partial \beta_{r\alpha}} = \sum_{i=1}^{39} \sum_{k=1}^3 \frac{\partial \ell}{\partial b_{ik}} \frac{\partial b_{ik}}{\partial \beta_{r\alpha}}, \quad r = 1, 2, \quad \alpha = 0, 1. \quad (25)$$

Likelihood equations The likelihood equations are obtained by equating to zero Equations (21), (22) and (25). The model has thus 8 parameters, a, p_1, p_2, p_3 and $\beta_{r\alpha}, r = 1, 2, \alpha = 0, 1$, which is quite a lot for 39 observations. For this reason, some constraints were added.

Constraints

1. To avoid numerical difficulties, $a \geq 0.1$ is specified.
2. The constraint $ap_k \geq 2.1$ is introduced, in order to guarantee the existence of the second order moments of ratios of parts, see Equation (16).
3. By definition, the constraint $p_k > 0$ must be fulfilled.

Nonlinear optimization with constraints A good introduction to nonlinear optimization with constraints is given by Madsen et al. (2004). The R package **alabama** (Varadhan, 2015) permits to maximize the likelihood with linear and nonlinear equality and inequality constraints. **alabama** is an acronym for "Augmented Lagrangian Adaptive Barrier Minimization Algorithm". Its purpose is the optimizing of smooth nonlinear objective functions with constraints. At each outer iteration, the algorithm defines adaptive barriers and invokes the optimization function **optim** from package **stat**. We use the function **auglag** from package **alabama** in the example, to estimate the parameters under three models with different constraints. A very handy feature of **auglag** is that the initial values need not to satisfy the constraints and that constraints on parameters can be introduced. A list of computation algorithms specifically for Dirichlet likelihood can be found in (Giordan and Wehrens, 2015).

Models Three models were estimated (see Tables 1-2)

Model (1), a model with fixed shape parameters ($a = 1, p_1 = p_2 = p_3 = 2.1$).

Model (2), the full model, under constraints 1. to 3.

Model (3), the scaled Dirichlet model, under constraints 1. to 3. and $a = 1$.

Table 1: Results for the Arctic lake example

	\hat{a}	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	\hat{p}_1	\hat{p}_2	\hat{p}_3	<i>AIC</i>
Initial values	1.00	-4.89	1.65	-7.24	1.92	2.10	2.10	2.10	
(1) Fixed shape parameters	1.00	-5.13	1.68	-6.72	1.80	2.10	2.10	2.10	199.30
(2) Full model	1.80	-8.26	1.74	-5.04	1.68	1.17	149.58	1.36	154.02
(3) Scaled Dirichlet	1.00	-9.18	1.70	-4.69	1.75	2.96	149.60	3.28	155.69

Inequality constraints in all cases: $a > 0.1$ and $p_j > 1.5/a, j = 1, 2, 3$;

(1) Shapes parameters a, p_1, p_2, p_3 fixed to their initial values;

(2) Full model, initial values given by (1);

(3) Scaled Dirichlet: $a = 1$, initial values given by (2), except a .

Table 1 shows the estimated parameters of the three models with the corresponding Akaike information criterion $AIC = -2\ell + 2(\# \text{ parameters})$. Initial values: the initial $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}$ are computed by linear regression on the ilr transforms, the shapes are chosen so that the inequality constraints are verified. The initial a is set to 1, in order to compare with the scaled Dirichlet case. The AIC of Model (1) is computed with 8 parameters. In Model (2), the initial values are given by the estimates of Model (1). In Model (3), the scaled Dirichlet case is estimated with starting values from Model (2), except a which is set equal to 1. According to the AIC criterion the scaled Dirichlet model is equivalent to the full model.

Table 2 summarizes the constraints involved in the different models. Inactive constraints have a Lagrangian parameter equal to zero. A blank means that the constraint was not specified in the corresponding model. It is interesting to see that in the full model (2), only one inequality constraint is active.

Figure 2 (top) depicts the best unbiased predictor of inverse of parts, see Equation (18). At the bottom, the predictor of parts defined in Equation (11) is in good agreement with the observed data.

4.2 Imputation

Sample 12 has an outlier value of 1/clay, see Figure 2, top. The best unbiased predictor given the sub-composition (sand,silt) is obtained from Equation (20). The 1-part sub-composition (clay)

Table 2: Values of the Lagrangian parameters for the three models

Constraints	Model (1)	Model (2)	Model (3)
$a > 0.1$	0.00	0.00	0.00
$ap_1 > 2.1$	0.00	1.34	0.00
$ap_2 > 2.1$	0.02	0.00	0.00
$ap_3 > 2.1$	0.00	0.00	0.00
$p_1 > 0$	0.00	0.00	0.00
$p_2 > 0$	0.00	0.00	0.00
$p_3 > 0$	0.00	0.00	0.00
$a = 1$	-36.32		-4.50
$p_1 = 2.1$	-2.18		
$p_2 = 2.1$	-6.09		
$p_3 = 2.1$	-1.12		

implies that $v = 1$. The sub-composition \mathbf{w} is (sand,silt)/(sand+slit) for sample 12. So X is the same as the value of the part for clay in the original composition. The scale factor for clay in sample 12 is $\mathbf{b}_1 = b_{12,3}$, and $\mathbf{b}_2 = (b_{12,1}, b_{12,2})$. Thus $\|\mathbf{v}/\mathbf{b}_1\|_a = ((1/b_{12,3})^a)^{1/a} = 1/b_{12,3}$.

$$E\left(\frac{1-X}{X} \middle| \mathbf{W} = \mathbf{w}\right) + 1 = E\left(\frac{1}{X} \middle| \mathbf{W} = \mathbf{w}\right) = 1 + \frac{1/b_{12,3}}{\|\mathbf{w}/\mathbf{b}_2\|_a} \frac{\Gamma(P_1 - 1/a)\Gamma(P_2 + 1/a)}{\Gamma(P_1)\Gamma(P_2)}.$$

In Table 3, the column "Observed inverse" is the observed value of 1/clay for samples 11 (inlier) and 12 (outlier), the following column contains the unconditional model prediction (i.e. based on log(depth) only) and the last, the prediction given log(depth) and \mathbf{w} . The knowing of the sub-composition \mathbf{w} almost doubles the unconditional predicted value of 1/clay for this sample. By comparison, the unconditional model prediction for sample 11 is similar to the one for sample 12, but the conditional prediction is quite near to the observed value. The evaluations are similar when based on the full model (2) or on the scaled Dirichlet model (3).

Table 3: Imputation of 1/clay for sample 11 (inlier) and 12 (outlier).

	Sample	Observed inverse	Unconditional prediction	Conditional prediction
Full model (2)	11	15.385	9.624	15.122
Scaled Dirichlet (3)	11	15.385	9.788	15.551
Full model (2)	12	166.667	8.479	14.348
Scaled Dirichlet (3)	12	166.667	8.594	14.615

4.3 Simulation

There is clearly an over-fit, when 8 parameters are fitted on 39 observed compositions. In order to evaluate the obtained results, the following simulation was set up.

1. The log(depths) are approximately uniformly distributed. The ordered log(depth) are regressed against the uniform quantiles $1/40, \dots, 39/40$. A sample of size 7800 ($= 200 \times 39$) is generated from $U(0, 1)$ and transformed by the regression equation, resulting in 7800 random log(depth), $d_i, i = 1, \dots, 7800$, say.
2. Independent triples of generalized Gamma variables $GG(\hat{a}, 1, \hat{p}_j), j = 1, 2, 3$ are generated using the parameters of the full model in Table 1. Each component of triple $i, i = 1, \dots, 7800$ is multiplied by b_{ij} , where b_{ij} is computed from Equations (23) - (24), using the $\hat{\beta}_{k\ell}, k = 1, 2, \ell = 0, 1$ from Table 1 and the d_i obtained above.

3. Each triple is divided by its sum.

First Karush-Kuhn-Tucker condition (KKT) was met 173 times out of 200 for the full model and 152 times for the model with the constraint $a = 1$. The algorithm converged in all 200 cases. Table 4 shows the results.

Table 4: Simulation results.

	a	β_{10}	β_{11}	β_{20}	β_{21}	p_1	p_2	p_3
Truth	1.80	-8.26	1.74	-5.04	1.68	1.17	149.58	1.36
Overall estimate, full model	1.76	-8.33	1.76	-5.08	1.70	1.20	142.13	1.39
Simulation, full model, mean	1.92	-7.47	1.76	-5.51	1.70	1.72	53.76	1.99
Simulation, full model, sd	0.85	0.85	0.14	0.65	0.14	1.31	38.85	1.53
Overall estimate, model with $a=1$	1.00	-9.66	1.76	-4.48	1.70	3.10	199.45	3.66
Simulation, model with $a=1$, mean	1.00	-9.12	1.77	-4.76	1.70	3.63	170.96	4.20
Simulation, model with $a=1$, sd	-	1.05	0.14	0.76	0.14	1.17	107.64	1.17

Table 4 shows the simulation results. The overall estimates (using the 7800 generated values simultaneously) for the full model are quite near to the "truth", i.e. the estimates of the original full model in Table 2. The mean and variance estimates for the 200 batches of 39 data points show a large variability as expected. Notice that the parameters with the lower standard deviation are the coefficients of $\log(depth)$.

4.4 Comparison of models

It may be of interest to compare the different models for the Arctic lake example.

1. In Aitchison's simplicial regression based on the logistic-normal distribution, the model is

$$\begin{aligned} \log(sand/clay) &= c_{10} + c_{11} \log(depth) + e_1, \\ \log(silt/clay) &= c_{20} + c_{21} \log(depth) + e_2, \end{aligned}$$

where the error terms e_i are independent normally distributed as $N(0, \sigma_i)$, $i = 1, 2$. With $\mathbf{c}_0 = (c_{10}, c_{20}, 1)$ and \mathbf{c}_1 likewise,

$$E_A(sand, silt, clay) = \mathcal{C} \exp(c_{10}, c_{20}, 1) \oplus \mathcal{C} \exp(c_{11}, c_{21}, 1) \odot \log(depth)$$

Aitchison's model has six parameters $(c_{10}, c_{11}, c_{20}, c_{21}, \sigma_1, \sigma_2)$.

2. In Monti et al. (2015), the scaled Dirichlet is used, with the shape parameters depending on $\log(depth)$. The model is

$$p_k = \delta_{k0} + \delta_{k1} \log(depth) \quad k = 1, 2, 3.$$

Further two scale parameters b_1, b_2 (not depending on depth) are introduced. They correspond to the scales of the ilr components. In this setting, the expected composition is

$$E_A(sand, silt, clay) = \text{ilr}^{-1}(b_1, b_2) \oplus \mathcal{C} \exp(\psi(p_1), \psi(p_2), \psi(p_3))$$

This model has eight parameters.

3. In the present SGB model, the p_i 's are constant shape parameters and the scales depend on $\log(depth)$, see Section 4.1. The expectation is, with $d = \log(depth)$,

$$\begin{aligned}\mathbf{b} &= \mathcal{C} \exp \left(0, \beta_{10} + \beta_{11}d, \frac{1}{2}\beta_{10} + \beta_{20} + (\frac{1}{2}\beta_{11} + \beta_{21})d \right) \\ &= \mathcal{C} \exp \left(0, \beta_{10}, \frac{1}{2}\beta_{10} + \beta_{20} \right) \oplus \mathcal{C} \exp \left(0, \beta_{11}, \frac{1}{2}\beta_{11} + \beta_{21} \right) \odot d \\ &= \boldsymbol{\beta}_0 \oplus \boldsymbol{\beta}_1 \odot d. \\ \mathbf{E}_A(sand, silt, clay) &= (\boldsymbol{\beta}_0 \oplus \boldsymbol{\beta}_1 \odot d) \oplus \mathcal{C} \exp(\psi(p_1)/a, \psi(p_2)/a, \psi(p_3)/a) \\ &= \tilde{\boldsymbol{\beta}}_0 \oplus \boldsymbol{\beta}_1 \odot d,\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_0 = \mathcal{C} \exp(\psi(p_1)/a, \psi(p_2)/a, \psi(p_3)/a) \oplus \boldsymbol{\beta}_0$.

This model has eight parameters also. The expected composition has the same form as in Aitchison's model, but the parameters have another interpretation.

5 Discussion

In this paper, compound distributions of a random vector \mathbf{Y} are considered, where the compounding parameter acts on the scale. We have given the conditions under which the closed random vector $\mathcal{C}(\mathbf{Y})$ does not depend on the mixing distribution. Then an extension of the Dirichlet family of distributions has been proposed. In this family, starting with independent Y_k , ($k = 1, \dots, D$) following generalized Gamma(a_k, b_k, p_k) distributions, it has been shown that if and only if $a_k = a$ for all k , the distribution of $\mathcal{C}(\mathbf{Y})$ does not depend on the mixing scheme. In the $a =$ constant case, the distribution of the composition is easily obtained and its properties generalize those of the scaled Dirichlet of Monti et al. (2015, 2016). Those authors use another model where the covariate acts on p_1, \dots, p_D . The present model is similar to Aitchison's simplicial regression, but with other distributional assumptions. The shape parameters a, p_1, \dots, p_D modify the dispersion and skewness of the compositional parts and implies a perturbation on the expected composition.

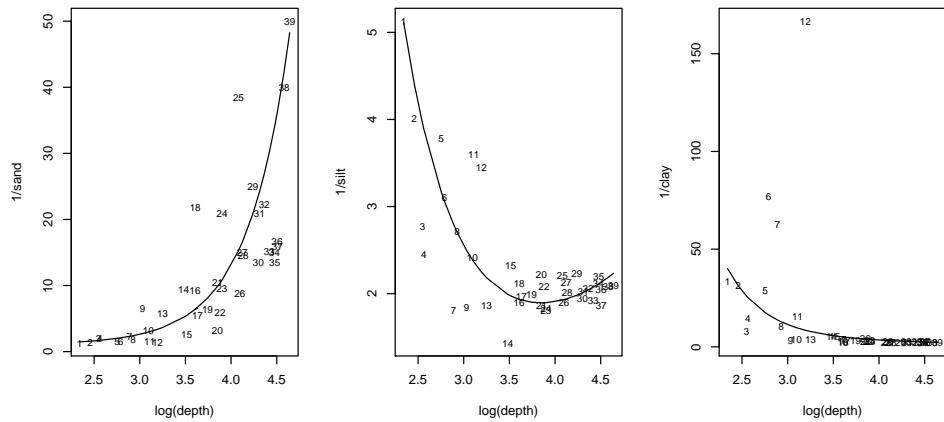
REFERENCES

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by the Blackburn Press, London (UK)).
- Connor, R. J. and J. E. Mosimann (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association* 64 (325), 194–206.
- Craiu, M. and V. Craiu (1969). Repartitia Dirichlet generalizatá. *Analele Universitatii Bucuresti, Mathematicá-Mecanicá* 18, 9–11.
- Giordan, M. and R. Wehrens (2015). A comparison of computational approaches for maximum likelihood estimation of the Dirichlet parameters on high-dimensional data. *SORT* 39(1), 109–126.
- James, I. R. (1975). Multivariate Distributions which have Beta Conditional Distributions. *Journal of the American Statistical Association* 70(351), 681–684.
- Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions, Volume 1, Models and Applications*. John Wiley & Sons.
- Madsen, K., H. Nielsen, and O. Tingleff (2004). Optimization With Constraints. Informatics and Mathematical Modelling, Technical University of Denmark.

- Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2003). Distributions on the simplex. In S. Thió-Henestrosa and J. M. Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*.
- Monti, G., G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. Egozcue (2015). Shifted-Dirichlet Regression vs Simplicial Regression: a comparison. In S. Thió-Henestrosa and J. M. Fernández (Eds.), *Proceedings of the 6th International Workshop on Compositional Data Analysis*.
- Monti, G., G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. Egozcue (2016). A regression model for compositional data based on the Shifted-Dirichlet distribution. Submitted.
- Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). Notes on the scaled Dirichlet distribution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis. Theory and applications*. Wiley.
- Ongaro, A., S. Migliorati, and G. Monti (2008). A new distribution on the simplex containing the Dirichlet family. In J. Daunis-i Estadella and J. E. Martín-Ferntández (Eds.), *Proceedings of the 3rd International Workshop on Compositional Data Analysis*.
- Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2007). Lecture Notes on Compositional Data Analysis.
- Rayens, W. S. and C. Srinivasan (1994). Dependence Properties of Generalized Liouville Distributions on the Simplex. *Journal of the American Statistical Association* 89(428), 1465–1470.
- van den Boogaart, K. G., R. Tolosana, and M. Bren (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1.
- Varadhan, R. (2015). *alabama: Constrained Nonlinear Optimization*. R package version 2015.3-1.
- Yang, X., E. Frees, and Z. Zhang (2011). A generalized Beta-copula with applications in modeling multivariate long-tailed data. *Insurance: Mathematics and Economics* 49(2), 265–284.

Relationship between inverse of parts and depth

Estimated parameters, full model



Relationship between parts and depth

Estimated parameters, full model

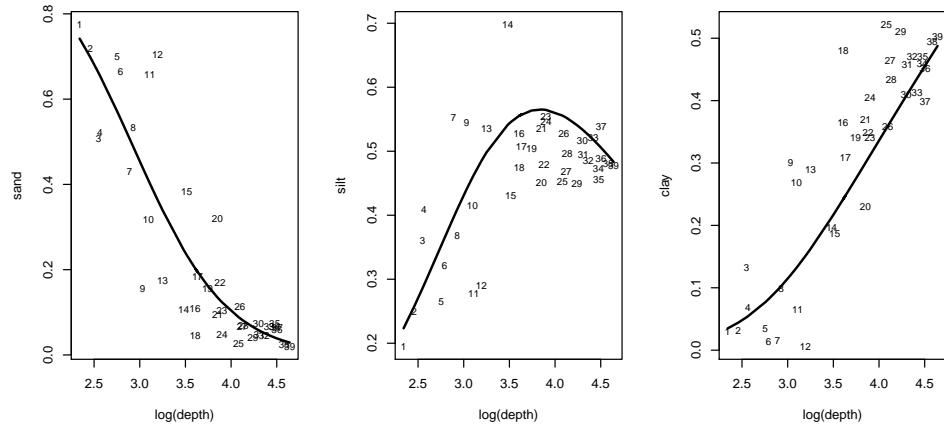


Figure 2: Top: Inverse of parts. Lines according to Equation (18). Bottom: parts and conditional expectation given depth. Lines according to Equation (10). Observations are depicted by their sample number.

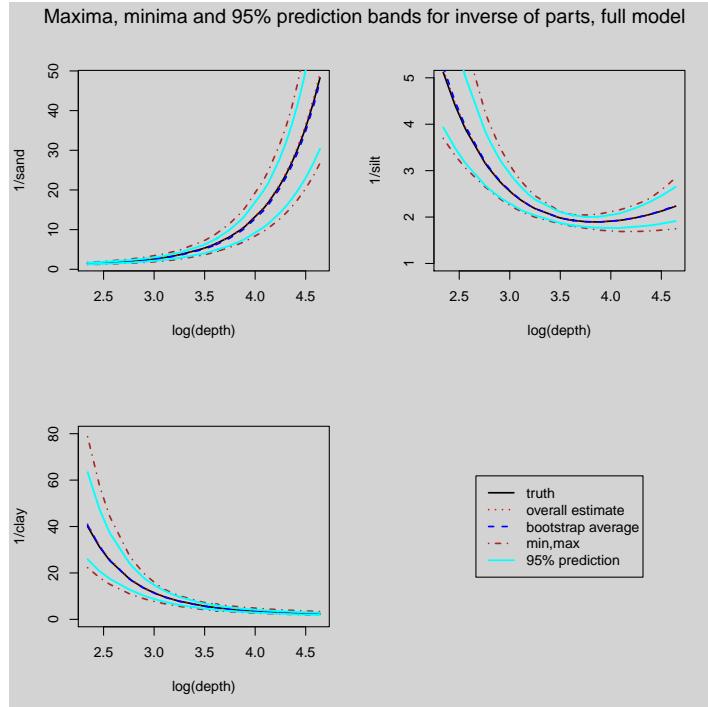


Figure 3: Parametric bootstrap estimates. Black: original estimate; red: estimate based on all 7800 samples; blue: mean of the 200 bootstrap samples; brown: minimum and maximum of the bootstrap samples; cyan: 95% bootstrap confidence limits.

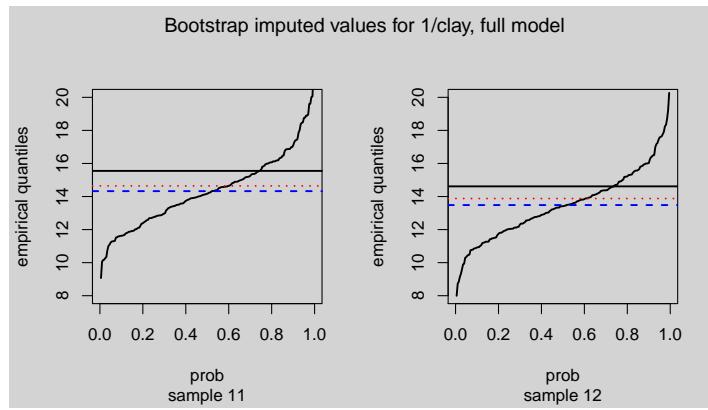


Figure 4: Parametric bootstrap conditional estimates of $1/\text{clay}$ given the sub-composition (sand,silt). Lines: (black, continuous), original estimate (see table 3); (red,dotted), estimate based on all 7800 samples, (blue,dashed): mean of the 200 bootstrap samples.

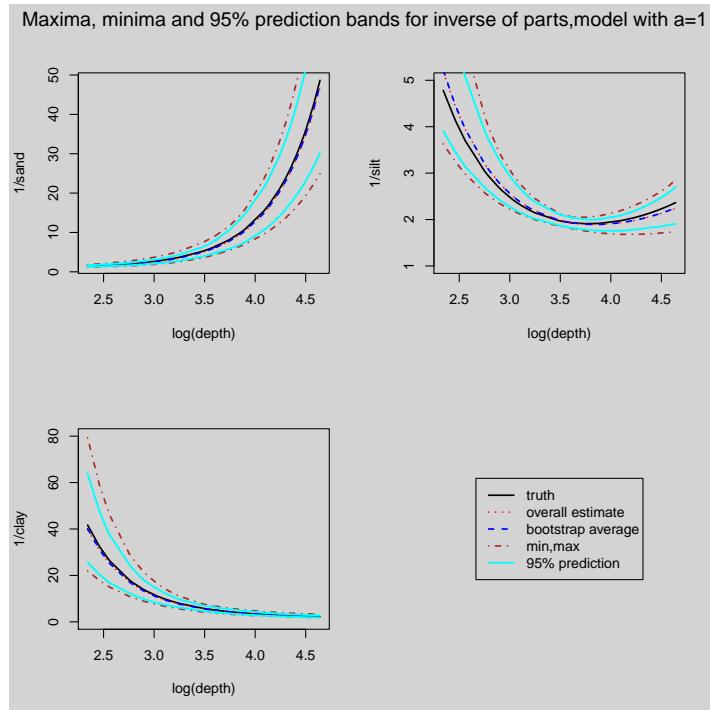


Figure 5: Parametric bootstrap estimates. Black: original estimate; red: estimate based on all 7800 samples; blue: mean of the 200 bootstrap samples; brown: minimum and maximum of the bootstrap samples; cyan: 95% bootstrap confidence limits.

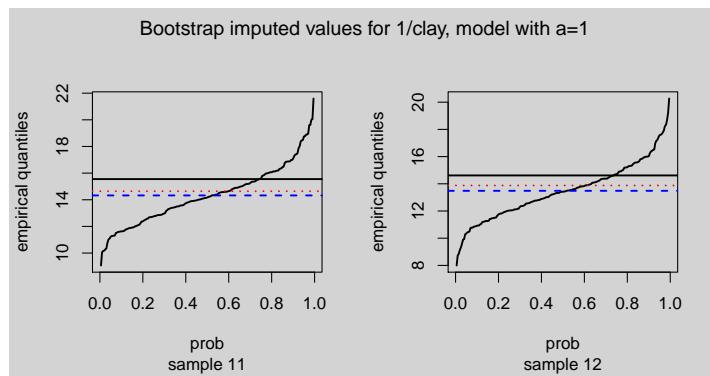


Figure 6: Parametric bootstrap conditional estimates of $1/\text{clay}$ given the sub-composition (sand,silt). Lines: (black, continuous), original estimate (see table 3); (red,dotted), estimate based on all 7800 samples, (blue,dashed): mean of the 200 bootstrap samples.

Application of Aitchison metrics on magnetic resonance images with multiple contrasts at ultra high field (7 Tesla) to investigate compositional characteristics of brain tissues in living humans

O.F. Gulban

Maastricht University, Maastricht, The Netherlands

Abstract

Images of living human brains can be acquired non-invasively by using magnetic resonance imaging (MRI). Different scanning parameters weight the image contrast to different tissue properties. A few examples of these differently weighted images are; T1 weighted (T1w) images to maximize the contrast between white matter and gray matter tissues, proton density weighted (PDw) for measuring concentration of hydrogen atoms and T2* weighted (T2*w) for creating a contrast highlighting the iron content. These multi-modal images are commonly combined pairwise (using e.g. simple ratio images) to mitigate intensity inhomogeneities or to reveal specific tissue properties such as gray matter myelination. A principled way to combine more than two images at once is to consider multi-modal MRI data as compositions. The present study applies the concepts of compositional data analysis to multi-modal MRI data in order to simultaneously reduce artefactual intensity inhomogeneities and highlight specific tissue characteristics. To this end, brain images of a living human with three different contrasts (T1w, PD, T2*w) were acquired at ultra high field (7 Tesla) MRI scanner and Aitchison metrics were used to create virtual MR contrasts similar to conventional ratio images. In addition, isometric logratio transformed coordinates of the tissue compositions were explored with two-dimensional transfer functions to probe meaningful compositional characteristics of brain tissues.

Key words: compositional data analysis, magnetic resonance imaging, MRI, image fusion, RGB, HSI, color space

1 Introduction

Compositional data analysis can be applied to images with multiple values for every picture element such as color images. Color images are commonly stored as triplets of non-negative integers representing the additive primary colors; red, green, blue (RGB). The creation of color images by assigning different sources of information to RGB channels to enhance the interpretation of multi-modal data is the simplest image fusion method (Pohl and Genderen, 2016, p. 96). Once the color image is formed in RGB format, the color space transformation from RGB to hue, saturation, intensity (HSI) is commonly the first step to improve image visualization. This transformation can be depicted as the projection of a cube representing the possible combinations of RGB values (also called as the RGB color cube) to a hexcone model (Schowengerdt, 2012, p. 225). In the hexcone model the height of the cone represents the intensity (sum of the RGB magnitudes) and the hexagon represents the scale invariant saturation and hue components (spectral components). An alternative way to think about saturation and hue in the RGB to HSI transformation is to consider the isometric logratio (ilr) transformation after applying closure to the RGB color cube. The real space coordinates can be used in the conventional way to derive hue and saturation since this transformation of the RGB color cube also forms a hexagon (the ilr transformation and closure operator are standard operations in compositional data analysis; Pawlowsky-Glahn et al., 2016).

Although useful, the representation of multi-modal data by means of RGB or HSI channels limits this application to cases with up to three values per picture elements. In contrast, the compositional data analysis interpretation of the RGB to HSI color space transformation provides a principled framework to operate on images with more than three information channels, and thereby overcomes this limitation of the HSI-based method. For instance, the saturation and hue of an image with four values in every pixel can be derived from the ilr transformed coordinates of the channel compositions, while the intensity can still be calculated as the sum of the channel magnitudes.

Magnetic resonance imaging (MRI) data is suitable to illustrate the compositional data analysis on multi-modal images because (I) different scanning parameters weight the image contrast to different tissue properties thus providing multitude of informative images; (II) although the MRI data is of complex type, the magnitude images are often analyzed which means that the images are bound to have non-negative values. A few of these images with different contrast weightings are: T1 weighted (T1w) images to maximize the contrast between white matter and gray matter tissues, proton density weighted (PDw) for measuring concentration of hydrogen atoms and T2* weighted (T2*w) for creating a contrast highlighting the iron content. It should be noted that the combination of information coming from different medical imaging modalities (e.g MRI, computed tomography, positron emission tomography) has been widely explored (for a review see: James and Dasarathy, 2014) but not under the compositional data framework. When considering MRI information alone, multi-modal MRI data have been used for tissue segmentation purposes (for a review see: Helms, 2016) and ratio images between pairs of modalities have been used to enhance/improve contrast visualization. For instance the ratio between T1w and PD images is used to mitigate artefactual image intensity inhomogeneities due to variations in sensitivity of the measurement method across brain areas (Van de Moortele et al. 2009) and the ratio between T1w and T2w or T2*w images is used to reveal relevant tissue properties such as myelination in the cortex (Glasser and Van Essen 2011; De Martino et al., 2014).

The compositional data framework is ideally suited for the analysis and visualization of multi-modal MRI data as it provides a principled way to combine multiple images (i.e. more than a pair as in ratio type of approaches). The present study applies the concepts of compositional data analysis to multi-modal MRI data in order to simultaneously reduce artefactual intensity inhomogeneities and highlight specific tissue characteristics. To this end, brain images of a healthy volunteer were acquired with three different contrasts (T1w, PD, T2*w) at ultra-high field 7 Tesla MRI scanner (for a review see: Ugurbil, 2014) and Aitchison metrics (Pawlowsky-Glahn et al., 2015) were used to create virtual MR contrasts. In addition, the ilr transformed coordinates of the tissue compositions were explored with 2D transfer functions to probe meaningful compositional

characteristics of brain tissues.

2 Methods

2.1 Data acquisition

Whole head T1 weighted, PD and T2* weighted images at 0.7 mm isotropic voxel (3D picture elements are referred to as voxels in the context of MRI data) resolution were acquired in one male participant using a 3D magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence with a 32-channel head-coil (Nova Medical) on a 7 Tesla whole-body scanner (Siemens; for details regarding the acquisition parameters see De Martino et al., 2014).

2.2 Preprocessing

Brain extraction was performed based on the PD image using FSL-BET (version 2.1; Smith, 2002), the resulting brain mask was applied to the T1w and T2*w images. Voxels with a value equal to zero were set to one because one is the smallest non-zero magnitude that the images can have considering the 16-bit unsigned integer precision of the data.

2.3 Analysis

The three three-dimensional images (T1w, PD, and T2*w) were concatenated, obtaining a four-dimensional image. Voxel-wise closure was performed to obtain the barycentric coordinates of the compositions. The distribution consisting of 4,543,582 compositions was centered and standardized inside the simplex to enhance the visualization of distribution. The ilr transformation was performed to acquire real space coordinates of the compositions. The ilr transformed coordinates of the compositions were stored in an additional image dimension so that they can easily be associated with spatial coordinates of the original brain images. The analysis pipeline is implemented in a free and open source Python package (available at <https://github.com/ofgulban/tetrahydra>; version 0.1.1; Gulban 2017)

In order to reveal the correspondence of compositional data clusters and brain tissue classes, the image space coordinates of the compositions were visualized in form of a 2D histogram with a logarithmic color map. 2D histogram was used to bin ilr transformed coordinates instead of binning the simplicial coordinates in a ternary plot for computational efficiency required in an interactive display using transfer function widgets. The clusters were interactively explored with two-dimensional transfer function widgets implemented as a part of the free and open source Python package Segmentator (available at: <https://github.com/ofgulban/segmentator>; version 1.1.1; Gulban and Schneider, 2016).

3 Results

Figure 1A depicts one transversal and one sagittal slice of the brain extracted MRI data in image space. Due to the different image contrasts, each image reflects different properties of the tissues. For instance, the cerebrospinal fluid in the ventricles is very dark in the T1w image, but bright in the T2*w image. On the other hand, the sagittal sinus (visible in the sagittal slice) is bright in the T1w and PD images and very dark in the T2*w image. It can also be noted that the smooth, artefactual intensity inhomogeneities are similar across images (note the overall intensity differences that covary across the three images). Figure 1B show the Aitchison norm image computed by

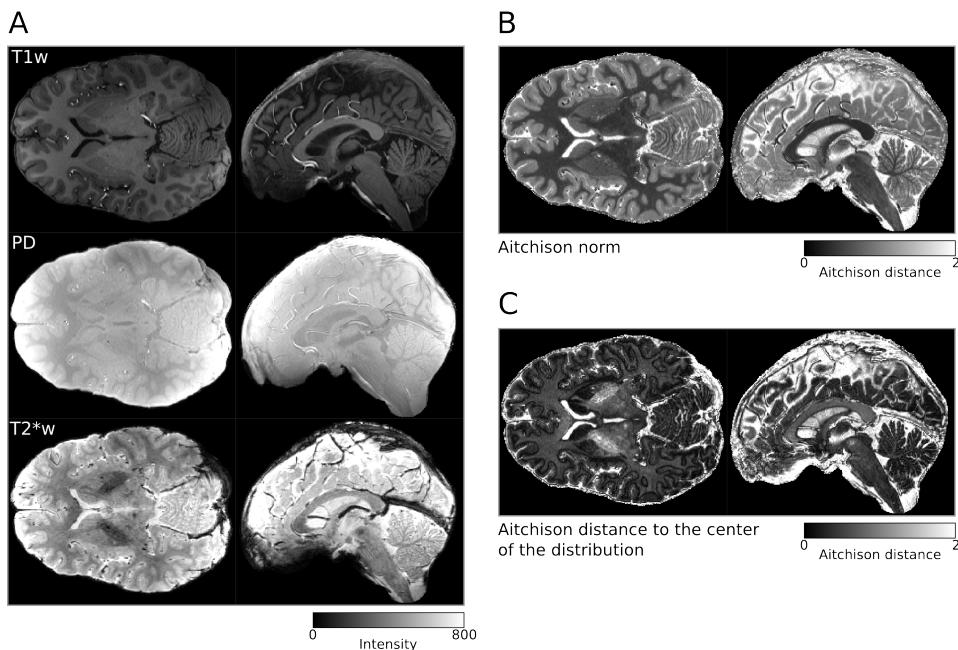


Figure 1: (A) Magnitude images of the T1w, PD and T2*w MRI data. (B) Voxel-wise Aitchison norm (aitchison distance to the center of the simplex) image of the brain extracted MRI data composition. (C) Aitchison distance to the center of the compositional distribution. Two slices of the brain extracted MRI data are visible in all panels, transversal slice (left hand side) and sagittal (right hand side) relative to the panels.

considering the T1w, PD and T2*w images as a composition after mapping the voxel values from the real space (R^3) to the simplex space (S^3) with closure operation. This image is similar to a conventional ratio images, however different because it can be computed from more than two images. The Aitchison distance to the center of the distribution inside the simplex is depicted in Figure 1C. In this image, the interface between white matter and gray matter is very dark, which demonstrates that the center of the distribution falls within the transition of white and gray matter tissues. The Aitchison distance image illustrates that the compositional data analysis framework can be used to create virtual contrast, which can be used to create tissue membership maps. For instance, the hypo-intense white matter/gray matter interface may be used to assist in the common task of segmenting those two tissue types.

The 2D histogram of the ilr transformed coordinates of the MRI brain image compositions (Figure 2A) contains 3 heavy clusters. Probing the real coordinates of the compositional data points using interactive 2D transform function widgets reveals that these clusters represent the white matter, gray matter and cerebrospinal fluid. At more peripheral coordinates, the arteries and sinuses can be seen. Although both of these are blood vessels, the difference between arteries and sinuses is meaningful because sinuses contain mostly deoxygenated hemoglobin, leading to a rapid decay of the MRI signal. In contrast, arteries contain oxygenated blood with slower MR signal decay. The color brain image (in Figure 2B) created by assigning T1w, PD, T2*w images to red, green, blue color display channels depicts the fused picture of the MRI data. The labels pointing to the tissues indicated in Figure 1A shows relation of the compositional characteristics of with the coloration. To illustrate this point, it should be noted that the arteries have mostly reddish-white colors and the sinuses appear in green, which corresponds to the areas delineated for these tissues

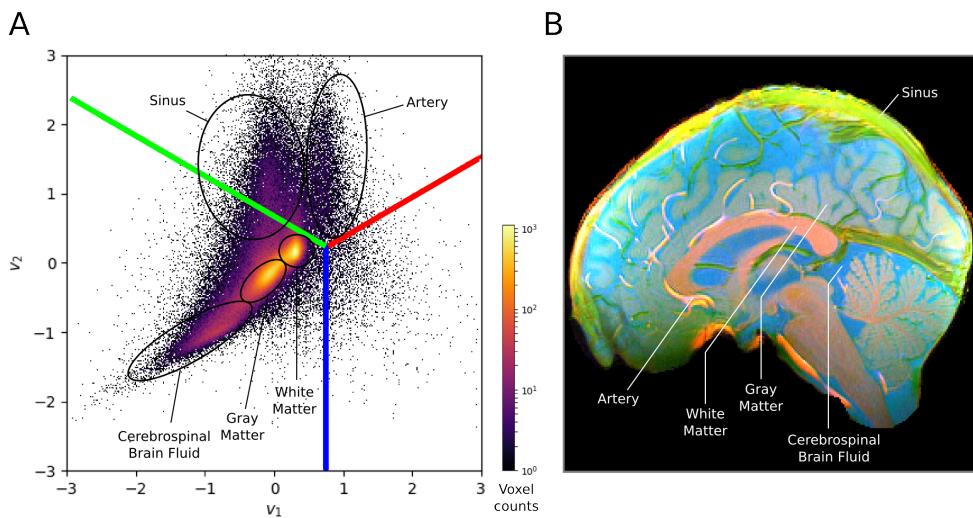


Figure 2: (A) 2D histogram of the ilr transformed coordinates of the MRI brain data compositions. Projections of the primary axes of RGB color cube are embedded (red line for T1w, green line for PD, blue line for $T2^*$ w) to provide an intuitive reference for the characteristics of the compositions because only having the real space coordinates (v_1 , v_2) are harder to interpret. (B) A sagittal slice of the brain extracted MRI data in false color. T1w image is assigned to the red channel, PD is assigned to the green channel, $T2^*$ is assigned to blue channel. The rendering of the image contrast is enhanced with the simplest color balance algorithm (Limare et al., 2011). The tissue labels are overlaid on top in both panels.

in Figure 1A when the positions of clusters are considered relative to the embedded RGB color cube axes. Similarly the compositional change from white matter to gray matter to cerebrospinal fluid can be seen as an approximately straight line in Figure 1A which corresponds to the change from red heavy color of white matter to cyan of cerebrospinal fluid in Figure 1B.

4 Discussion

The application of compositional data analysis methods to the multi-modal MR images provides a promising extension to color space based image fusion methods. The scale invariance principle of compositional data can be leveraged to mitigate intensity inhomogeneities. This is particularly beneficial in case of state of the art ultra high field MRI data, where intensity inhomogeneities are much more pronounced than at lower field strengths. However, proposed method assumes similar intensity inhomogeneity profiles across images. If this assumption is not fulfilled, intensity inhomogeneity correction algorithms may be employed as a preprocessing step, and the proposed method may still be used for data exploration and tissue classification. Moreover, image regions affected by strong artifacts, such as heavy signal drop-outs, should be discarded from the compositional analysis before attempting tissue classification.

Although the present study only considered MRI data with three contrasts, compositional data analysis provides a framework to analyze an arbitrary number of image modalities, overcoming a major limitation of color space based image fusion methods. Thus, combining multi-modal images using Aitchison metrics opens the possibility to create a generalization of conventional ratio images with an arbitrary number of image components. In addition, the proposed method may be applied

to investigate the compositional characteristics of temporal processes such as the blood oxygenation level dependent signal in multi-echo functional MRI data (in which multiple images of the brain are acquired in quick succession, resulting in a five-dimensional image, in which the fifth dimension contains additional information about metabolic processes, in addition to the three spatial and one temporal dimensions). Another potential use is the application of clustering methods to the compositional coordinates of multi-contrast MRI data for creating binary segmentations of cortical or sub-cortical tissues. Moreover, the method may prove useful in the classification of diseased tissue. Further investigation will be necessary to establish the relevance of the proposed applications of compositional data analysis to MRI data.

Acknowledgements

The data was acquired thanks to Federico De Martino, under the project supported by NWO VIDI grant 864-13-012. The author O.F.G. was also supported by the same grant. In addition, I thank Ingo Marquardt for language editing and proofreading.

References

- De Martino, F., Moerel, M., Xu, J., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E. (2014). High-Resolution Mapping of Myeloarchitecture In Vivo: Localization of Auditory Areas in the Human Brain. *Cerebral Cortex*, 25(10), 3394-405.
- Glasser, M. F., Van Essen, D. C. (2011). Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *Journal of Neuroscience*, 31(32), 11597-616.
- Gulban, O.F., (2017) Tetrahydra v0.1.1. *Zenodo*. doi:10.5281/zenodo.571207.
- Gulban, O.F., Schneider, M. (2016). Segmentator v1.1.0. *Zenodo*. doi:10.5281/zenodo.157996.
- Helms, G. (2016). Segmentation of human brain using structural MRI. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 29(2), 111-124.
- James, A. P., Dasarathy, B. V. (2014). Medical image fusion: A survey of the state of the art. *Information Fusion*, 19(1), 4-19.
- Limore, N., Lisani, J., Morel, J., Petro, A. B., Sbert, C. (2011). Simplest Color Balance. *Image Processing On Line*, 1(1), 125-133.
- Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. (2015). *Modelling and Analysis of Compositional Data*. Chichester, UK: John Wiley & Sons, Ltd.
- Pohl, C., van Genderen, J. (2016). *Remote Sensing Image Fusion*. Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742: CRC Press.
- Schowengerdt, R. A. (2012). *Remote Sensing: Models and Methods for Image Processing*, 3rd edition. Burlington: Academic Press
- Smith, S.M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17(3), 143-155.
- Ugurbil, K. (2014). Magnetic Resonance Imaging at Ultrahigh Fields. *IEEE Transactions on Biomedical Engineering*, 61(5), 1364-1379.
- Van de Moortele, P.-F., Auerbach, E. J., Olman, C., Yacoub, E., Ugurbil, K., Moeller, S. (2009). T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. *NeuroImage*, 46(2), 432-46.

Forecasting patent filings at the European Patent Office (EPO) using compositional data analysis techniques

P. Hingley

European Patent Office, Bob-van-Benthem-Platz 1, 80469 Munich, Germany; *phingley@epo.org*

Abstract

A dynamic log-linear (DLL) regression model is fitted to time series data for forecasting future numbers of Total Filings at the European Patent Office. 28 source countries are included, with independent variables for Gross Domestic Product, Research and Development Expenditures and autoregressive terms. A breakdown of Total Filings according to industrial or technical areas of the underlying inventions induces additional compositions of the data to the breakdown by countries.

Some models of compositions with a total are considered. The additional forecasting power for modelling Total filings is assessed by considering a breakdown of the data into proportions according to three industrial areas (IAs): Electricals, Chemicals and Traditionals. The IAs Electricals and Traditionals have grown markedly since year 2000, while Chemicals has grown more slowly. Isometric log-ratio (ilr) terms were included in the DLL model to represent the IAs Electricals and Chemicals. This gave an improved least squares fit with statistically significant ilr terms.

A simpler CoDa based straight line regression model was fitted to the ilr terms to give forecasts for the proportions of the three IAs. The total was then included as an additional predictor for proportions. As a complementary experiment, a straight line regression model was fitted to Total Filings and then the ilr terms were added as additional predictors. These results are contrasted with each other and with the results from the DLL model.

CoDa techniques have a small beneficial effect on forecasting Total Filings. They can also be useful for analysing proportions of breakdowns of the total into classes in case there are many classes or some of the classes have high or low proportions.

Key words: CoDa with a total, dynamic log-linear regression, industrial areas, isometric log ratio, patent filings.

1 Introduction

In order to manage the financial budget of the European Patent Office (EPO), at the start of each year forecasters prepare a scenario for the numbers of Total Filings (TFs) for the current year and the following five years. Several sets of forecasts are obtained by different methods (Hingley and Nicolas, 2004, 2006), including a survey (EPO, 2017). One of the models is a dynamic log-linear (DLL) model, that considers the 27 leading countries of residence of the patent applicants and a residual rest-of-the-world term. This model has independent variables and autoregressive terms, with most variables being standardised by dividing by workforce population sizes before taking logarithms. The model has been shown to produce useful scenarios, even though the forecasts are found to be more optimistic than with some other methods. The goodness of fit is increased by splitting the variable for Gross Domestic Product (GDP) into trend and (business) cycle terms (Hingley and Park, 2017).

Applications for patents for inventions can be made for many commercially viable technical areas that can be classified under hierarchical schemes such as the International Patent Classification (IPC) (WIPO, 2017). The worldwide markets for processes and products according to the main classes can differ and so there can be different dynamics for the development of their time series. In theory, there is more information in the individual sub-time series so that modelling them and adding the results could give better forecasts for TFs. Aspects of this are explored here, by working with a simple breakdown of Total Filings (TFs) into three Industrial Areas (IAs):- Electricals, Chemicals and Traditionals.

One way to consider the IAs is to employ compositional data analysis (CoDa) on their proportions. The main aim remains to forecast the TFs time series, which is the sum of the filings over the IAs, rather than to study the evolution of the proportions. So it is appropriate to consider CoDa models of compositions with a total. Studies by Pawlovsky-Glahn (2014) and Coenders et al. (2015) suggested using geometric means from which a kind of total was obtained by multiplying by the sample size.

The approach taken here is to abstract isometric log ratio (ilr) terms for the IAs in order to see whether the performance of the DLL model can be improved by including them. This is assessed both in terms of goodness of fit of the model to the data and in terms of the plausibility of the forecasts. An analysis of a simple trend model is also described in order to consider the effects of adding the total as a predictor for the proportions of the IAs. Finally there is a complementary analysis to see what happens to a trend model for Total Filings when the ilr terms for the IAs are added as predictors for the total.

2 A DLL model for EPO Total filings for patents

A description of the use of this model is given by Hingley and Park (2015, 2016, 2017). The general form of the model, for transformed Total Filings from country i at year t without CoDa terms, is as follows.

$$\Delta \ln \left(\frac{P}{L} \right)_{it} = \alpha_i + \alpha_1 \Delta \ln \left(\frac{P}{L} \right)_{it-1} + \alpha_2 \Delta \ln \left(\frac{P}{L} \right)_{it-2} + \alpha_3 \Delta \ln \left(\frac{R}{L} \right)_{it} + \alpha_4 \Delta \ln \left(\frac{Y^T}{L} \right)_{it} + \alpha_5 \Delta u + \varepsilon_{it} \quad (1)$$

Where P is the number of EPO Total filings (TFs) from a source country, which is the sum of Euro-direct and PCT international phase filings;

L is the number of workers in the source country;

-1 and -2 indicate lags of one year and two years respectively;

R is Research and Development (R&D) expenditures, considered here as a stock variable comprising components from 0 to 5 year lags;

The GDP of the source country Y is split into two components:- Y^T is the “trend” level of output, and u is a business cycle variable (a ratio of cyclical GDP to trend GDP);

ε_{it} is an error term, assumed to be normal with constant variance;

$\ln()$ denotes natural logarithm;

Δ indicates year to year differences.

The counts P are available with breakdowns into the three IAs: Electricals (IA_E), Chemicals (IA_C) and Traditionals (IA_T). A current lack of detailed information at the level of each IA for the independent variables L , R , Y^T and u means that, at this first level of modelling, the same common values of these inputs are taken for fitting the models to all three breakdowns.

Standard CoDa regression models are fitted to these data in Section 4 below.

In order to incorporate compositional terms into the operational DLL model of Equation (1), use is made here of the suggestion by Pawlowsky-Glahn et al (2014) that the total and the proportions can be modelled together. Consider the three IAs and suppress the subscripts i and t . Call x_E , x_C and x_T the subtotals for country i at time t for IA_E , IA_C and IA_T respectively. The total to be modelled is $P = x_E + x_C + x_T$.

With three components (E, C, T), the ilr for component j is $ilr_j = \log\left(\frac{x_j}{(x_Ex_Cx_T)^{1/3}}\right)$. But the total depends on the sum of the components. To avoid collinearity when modelling the total, it is appropriate to estimate parameters for only two out of the three available ilr_j components - ilr_E (Electricals) and ilr_C (Chemicals) are chosen, avoiding ilr_T (Traditionals). The two additional ilr terms may improve the fit of the DLL model, because they allow for uneven changes of proportions according to IAs. The enhanced DLL model that is to be fitted is as follows.

$$\begin{aligned} \Delta \ln\left(\frac{P}{L}\right)_{it} = & \alpha_1 + \alpha_1 \Delta \ln\left(\frac{P}{L}\right)_{it-1} + \alpha_2 \Delta \ln\left(\frac{P}{L}\right)_{it-2} + \alpha_3 \Delta \ln\left(\frac{R}{L}\right)_{it} + \alpha_4 \Delta \ln\left(\frac{Y^T}{L}\right)_{it} + \alpha_5 \Delta u + \\ & \alpha_6 \Delta ilr_E + \alpha_7 \Delta ilr_C + \varepsilon_{it} \quad (2) \end{aligned}$$

The variables u and the ilr terms are essentially dimensionless and are not standardised by L .

A variant of model (2) is fitted in the next section. For the independent variables, some values are not known a-priori for later than about two years in the past. While the value of P for the second order auto-regressive term is in a sense known out to two years in the future, L has at least a two year reporting lag and so the standardised form of P is not fully known even for the preceding year. R is also reported with a two year lag. GDP per country (from which Y^T and u are derived using a Hodrick-Prescott filter) can be assumed known for the current year and is modelled for the future by a second order autoregressive process, rather than by using agency estimates. Timeliness of the ilr terms is discussed in the next section.

To determine inputs for the independent variables to the fitted model for forecasts in the future, linear trends are generally fitted to the 10 annual values up to the last reliable known data value. The projected values of the independent variables are considered fixed for making the forecasts and confidence limits for the forecasts are generated using the theoretical standard errors of the fitted values under this assumption. While it is recognised that these limits are too narrow, it is reasonable to take residual error estimates under these assumptions as a basis for comparing the fits of variants of the DLL models. The comparative widths of the confidence limits for the forecasts can also be considered.

3 Fitting the DLL model to EPO Total filings for patents

In this section, modelling is done with the data that were available in early 2017, although the forecasts to be described are not the current official forecasts. These analyses were done in R, with further processing in Excel.

The model is fitted to the transformation $\Delta \ln\left(\frac{P}{L}\right)$. The time series for the training data runs from

year 2001 up to year 2015, after preliminary testing showed no advantage by starting at earlier years. The practice is to fit models to data up to the end of the calendar year two years previously, because counts for the latest year may not yet be stable. A comparison of the one year forecasts to the outcome for the latest year can be used to compare effectiveness of alternative forecasting methods.

The following results involve aggregation over the 28 countries used in the models. Figure 1 shows the historical data broken down by IAs. Data for numbers of filings in IAs are considered to be reliable up to 2013 only, but proportions due to IAs for 2014 and 2015 are derived by ignoring the unassigned data when calculating them. Ilr values are trended thereafter in the forecasting region.

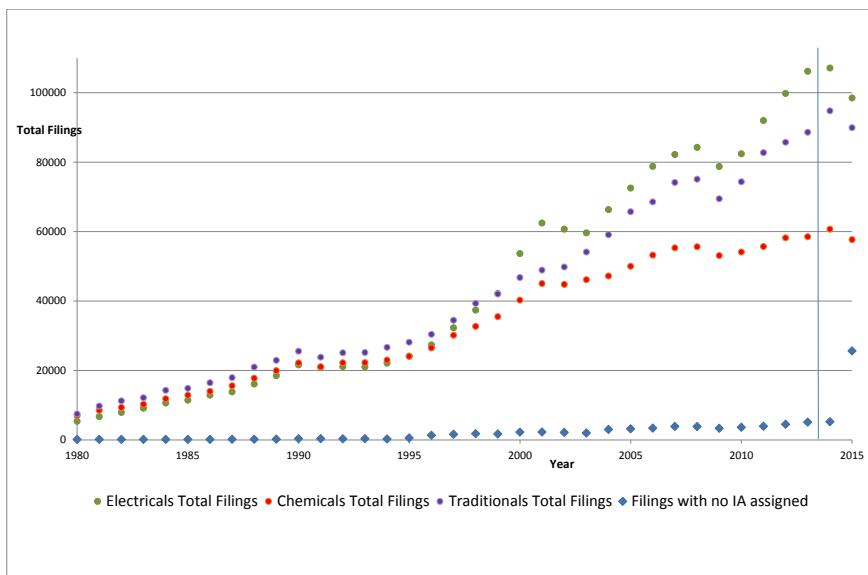


Figure 1: Filings by three industrial areas (IAs). The numbers of Filings with no IA assigned increases after 2013.

Firstly it was determined pragmatically that the AR1 term α_1 was not significant and so this was dropped from the model. Table 1 shows the parameter estimates that were obtained after fitting to each IA separately, to the model (1) without IA based ilr terms and to the model (2) that includes IAs. The estimates for the 28 country intercepts are not shown (but see Hingley and Park, 2016). Approximate significance at the 95 percent level is assigned when the absolute value of a parameter is more than 1.96 times its estimated standard error.

Table 1 demonstrates that model (2) fits better than model (1), because the observation standard error is smaller. Consider the statistical significance of the parameter estimates in Table 1, assessed approximately by checking whether the absolute values of the estimates are more than 1.96 times their standard errors. Five of the six parameter estimates in model (2) are statistically significant, which is a higher proportion significant than in model (1) or any of the sub-models fitted to the three separate IAs. The three sub-models show some differences to each other. Electricals has no significant parameters. Traditionals has a significant α_4 parameter, while Chemicals has significant

Table 1: Parameter estimates after fitting the various models. * indicates approximate significance at the 95 percent level (see text for details).

Parameter estimates	Separate Industrial Areas (IAs)			Total Filings	
	Electricity (E)	Chemistry (C)	Traditional (T)	With IAs (2)	Without IAs (1)
Autoregression	α_2	-0.01	-0.18*	-0.05	-0.09*
R&D Stock	α_3	0.29	0.27	0.09	0.36*
GDP trend	α_4	-0.25	2.51*	1.63*	1.26*
GDP cycle	α_5	0.39	0.2	0.44	0.53
ilr for Electricals	α_6				0.30*
ilr for Chemicals	α_7				0.21*
Observation standard error		0.1877	0.1653	0.1325	0.111
Residual degrees of freedom		388	388	388	388

α_2 and α_4 parameters. Both models (1) and (2) have significant α_2 , α_3 and α_4 parameters. Thus the R&D expenditures component that is represented by α_3 is only significant for the models for TFs and not for any of the models for the separate IAs. In model (2), the additional parameters α_6 (for ilr_E) and α_7 (for ilr_C) are both significant.

Table 2 shows the forecasts that were generated by the models for years 2016 to 2022. Figure 2 shows the Total Filings data and the forecasts. It can be seen that there is little difference between the forecasts for TFs with and without ilr terms included. On the other hand, the forecasts generated by applying the model (without ilr terms) separately to each IA and then cumulating forecasts, is more positive, especially towards the end of the period.

Table 2: Forecasts generated by the various models for the period from 2016 to 2023.

Year	Separate industrial areas (IAs)			Cumulated IAs		Total Filings with IAs (2)		Total Filings without IAs (1)		Observed Total
	Electricity (E)	Chemistry (C)	Traditional (T)	Total	SE	Total	SE	Total	SE	
2016	118 190	62 842	99 980	281 012	2 368	282 748	3 852	282 124	3 809	287 129
2017	127 819	62 839	98 883	289 541	3 883	288 848	4 937	288 698	4 941	
2018	141 126	64 707	104 419	310 252	5 217	305 443	5 957	305 276	5 318	
2019	156 779	66 342	109 418	332 539	6 653	320 700	6 989	320 928	7 055	
2020	177 343	68 200	115 410	360 953	8 304	339 159	7 997	339 617	8 093	
2021	203 148	70 239	121 940	395 327	10 263	359 783	9 034	360 514	9 159	
2022	236 029	72 307	128 813	437 149	12 853	381 863	10 149	383 037	10 308	
2023	278 022	74 577	136 471	489 070	16 144	406 589	11 327	408 301	11 523	

Although it is difficult to be sure before the outcomes appear in the future years, the model (2) for TFs including ilr terms seems better than the other models. Its forecasts are more believable than the higher forecasts generated by cumulating the IAs in a way that does not take account of correlations over time.

4 Fitting an ilr regression model to IA proportions

This section moves away from the DLL model for TFs in order to cover a simpler application of CoDa to patent data. An ilr based CoDa straight line regression is carried out on the proportions of IAs in Total Filings. The totals for each IA are analysed with no further breakdown by countries of origin. Analyses were done in Excel.

The quantities that are modelled are the counts (P) from 2001 to 2013, broken down by the three IAs and expressed as proportions of their total. The terms ilr_E , ilr_C and ilr_T are calculated as in

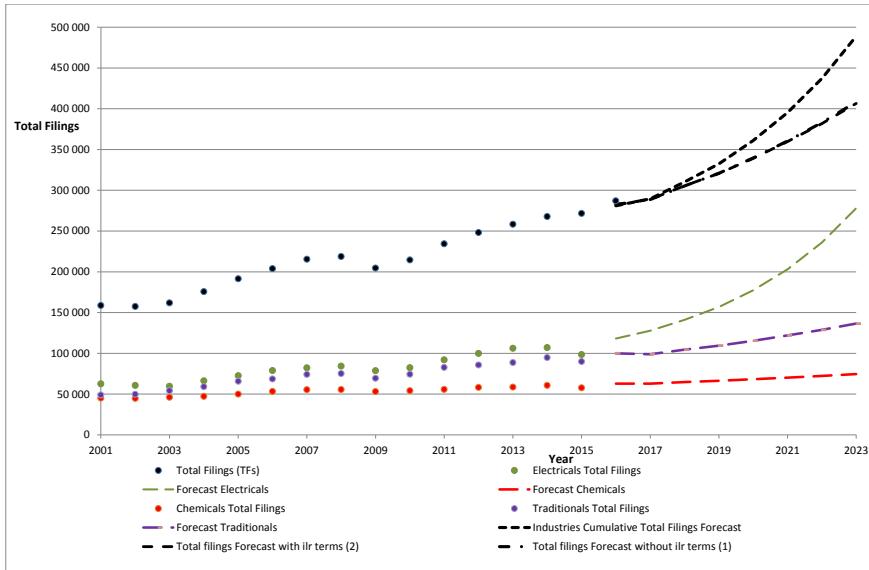


Figure 2: Total Filings forecasts for 2016 to 2023, with and without ilr terms for IAs. For Total filings forecasts, both with and without ilr terms for IAs, training data run from 2001 to 2015. For IAs forecasts, individual and cumulative, training data run from 2001 to 2013.

Section 2. Straight line regression was done independently on ilr_E and ilr_C and projections were made out to year 2022. The projected ilr terms were back-transformed to estimated proportions for Electricals and Chemicals IAs, with the proportion for the Traditionals IA then calculated by subtracting the sum from one.

Direct straight line regressions were also done on the raw proportions. Figure 3 shows the data used and forecasts.

There is little difference in the fitted proportions and in the goodness of fit statistics between direct regression on the proportions and back-transformations of the regressed ilr terms. In the later forecasted years, there is slightly less range between the forecasted proportions by the ilr based method rather than by the direct method.

This result does not rule out the unsophisticated use of straight line regressions on proportions. But CoDa based regressions would be advisable for proportions of IAs in future analyses of patent data, particularly where breakdowns are done into larger numbers of classes (see for example Youn et al., 2015) or where one of the classes accounts for a low or a high proportions of the data.

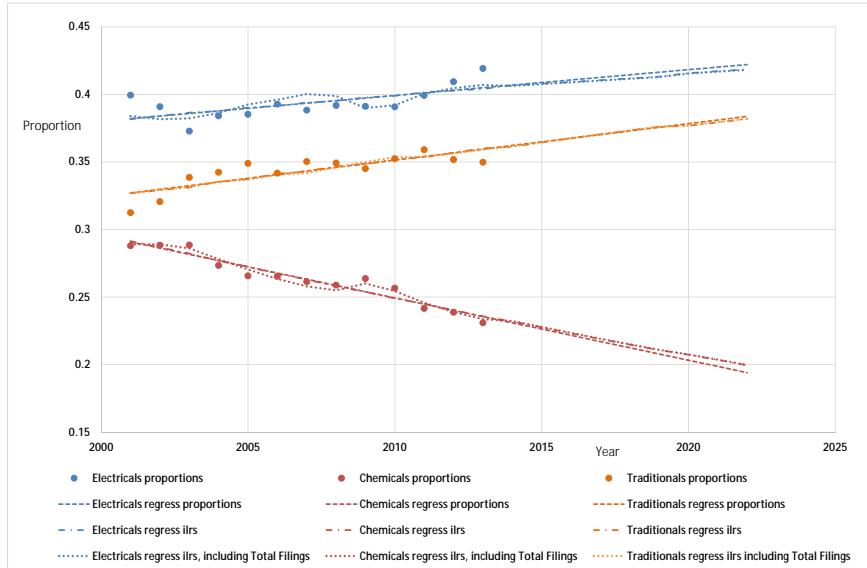


Figure 3: Forecasts of proportions of IAs in Total Filings out to 2022, based on: direct fitting of straight line regressions; straight line regressions after transforming proportions to ilrs; and straight line regressions after transforming proportions to ilrs and including Total Filings as an additional predictor. Training data run from 2000 to 2013.

5 Fitting ilr regression models to both IA proportions and to Total Filings

Two options are explored in this section for combining ilr regressions of transformed proportions with forecasts of a total. Firstly, in order to inspect the possibility of improving the forecasts of proportions, the method of the previous section is augmented to see whether the linear regression of ilr transforms gives a better fit when the total is included as an additional regressor. Secondly, in a simpler analogy of forecasting TFs by the DLL approach, ilr terms are incorporated in a straight line regression of the total to see whether this gives a better fit to the data. These analyses were done in Excel.

5.1 Augmenting the model for IA proportions with the total as predictor

The straight line regression model on ilrs that was fitted in Section 4 was extended to include the total as an additional predictor. A straight line regression on TFs from 2001 to 2013 gave fitted values to input to the model (see the column "Total Filings without IAs" in Table 3 for these inputs).

The fitted values and forecasts are shown in Figure 3. For IA_E and IA_C , incorporation of TFs as a predictor improves the fit of the model, while for IA_T the effect is at best only marginal. Beyond

2013, the forecasted proportions are essentially the same as those for the earlier model where the projection was done only by straight line regression of the ilr terms. There are small variations that are not visible in Figure 3. For example, for 2022 the forecasted proportions for IA_E , IA_C and IA_T are 0.4185, 0.1995 and 0.3819 respectively for straight line regression and 0.4182, 0.1997 and 0.3821 respectively for straight line regression incorporating the TFs as predictors. (It was also found that inputting Totals that had themselves been forecasted by making use of the ilrs, that is from the column "Total Filings with IAs" in Table 3, as described in Section 5.2, gave forecasts of proportions that are identical to the model in Section 4.)

5.2 Augmenting the model for the total with IA proportions as predictors

In Sections 2 and 3 it was shown that the fit of the DLL model can be improved by including ilr terms for proportions of IAs. Here a straight line regression model for TFs is examined to see whether this also applies to a simpler construction.

The straight line regression model was fitted to TFs (untransformed) from 2001 to 2013. Then the model was extended to include ilr_E and ilr_C as regressors. The fitted linear model was as follows.

$$P_t = \alpha_A + \alpha_B t + \alpha_E \cdot ilr_E + \alpha_C \cdot ilr_C + \varepsilon_t$$

Table 3 and Figure 4 show the augmented results compared with the results without the ilr terms.

Table 3: Total Filings generated by straight line regression and by straight line regression with ilr terms for two IAs. Parameter estimates after fitting the two models. Forecasts generated by the two models for the period from 2014 to 2022. * indicates approximate significance at the 95 percent level.

Parameter estimates		Total Filings without IAs	Total Filings with IAs	
Intercept	α_A	-16 334 992*	-4 235 243	
Slope	α_B	8 240*	2 168	
ilr for Electricals	α_E		18 156	
ilr for Chemicals	α_C		-371 482*	
Observation standard error		2 719	1 706	
Residual degrees of freedom		11	9	
Forecasts		Total Filings without IAs	Total Filings with IAs	Observed Total
	Year			
	2014	260 368	260 050	287 129
	2015	268 608	268 290	
	2016	276 848	276 530	
	2017	285 088	284 770	
	2018	293 328	293 010	
	2019	301 568	301 250	
	2020	309 808	309 489	
	2021	318 048	317 729	
	2022	326 288	325 969	

Incorporation of the ilr terms as predictors improves the fit of the model, which happens mainly because of the inadequacy of the straight line as predictor on its own. The estimates for intercept and slope are both significant for the straight line model, while the addition of ilr terms gives a significant estimate only for the ilr_C term. Regarding the forecasts for TFs, these are only slightly different between the models (Table 3 shows a difference of 299 in 2022 at the end of the period). The generated forecasts here for future numbers of TFs are lower than with the DLL model (comparing Figure 2 with Figure 4). However it is not suggested that the straight line is a better model than the DLL model for making forecasts.

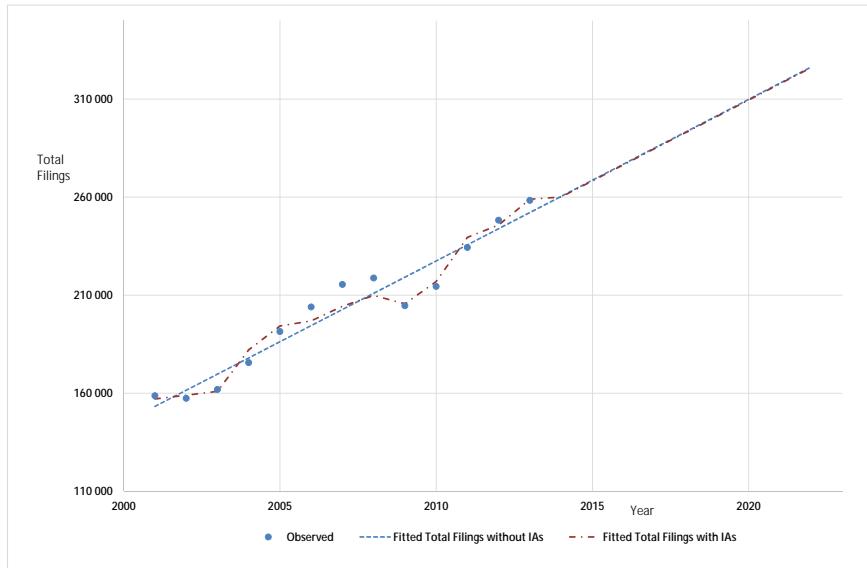


Figure 4: Total Filings forecasts for 2014 to 2022 by straight line regression, with and without ilr terms for IAs. Training data run from 2001 to 2013.

6 Discussion

Regarding the main task of forecasting future numbers of TFs, the results under the DLL model suggest that it is useful to include additional terms based on ilr transformations of sub-counts due to the three IAs. Taking a simpler straight line regression model for forecasting TFs, inclusion of ilr IA terms improves the model fit but leads to almost identical forecasts. For breakdowns of proportions according to IAs, straight line regressions on ilr transforms of counts by IAs give almost the same forecasts as simply fitting straight lines to the raw proportions. For this latter analysis, there is scope for CoDa techniques to be profitably employed with a larger number of classes or if some classes have low frequencies.

A similar remark can be made for analysing trends in proportions of other possible breakdowns of TFs, for example by countries of residence of the applicants or different types of patent filings. One breakdown that is important for forecasting is the distinction between first and subsequent filings. An inventor firstly files for a patent at one office and is then allowed to make subsequent filings within 12 months to any other patent office of interest. Thus a growth of Total Filings can be due either to more first filings that represent the output of inventive enterprise, or to more subsequent filings that represent willingness to gain protection in worldwide markets. The CoDa based techniques introduced here may be relevant for studying patent families that represent the international flows of subsequent filings from their countries of first filing (IP5, 2016).

Experiments were reported above to include complementary terms, meaning the total within a model for the proportions, or proportions within a model of the total. This could be extended

to attempt to model the total and the proportions simultaneously by a combined approach, that could even involve successive iterative applications of the two separate models. However the data here have not given very different forecasts when adding the complementary terms.

This issue could be approached by theoretical analysis, supported by simulated data sets that are constructed to demonstrate useful behaviours. The suggestion by other authors to use a “Total” that relates to the geometric mean rather than the true total could be included in this framework by taking multiples of the geometric mean (EG in the current study this would be “Total” = $3 * (x_{EXCXT})^{1/3}$). In Section 5, the near identities that were found between forecasts with and without the inclusion of the complementary terms might be even closer to true identities when using the proxy total. If so, then the small discrepancies that were found may be related to the inherent approximation involved in estimating the total from the geometric mean.

References

- Coenders, G., Ferrer-Rosell, B., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2015). MANOVA of Compositional Data with a Total. In S. Thió-Henestrosa and J.A. Martín Fernández (Eds.), *Proceedings of the 6th International Workshop on Compositional Data Analysis*.
- EPO (2017). European Patent Office. Patent Filings Survey 2016. <http://www.epo.org/service-support/contact-us/surveys/patent-filings.html>
- Hingley, P. and Nicolas, M. (2004). Methods for forecasting numbers of patent applications at the European Patent Office. *World Patent Information* 26(3), pp. 191-204.
- Hingley, P. and Nicolas, M. (Eds.) (2006). *Forecasting Innovations, Methods for Predicting Numbers of Patent Filings*. Heidelberg: Springer.
- Hingley, P. and Park, W. (2015). A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office. *World Patent Information* 42(3), pp. 19-27.
- Hingley, P. (2016). Forecasting Total Numbers of Filings at the European Patent Office (EPO):- How useful are Breakdowns by Technologies? International Institute of Forecasters Workshop on forecasting New Products and Technologies, Milan, 12 – 13 May 2016.
- Hingley, P. and Park, W. (2016). Forecasting Patent Filings at the European Patent Office (EPO) with a Dynamic Log Linear Regression Model: Applications and Extensions. In S. Soh (Ed.), *Selected Papers from the Asia Conference on Economics & Business Research 2015*. Heidelberg: Springer, pp. 63-83.
- Hingley, P. and Park, W. (2017). Do business cycles affect patenting? Evidence from European Patent Office filings. *Technological Forecasting and Social Change* 116(March), pp. 76-86.
- IP5 (2016). European Patent Office, Japan Patent Office, Korean Intellectual Property Office, State Intellectual Property Office of the People's Republic of China, United States Patent & Trademark Office. IP5 Statistics Report 2015 Edition. <http://www.fiveipoffices.org/statistics/statisticsreports/2015edition.html>
- Pawlowsky-Glahn, V., Egozcue, J. and Lovell, D. (2014). Tools for compositional data with a total. *Statistical Modelling* 15(2), pp. 175-190.
- WIPO (2017). World Intellectual Property Organization. International Patent Classification (IPC). <http://www.wipo.int/classifications/ipc/en/>
- Youn, H., Strumsky, D., Bettencourt, L. and Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *J. R. Soc. Interface* 12. <http://dx.doi.org/10.1098/rsif.2015.0272>

Association rules and compositional data analysis: implications to big data

R.S. Kenett^{1,2,3}, J.A. Martín-Fernández⁴,
S. Thió-Henestrosa⁴, and M. Vives-Mestres⁴

¹KPA Group, Raanana, Israel

²University of Torino, Torino, Italy

³Neaman Institute, Technion, Israel

⁴Department of Computer Science, Applied Mathematics and Statistics,
University of Girona, Girona, Spain

Abstract

Many modern organizations generate a large amount of transaction data, on a daily basis. Transactions typically include semantic descriptors that require specialised methods for analysis. Association rule (AR) mining is a powerful semantic data analytic technique used for extracting information from transaction databases. AR was originally developed for basket analysis where the combination of items in a shopping basket is evaluated to determine prevalence. To generate an AR, the collection of more frequent itemsets—a set of two or more items—must be detected. Then, as a second step, all possible ARs are generated from each itemset. The ARs are then ranked using measures of association labelled, in this context, “measures of interestingness”. The R package “arules” provides more than a dozen such measures including the relative linkage disequilibrium (RLD) which normalises classical Euclidean distances of the itemset from a surface of independence. In this work we study AR and RLD from compositional data (CoDa) perspective. It is well known that CoDa methodology provides nice properties such as subcompositional coherence and scalability. In this work we explore their implications to AR mining in big data analysis. The aim is to analyse if these properties ensure that the AR characteristic is not scale dependent and that if we consider a subset of the original items, we still keep similar behaviour. The work focuses on such aspects, including the dynamic visualization of CoDa-AR measures on a simplex representation of the itemsets and its multidimensional extension.

Key words: Aitchison geometry, text analysis, association rules, isometric logratio coordinates, itemsets, measures of interestingness.

1 Introduction

This paper is about the application of compositional data analysis methods to text or unstructured semantic data. Using association rule (AR) mining one can detect and extract useful information from unstructured semantic data (Agrawal et al., 1993) commonly organized in large databases. A database is formed by *attributes* and *transactions*. The attributes are binary variables $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_n\}$ called *items*; and the transactions are the row vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$. For example, in web clickstream analysis the interest are the web pages visited (items) in a web session (transaction). In market basket analysis a transaction is a single visit of a customer to the supermarket as and the attributes are the list of products or items bought.

Given $\mathbf{A}, \mathbf{B} \subseteq \mathbf{I}$ two *itemsets* (sets of items) with $\mathbf{A} \cap \mathbf{B} = \emptyset$, a *rule* is an implication of the form $\{\mathbf{A} \Rightarrow \mathbf{B}\}$. Here, the itemsets \mathbf{A} and \mathbf{B} are respectively called the *antecedent* and *consequent* itemsets. One said that has an AR between \mathbf{A} and \mathbf{B} when its rule is “important”. The AR $\{\text{onions, potatoes} \Rightarrow \{\text{burger}\}$ is a typical example of AR in market basket analysis. Despite the analysis commonly deals with original binary variables, continuous rules can be also defined: $\{\text{age} > 25\} \Rightarrow \{\text{total purchase} > € 50\}$. A major challenge in big data applications is to discover the “important” AR. Using the AR expression as a contingency table, Kenett (1983) proposed its representation on the simplex. In addition he defines a measure association called relative linkage disequilibrium (RLD) to calculate the distance of an AR to the surface of independence. Applications of this approach to the analysis of accident data and telecommunication failures are presented in Kenett and Salini (2008), whereas Kenett and Salini (2010) present applications to web searches using Google insight. RLD was implemented in the arules R package (<https://cran.r-project.org/web/packages/arules/index.html>). In this work, we extend this approach using compositional data (CoDa) analysis methods (<http://www.compositionaldata.com/>) providing examples of how CoDa can be used in test analysis.

1.1 Measuring the strength of an AR: measures of interestingness

Let $\{\mathbf{A} \Rightarrow \mathbf{B}\}$ be the AR of interest. Let x_1 be the support (relative frequency of occurrence) of both \mathbf{A} and \mathbf{B} ; x_2 the support of only \mathbf{A} ; x_3 the support of only \mathbf{B} ; and x_4 the relative frequency of transactions where neither \mathbf{A} or \mathbf{B} occur. In other words, let n_k be the number of transactions which satisfy the conditions in x_k , $k=1,\dots,4$, the total number of transactions is $\sum n_k = m$, and $x_k = n_k/m$. Table 1 shows that x_k respectively estimates $P(\mathbf{A} \cap \mathbf{B})$, $P(\mathbf{A} \cap \mathbf{B}^c)$, $P(\mathbf{A}^c \cap \mathbf{B})$, $P(\mathbf{A}^c \cap \mathbf{B}^c)$. Consequently, $\sum x_k = 1$ and $\mathbf{x} = (x_1, x_2, x_3, x_4)$ can be considered as a *composition*.

Table 1: AR contingency table for the AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$

		\mathbf{B}	\mathbf{B}^c
\mathbf{A}	x_1	x_2	
\mathbf{B}^c	x_3	x_4	

We present below four measures of interestingness, implemented in the R *arules* package (Hahsler et al., 2014):

- support $\{\mathbf{A} \Rightarrow \mathbf{B}\} = x_1 = N\{\mathbf{A} \Rightarrow \mathbf{B}\}/m$, where $N\{\mathbf{A} \Rightarrow \mathbf{B}\}$ is the number of transactions verifying the rule, informs of the proportion of transactions that verify the AR.
- confidence $\{\mathbf{A} \Rightarrow \mathbf{B}\} = x_1/(x_1+x_2) = N\{\mathbf{A} \Rightarrow \mathbf{B}\}/N\{\mathbf{A}\}$, where $N\{\mathbf{A}\}$ is the number of transactions containing the antecedent. Note that because confidence can be interpreted as a conditional probability.
- lift $\{\mathbf{A} \Rightarrow \mathbf{B}\} = x_1/[(x_1+x_2) \cdot (x_1+x_3)] = \text{confidence}\{\mathbf{A} \Rightarrow \mathbf{B}\}/\text{support}\{\mathbf{B}\}$. One can interpret this measure as a deviation under independence of the itemsets (Kenett and Salini, 2011). When lift are respectively smaller and greater than 1, the knowledge that \mathbf{A} holds causes a negative and positive effect on the probability of \mathbf{B} . For lift = 1, no effect is indicated, that is, there is no association between the itemsets.
- RLD $\{\mathbf{A} \Rightarrow \mathbf{B}\}$, introduced in Kenett and Salini (2008), captures the level of dependence of the AR because measures the relative Euclidean distance of the AR from its linear projection on a surface with lift=1 (next section provides more details on RLD).

1.2 Compositional data (CoDa)

Nowadays is a general agreement among researchers (Pawlowsky-Glahn and Buccianti, 2011) that the simplex has its own geometry, different from the classical Euclidean. The three basic operations of this particular geometry are: perturbation, powering, and inner product. These basic elements provide an Euclidean structure to the simplex which allows to analyse CoDa applying all the multivariate methods on the orthonormal log-ratio coordinates. In consequence, an important previous step to use these statistical techniques is to build orthonormal bases to express any composition \mathbf{x} in its corresponding coordinates, called *isometric log-ratio* coordinates: $ilr(\mathbf{x})$. The Sequential Binary Partition (SBP) (Pawlowsky-Glahn and Buccianti, 2011, Chapter 2) is an easy and interpretable way to build a basis. A SBP of the parts of a composition consists of $D-1$ steps, where an orthonormal coordinate is built in each step of the partition. In a first step, a SBP consists of splitting parts of the composition \mathbf{x} into two groups, which are indicated by +1 and -1. In consecutive steps, each previously created group of parts is split again into two groups. The partition ends when the groups are made up of a unique part. In the j^{th} step of a SBP, denoting by \mathbf{x}^+ the group of r parts marked with a +1 and by \mathbf{x}^- the group of s parts marked with a -1, the corresponding coordinate, $ilr_j(\mathbf{x})$, is

$$ilr_j(\mathbf{x}) = \sqrt{\frac{r+s}{r+s}} \ln \left(\frac{g(\mathbf{x}^+)}{g(\mathbf{x}^-)} \right),$$

where $g(\cdot)$ is the geometrical mean of involved parts of \mathbf{x} . One important criterion to select the basis is that the ilr-coordinates are useful for interpretation purposes. For example, if we identify the Table 1 by the composition \mathbf{T} , this table can be expressed in terms of the ilr-coordinates

$$ilr(\mathbf{T}) = \left(\frac{1}{2} \ln \left(\frac{x_1 x_4}{x_2 x_3} \right), \frac{\sqrt{2}}{2} \ln \left(\frac{x_1}{x_4} \right), \frac{\sqrt{2}}{2} \ln \left(\frac{x_2}{x_3} \right) \right). \quad (1)$$

2 CoDa-measures for independence in a table

The multiplicative column and row marginal vectors (Egozcue et al., 2015) of any table \mathbf{T} (Table 1) are respectively

$$Gr = C(\sqrt{x_1 x_2}, \sqrt{x_3 x_4}) \text{ and } Gc = C(\sqrt{x_1 x_3}, \sqrt{x_2 x_4})$$

where C means the closure operation $C(\mathbf{x}) = (\frac{x_1}{\sum x_k}, \dots, \frac{x_4}{\sum x_k})$. These vectors allow to construct the independent probability table \mathbf{T}_{ind} (Table 2) associated to table \mathbf{T} .

Table 2: Independent probability table \mathbf{T}_{ind} of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$ (without closure for simplicity)

	B	${}^c B$
A	$x_1 \sqrt{x_2 x_3}$	$x_2 \sqrt{x_1 x_4}$
${}^c A$	$x_3 \sqrt{x_1 x_4}$	$x_4 \sqrt{x_2 x_3}$

Note that a \mathbf{T}_{ind} table verifies that is “independent” (Egozcue et al., 2015) because $\mathbf{T}_{\text{ind}} = (\mathbf{T}_{\text{ind}})_{\text{ind}}$. The interaction probability table \mathbf{T}_{int} (Table 3) associated to table \mathbf{T} is obtained when one applies the perturbation operation to subtract table \mathbf{T}_{ind} from \mathbf{T} .

Table 3: Interaction probability table \mathbf{T}_{int} of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$ (without closure for simplicity)

	B	${}^c B$
A	$1/\sqrt{x_2 x_3}$	$1/\sqrt{x_1 x_4}$
${}^c A$	$1/\sqrt{x_1 x_4}$	$1/\sqrt{x_2 x_3}$

Table 4 shows the ilr-coordinates of tables \mathbf{T} , \mathbf{T}_{ind} and \mathbf{T}_{int} . Note that it holds $ilr(\mathbf{T}) = ilr(\mathbf{T}_{\text{ind}}) + ilr(\mathbf{T}_{\text{int}})$.

Table 4: ilr-coordinates of tables \mathbf{T} , \mathbf{T}_{ind} and \mathbf{T}_{int} of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$ using basis defined in Eq. (1).

ilr-coordinates	ilr ₁	ilr ₂	ilr ₃
\mathbf{T}	$\frac{1}{2} \ln \left(\frac{x_1 x_4}{x_2 x_3} \right)$	$\frac{\sqrt{2}}{2} \ln \left(\frac{x_1}{x_4} \right)$	$\frac{\sqrt{2}}{2} \ln \left(\frac{x_2}{x_3} \right)$
\mathbf{T}_{ind}	0	$\frac{\sqrt{2}}{2} \ln \left(\frac{x_1}{x_4} \right)$	$\frac{\sqrt{2}}{2} \ln \left(\frac{x_2}{x_3} \right)$
\mathbf{T}_{int}	$\frac{1}{2} \ln \left(\frac{x_1 x_4}{x_2 x_3} \right)$	0	0

Let $\|\mathbf{T}\|_a = \|ilr(\mathbf{x})\|$ be the Aitchison norm of a table \mathbf{T} , it holds that $\|\mathbf{T}\|_a^2 = \|\mathbf{T}_{\text{ind}}\|_a^2 + \|\mathbf{T}_{\text{int}}\|_a^2$, that is, one has a decomposition of the Aitchison norm of table \mathbf{T} .

Egozcue et al. (2015) present the Simplicial Deviance (SD) as a natural measure of independence in a table which for a table \mathbf{T} (Table 1) is

$$SD(\mathbf{T}) = \|\mathbf{T}_{\text{int}}\|_a^2 = \frac{1}{4} \log^2 \left(\frac{x_1 x_4}{x_2 x_3} \right) = ilr_1^2(\mathbf{T}) \quad (2)$$

We can interpret that the strength of the AR depends on how large is the value of coordinate ilr_1 . In other words, the closer ilr_1 gets to zero value, the more independence between itemsets \mathbf{A} and \mathbf{B} . However, the decomposition of $\|\mathbf{T}\|_a^2$ suggests that a same SD value may be obtained with different size of the norm of \mathbf{T} . Due to that fact, Egozcue et al. (2015) introduce the Relative Simplicial Deviance that normalizes SD

$$RSD(\mathbf{T}) = \frac{SD}{\|\mathbf{T}\|_a^2} = \frac{ilr_1^2(\mathbf{T})}{\|ilr(\mathbf{T})\|^2} \quad (3)$$

RSD takes values in an interval $[0,1]$, with $RSD = 0$ for the independence and $RSD = 1$ for the maximum association.

In addition, to contrast the independence evaluating the significance of both SD and RSD measures, Egozcue et al. (2015) introduce a bootstrap algorithm consisting of following steps:

- i) Calculate \mathbf{T}_{ind} , \mathbf{T}_{int} , SD and RSD .
- ii) Simulate 10000 multinomial samples ($\mathbf{T}^{(k)}$) assuming the independence hypothesis $H_0: \mathbf{T} = \mathbf{T}_{\text{ind}}$ is true. For each table $\mathbf{T}^{(k)}$, calculate $\mathbf{T}_{\text{ind}}^{(k)}$, $\mathbf{T}_{\text{int}}^{(k)}$, $SD^{(k)}$ and $RSD^{(k)}$.
- iii) Compare respectively the value of SD and RSD with the distribution of the 10000 values of $SD^{(k)}$ and $RSD^{(k)}$ to obtain the *percentile* p-value (*left tail*). Calculate the 0.05 significance critical points (5th quantile) in the left tail of each distribution.

3 Example

Consider a questionnaire ($m=5000$) where young people respond if they like Basketball and eat cereal for breakfast. Table 5 shows the three AR tables.

Table 5: Tables \mathbf{T} , \mathbf{T}_{ind} and \mathbf{T}_{int} : (a) \mathbf{T} in counts (corresponding proportions); (b) \mathbf{T}_{ind} ; (c) \mathbf{T}_{int}

\mathbf{T}	Cereal	Not cereal
Basketball	2000 (0.4)	1750 (0.35)
Not basketball	1000 (0.2)	250 (0.05)
(a)		
\mathbf{T}_{ind}	Cereal	Not cereal
Basketball	0.54	0.25
Not basketball	0.14	0.07
(b)		
\mathbf{T}_{int}	Cereal	Not cereal
Basketball	0.17	0.33
Not basketball	0.33	0.17
(c)		

The geometry of the simplex allows plotting (Figure 1) the decomposition of the vector \mathbf{x} corresponding to the table \mathbf{T} (blue vector) in the sample space S^4 (Fig 1a) and in the ilr-coordinates space (Fig. 1b). It holds that $\mathbf{T}=(0.4, 0.35, 0.2, 0.005) = \mathbf{T}_{\text{ind}} \oplus \mathbf{T}_{\text{int}} = (0.54, 0.25, 0.17, 0.07) \oplus (0.17, 0.33, 0.33, 0.17)$, where “ \oplus ” is the perturbation operation. The vector of ilr-coordinates $\text{ilr}(\mathbf{T}) = (-0.63, 1.47, 0.40)$ decomposes into the vector $\text{ilr}(\mathbf{T}_{\text{int}})$ (green colour) and its orthogonal projection to the plane $\langle \text{ilr}_2, \text{ilr}_3 \rangle$, the $\text{ilr}(\mathbf{T}_{\text{ind}})$ (red colour). The simplicial deviance is equal to $SD = 0.39$ that normalizes to $RSD = 0.14$. When the testing procedure to contrast the independence was applied we obtained both p-values below $0.5 \cdot 10^{-4}$ indicating a significant interaction.

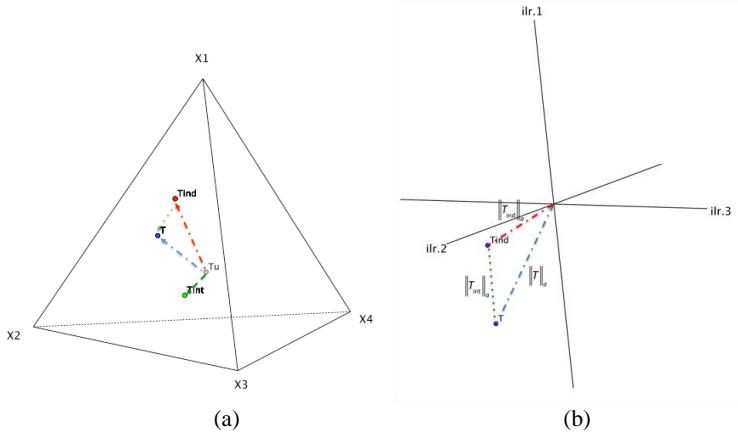


Figure 1: Table decomposition: (a) in the simplex $\mathbf{T}=\mathbf{T}_{\text{ind}} \oplus \mathbf{T}_{\text{int}}$; (b) in the ilr-coordinates space. The table \mathbf{T} (blue color) is orthogonally projected to table \mathbf{T}_{ind} (red colour) in the plane $\langle \text{ilr}_2, \text{ilr}_3 \rangle$. The dotted green line represents the norm of the table \mathbf{T}_{int} .

4 CoDa-measure for AR

4.1 Relative linkage disequilibrium (RLD) versus $\text{ilr}_1(\mathbf{T})$

The information of the measure lift for a table \mathbf{T} (Table 1) of an AR

$$\text{lift(AR)} = \frac{x_1}{(x_1+x_2)(x_1+x_3)} \quad (4)$$

comes from its comparison with the value 1. That is, it measures how similar is the value x_1 to the product of corresponding additive column and row marginal vectors $(x_1+x_2)(x_1+x_3)$. On the other hand, the RLD measure uses the approach to measure the similarity between the value x_1 and the product $(x_1+x_2)(x_1+x_3)$ via the subtraction $x_1 - (x_1+x_2)(x_1+x_3) = 0$ which is equivalent to the difference or disequilibrium (Kenett, 1983)

$$D(\text{AR}) = x_1x_4 - x_2x_3 = 0 \quad (5)$$

Importantly D takes values in $[-1,1]$ and it holds that

$$\text{lift(AR)} = 1 + \frac{D(\text{AR})}{(x_1+x_2)(x_1+x_3)}.$$

A value $D < 0$ indicates a negative effect; $D = 0$ the independence; and the positive effect corresponds to $D > 0$. The definition of D produces some difficulties because Kenett and Salini (2011, page 153) point out: “However, points closer to the edges of the simplex will have intrinsically smaller values of D ”. To solve this difficulty, they introduce the measure $RLD = D/D_M$, where D_M is the Euclidean distance between the table \mathbf{T} and the surface $D = 0$. Because Euclidean distance is not coherent with the simplicial

geometry (Palarea-Albaladejo et al., 2012), instead one should use the Aitchison distance between two compositions \mathbf{x} and \mathbf{y} : $d_a(\mathbf{x}, \mathbf{y}) = \|ilr(\mathbf{x}) - ilr(\mathbf{y})\|$.

From the definition of *RLD* one can easily deduce that the tables \mathbf{T} where one or more values in the vector \mathbf{x} are equal to zero are of no interest for the analysis. Indeed, if only one value in the vector \mathbf{x} is equal to zero then $RLD = 1$, that is, the point \mathbf{x} takes the maximal distance to the surface $D = 0$. One has the same situation when the pair $\{x_1, x_4\}$ or the pair $\{x_2, x_3\}$ is equal to zero. On the other hand, when the

other possible pairs are zero or three values are zero, then $D = 0$, that is, one has independence. However, for the case of three values equal to zero the index is misleading because when three values in a table \mathbf{T} are zero it means that the itemsets \mathbf{A} and \mathbf{B} are absolutely associated. In any case, hereafter one can assume that all values in the table \mathbf{T} are non-zero.

The Eq. (5) can be analyzed from another approach because the information provided by the table \mathbf{T} is of relative nature. The comparison $x_1x_4 - x_2x_3 = 0$, or $x_1x_4 = x_2x_3$ can be formulated as

$$\frac{x_1x_4}{x_2x_3} = 1 \Leftrightarrow \log\left(\frac{x_1x_4}{x_2x_3}\right) = 0 \Leftrightarrow ilr_1(\mathbf{T}) = 0 \Leftrightarrow SD = 0,$$

where $ilr_1(\mathbf{T})$ is the first ilr -coordinate. Therefore, using the relationship between ilr_1 and D it holds that

- $ilr_1(\mathbf{T}) < 0$: negative effect between itemsets (\mathbf{A} true, \mathbf{B} less likely true)
- $ilr_1(\mathbf{T}) = 0$: independence
- $ilr_1(\mathbf{T}) > 0$: positive effect (\mathbf{A} true, \mathbf{B} more likely true)

4.2 Odds ratio (OR) versus $ilr_1(\mathbf{T})$

Among the number of different measure of interestingness, Tan et al. (2004) describe the odds ratio (OR). Given a table \mathbf{T} (Table 1) one can calculate

$$odds(\mathbf{B}/\mathbf{A}) = (P(\mathbf{B}/\mathbf{A})/P(\mathbf{B}'/\mathbf{A})) = P(\mathbf{A} \cap \mathbf{B})/P(\mathbf{A} \cap \mathbf{B}') = x_1/x_2$$

and

$$odds(\mathbf{B}'/\mathbf{A}') = (P(\mathbf{B}'/\mathbf{A}')/P(\mathbf{B}/\mathbf{A})) = P(\mathbf{A}' \cap \mathbf{B}')/P(\mathbf{A}' \cap \mathbf{B}) = x_3/x_4,$$

where the odds-ratio (*OR*) is

$$OR(AR) = odds(\mathbf{B}/\mathbf{A})/odds(\mathbf{B}'/\mathbf{A}') = (x_1x_4)/(x_2x_3). \quad (6)$$

The value $OR(AR) = 1$ indicates independence, $OR(AR) > 1$ positive effect and $OR(AR) < 1$, negative effect. Note that it holds $ilr_1(\mathbf{T}) = 1/2 \cdot \log(OR(AR))$ and $OR(AR) = e^{(2ilr_1(\mathbf{T}))}$. This monotonic functional relation indicates that both values have the same ranking. Moreover, when a measure is unbounded some practical normalizations should be advisable (Tan et al., 2004): “A measure is normalized if its value ranges between -1 and $+1$. An unnormalized measure M that ranges between 0 and $+\infty$ can be normalized via transformation functions such as $(M-1)/(M+1)$ or $(\tan^{-1}\log(M))/(\pi/2)$ ”. For example, the measure *Yule's Q*

$$OR^*(AR) = Yule's\ Q(AR) = \frac{x_1x_4 - x_2x_3}{x_1x_4 + x_2x_3} \quad (7)$$

is a normalized version of the *OR* (Tan et al., 2004). Following Tan et al. (2004), a measure M should satisfy three key properties:

- P1: $M = 0$ if \mathbf{A} and \mathbf{B} are statistically independent;
- P2: M monotonically increases with $P(\mathbf{A} \cap \mathbf{B})$ when $P(\mathbf{A})$ and $P(\mathbf{B})$ remain the same;
- P3: M monotonically decreases with $P(\mathbf{A})$ (or $P(\mathbf{B})$) when the rest of the parameters ($P(\mathbf{A} \cap \mathbf{B})$ and $P(\mathbf{B})$ or $P(\mathbf{A})$) remain unchanged.

The *Yule's Q* measure (Eq. 7) verifies the three properties (Tan et al., 2004). By its definition $ilr_1(\mathbf{T})$ verifies the property P1. Because the unnormalized version of measure *OR(AR)* verifies properties P2 and P3 (Tan et al., 2004), then $ilr_1(\mathbf{T})$ also verifies these two properties. On the other hand, by its definition

the measure $SD(AR)$ does not verify these two properties.

4.3 CoDa-AR measure

We can aggregate the benefits of interpretation and of the three properties defining the unnormalized compositional measure of association

$$C(AR) = ilr_1(\mathbf{T}) \quad (8)$$

Because the measure C takes values in $(-\infty, \infty)$ it turns out more difficult to interpret the strength of the association. On the other hand, the value $C(AR) = 0$, or non significant different from zero, indicates that **A** and **B** are statistically independent (property P1 or $\mathbf{T} = \mathbf{T}_{ind}$). Following Prados et al. (2010), among the number of possibilities to normalize a measure that ranges between $-\infty$ and $+\infty$ as $C(AR)$ (Eq. 8), one can select the hyperbolic tangent function $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$ also used in It holds

$$C^*(AR) = \tanh(C(AR)) = OR^*(AR) = Yule's\ Q(AR)$$

that accordingly exhibits all the properties of the *Yule's Q* measure. The example in Fig. 1 takes $C(AR) = -0.63$ and $C^*(AR) = -0.56$, which corresponds to a negative effect, that is, given that a young likes basketball, it is less likely that he/she eats cereal for breakfast. The positive sign of $ilr_2(\mathbf{x}) = 1.47$ indicates that is more likely that a young like both products than none. Moreover, because $ilr_3(\mathbf{x}) = 0.40$ is positive we can assume that people that only like one of them, they prefer basketball.

4.4 CoDa-AR measure applied to a large data base

As an example of application we use a data base obtained through the website <https://treato.com/Nicardipine/?a=s>. On this web site users reported side effects and concerns after use certain medication. Data base consists of 6075 side effects from 882 different users.

With this data base we performed an AR analysis and we selected rules with a minimum support and confidence of 0.1 and 0.4 respectively. This process gives us a total of 50 rules, each one with their contingency table.

Using CoDaPack we calculated the ILR coordinates using basis defined in Eq. (1). The first ILR coordinate has a mean of 0.54 and a standard deviation of 0.19, being all observations greater than 0, that is, on every rule we have a positive effect because the product $x_1 \cdot x_4$ is greater than $x_2 \cdot x_3$. Also Figure 2 shows the histogram of this first ILR coordinate with a bias on the right.

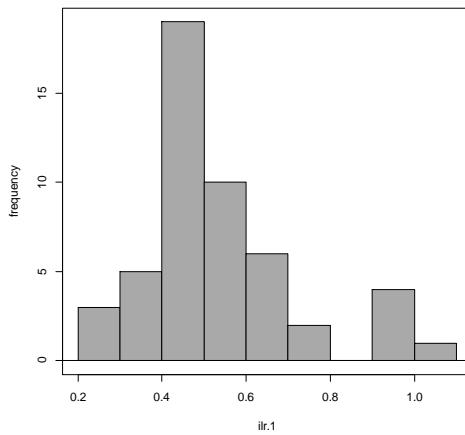


Figure 2: Histogram of ILR1.

Figure 3 shows the CLR-Biplot of these contingency tables, where rules are colored according to their consequent. The first principal component of this biplot has a cumulative proportion explained of 87.01% and shows the opposition between $CLR(x_3)$ and $CLR(x_2)$, that is, rules with high $CLR(x_3)$ has low $CLR(x_2)$ and vice versa.

Also, this biplot show the rules well separated according their consequent. This separation follows the direction of $CLR(x_3)$. For example we can see that rules with a consequent consisting of a side effect of “Hydrocephalus” have the higher value on $CLR(x_3)$.

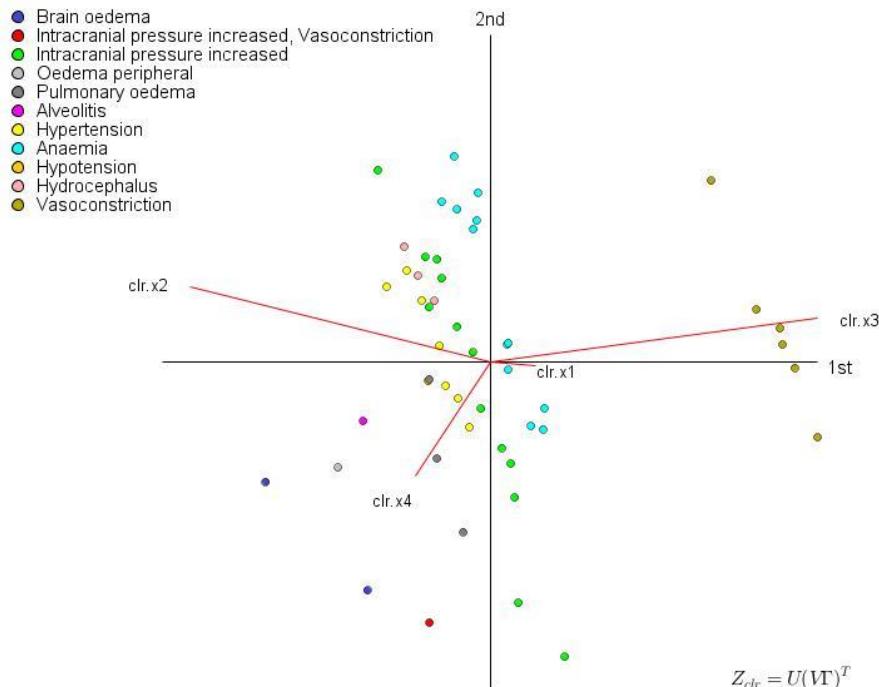


Figure 3: CLR Biplot ($n=50$) where parts are the four cells of the contingency table. Observations are colored according the consequent of the rules.

5 Final remarks

Compositional measures of independence SD and RSD are coherent with the simplicial geometry of the simplex, the sample space of contingency tables of AR. Moreover, a contrast is available to confirm the significance of an association between two variables. CoDa-AR measures of interestingness $C(AR)$ and $C^*(AR)$ exhibit these properties as well as the properties received from the OR and $Yule's\ Q$ indices. In addition, the relation between these CoDa-AR measures and other common measures facilitates the interpretation of negative and positive effects between itemsets. The CoDa geometry provides visualization techniques used here to plot the performance of these measures when all the significant AR of a large database are analysed. The principles of coherence and scalability that are fundamental to CoDa are relevant to big data text analysis. This paper demonstrates how this can be implemented.

Acknowledgements

This work has been partially financed by the project by the projects: “CODA-RETOS” (MCI; Ref: MTM2015-65016-C2-1-R) and Programa Salvador de Madariaga-Fulbright (MECD, Ref.: PRX16/00258).

References

- Agrawal, R., Imielinski, T., Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: *Proceedings of the Conference on Management of Data*, pp. 207–216, ACM Press, New York.
- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London, UK.
- Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M., Hron, K. (2015). Independence in Contingency Tables Using Simplicial Geometry. *Commun Stat-Theor M*, 44, 3978—3996.
- Hahsler, M., Buchta, C., Gruen, B. and Hornik, K. (2008). arules R Package, Version 0.6-6, Mining Association Rules and Frequent Itemsets, <https://cran.r-project.org/web/packages/arules/index.html>.
- Kenett, R.S. (1983). On an Exploratory Analysis of Contingency Tables. *J R Stat Soc Series D*, 32, 395—403.
- Kenett, R.S., Salini, S. (2008). Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks. *T Mach Learn and Data Min*, 1(2), 83—96.
- Kenett, R.S., Salini, S. (2010). Relative linkage disequilibrium in tracking web search patterns. *CLADAG*, Florence, Italy.
- Kenett, R.S., Salini, S. (2011). Measures of Association Applied to Operational Risks (Chapter 9). In: Kenett, R.S., Raanan, Y. (eds.) *Operational Risk Management*, pp. 149-167. John Wiley & Sons.
- Martín-Fernández, J.A., Vives-Mestres, M. and Kenett, R.S. (2016). Understanding association rules from a compositional data approach. *SIS 2016, 48th Meeting of the Italian Statistical Society*, Università di Salerno, Italy, June 8-10th.
- Palarea-Albaladejo, J., Martín-Fernández, J.A., Soto, J. (2012). Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *J Classif*, 29, 144-169.
- Pawlowsky-Glahn, V., Buccianti, A. (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd., Chichester, UK.
- Prados ,F., Boada, I., Prats, A., Martín-Fernández, J.A., Feixas, M., Blasco, G., Puig, J., Pedraza, S., (2010). Analysis of New Diffusion Tensor Imaging Anisotropy Measures in the 3P-plot, *Journal of Magnetic Resonance Imaging*, 31(6), pp. 1435-1444
- Tan, P.-N., Kumar, V., Srivastava, J. (2004). Selecting the Right Objective Measure for Association Analysis. *Inform Syst*, 29(4), 293—313.

Dominic LaRoche

this is page 1

Title: Quality control metrics for extraction-free targeted RNA-Seq: methods afforded by a compositional framework.

D.D. LaRoche^{1,2,*}, D.D. Billheimer¹, K.. Michels², and B.L. LaFleur²

¹University of Arizona, Tucson, Arizona, USA;

²HTG Molecular Diagnostics, Tucson, Arizona, USA

* *dlaroche@email.arizona.edu*

Abstract

We develop quality control diagnostics for targeted RNA-Seq using the theory of compositional data. Targeted sequencing using extraction-free sample preparation allows researchers to efficiently measure transcripts of interest for a particular disease by focusing sequencing efforts on a select subset of transcript targets from small sample volumes. However, extraction free technologies create the need for post-sequencing quality control metrics since poor quality samples, which would likely be removed after unsuccessful RNA extraction in extraction-based technologies, can still be sequenced. We capitalize on the relative frequency property of RNA-Seq data to identify poor quality samples, samples that violate the relative frequency expectation, and batch effects using only post-sequencing data.

Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a given sample within a sequencing run. We propose a method, based on outlier detection of Centered Log-Ratio (CLR) transformed counts, for objectively identifying problematic samples based on the total number of reads allocated to the sample. Similarly, most RNA-Seq analyses assume that the relative frequencies of target read counts are not affected by the total number of reads allocated to the sample, a property known as Compositional Invariance. We develop a method for evaluating sequencing runs or experiments for violations of compositional invariance. Finally, batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems. We show that CLR transformed RNA-Seq data is appropriate for evaluation in a PCA biplot and improves batch effect detection over current methods.

Key words: RNA-Seq, Transcriptome, Quality Control, Sequencing.

Dominic LaRoche

this is page 2

1 Introduction

We develop quality control diagnostics for targeted RNA-Seq using the theory of compositional data. Targeted sequencing allows researchers to efficiently measure transcripts of interest for a particular disease by focusing sequencing efforts on a select subset of transcript targets. Targeted sequencing offers several benefits over traditional whole-transcriptome RNA-Seq for clinical use including the elimination of amplification bias, reduced sequencing cost, and a simplified bioinformatics workflow. Moreover, extraction-free targeted sequencing technologies, such as HTG EdgeSeq, permit the use of very small sample volumes. However, extraction free technologies create the need for post-sequencing quality control metrics since poor quality samples, which would likely be removed after unsuccessful RNA extraction in extraction-based technologies, can still be sequenced. The post-sequencing methods described here should be easily extensible to traditional extraction-based RNA-Seq because targeted and traditional RNA-Seq data share many of the same properties.

Relative frequency measures are characterized as a vector of proportions of some whole. These proportions are necessarily positive and sum to a constant which is determined by the measurement system and not the measurand. Targeted and whole transcriptome RNA-Seq measurements from NGS-based instruments provide only relative frequencies of the measured transcripts. The measurement technology, along with sample preparation, preclude the measurement of absolute abundance. The total number of reads in a sequencing run for high-throughput RNA-Seq instruments is determined by the maximum number of available reads and not the absolute number of reads in a sample. For example, the Illumina Mi-Seq is limited to 25 million reads in a sequencing run while the Roche 454 GS Junior (TM), with longer read lengths, claims approximately 100,000 reads per run for shotgun sequencing. These reads are distributed across all of the samples included in a sequencing run and, therefore, impose a total sum constraint on the data. This constraint cascades down to each probe or tag within a sample which is, in turn, constrained by the total number of reads allocated to the sample thereby creating a natural hierarchical structure to RNA-Seq data.

Previous authors have identified the relative abundance nature of RNA-Seq data ([Robinson and Smyth 2007](#); [Anders and Huber 2010](#); [Robinson and Oshlack 2010](#); [Law et al. 2014](#); [Lovell et al. 2015](#)). For example, Robinson and Smyth ([2007](#)) consider counts of RNA tags as relative abundances in their development of a model for estimating differential gene expression implemented in the Bioconductor package edgeR. Similarly, Robinson and Oshlack ([2010](#)) explicitly acknowledge the mapped-read constraint when developing their widely used Trimmed-Mean of M-values (TMM) normalization method for RNA-Seq data. Finally, the commonly used log₂ Counts per Million (CPM) re-scaling transformation proposed by Law et al. ([2014](#)) divides each sequence count by the total number of reads allocated to the sample thereby transforming the data for each sample into a vector of proportions.

The positivity and summation constraint complicate the analysis of relative frequency data. As early as 1896 Karl Pearson ([Pearson 1896](#)) identified the spurious correlation problem associated with compositions. John Aitchison observed that relative frequency data is compositional and developed a methodology based on the geometric constraints of compositions ([Aitchison 1986](#)). Recent authors have argued that ignoring the sum constraint can lead to unexpected results and erroneous inference ([Lovell et al. 2011](#)). Despite the evidence that RNA-Seq data are compositional in nature, few researchers have extended the broad set of compositional data analysis theory and operations for use in RNA-Seq analysis problems.

We provide a brief background on compositional methods. We then extend existing compositional data methodology to develop two quality control metrics and improve batch effect detection for RNA-Seq data.

Dominic LaRoche

this is page 3

2 Methods

2.1 Compositional Data

Compositional data is defined as any data in which all elements are non-negative and sum to a fixed constant ([Aitchison 1986](#)). For RNA-seq data, the total sum constraint is imposed by the limited number of available reads in each sequencing run. Since this total differs between sequencing platforms we will refer to the total number of available reads as \mathbb{T} . These reads are distributed among the D samples in a sequencing run such that:

$$\sum_{i=1}^D t_i = \mathbb{T} \quad (1)$$

where t_i represents the total reads for sample i . Because of the total sum constraint, the vector \mathbf{t} is completely determined by $D - 1$ elements since the D^{th} element of \mathbf{t} can be determined from the other $d = D - 1$ elements and the total \mathbb{T} :

$$t_D = \mathbb{T} - \sum_{i=1}^d t_i \quad (2)$$

In [2](#), any of the elements can be chosen for t_D with the remaining elements labeled $1, \dots, d$ in any order ([Aitchison 1986](#)). Similarly, the total reads for each sample (t_i) are distributed among the P transcript targets in the assay such that $\sum_{j=1}^P p_{ij} = t_i$, where p_{ij} is the number of reads allocated to target j in sample i . We highlight the hierarchical structure of RNA-Seq data as it leads to useful properties when developing quality control metrics.

From equations [1](#) and [2](#) it is clear that the total reads allocated to each of the D samples represent a $D - 1 = d$ dimensional simplex (\mathcal{S}^d). This leads to problems when using methods developed for standard Euclidean sample spaces such as interpreting the traditional $D \times D$ covariance structure or measuring the distance between vectors. In particular, it is clear that for a D -part composition \mathbf{x} , $\text{cov}(x_1, x_1 + \dots + x_D) = 0$ since $x_1 + \dots + x_D$ is a constant. Moreover, the sum constraint induces negativity in the covariance matrix,

$$\text{cov}(x_1, x_2) + \dots + \text{cov}(x_1, x_D) = -\text{var}(x_1). \quad (3)$$

Equation [3](#) shows that at least one element of each row of the covariance matrix must be negative. Aitchison refers to this as the “negative bias difficulty” (although ‘bias’ is not used in the traditional sense; ([Aitchison 1986](#), p. 53)). The structurally induced negative values create problems for the interpretation of the covariance matrix. Similarly, the use of naive distance metrics in the simplex may not be interpretable as in Euclidean space. Because of these difficulties, standard statistical methodology is not always appropriate ([Aitchison 1986](#)) and can produce misleading results ([Lovell et al. 2015](#)).

To overcome these obstacles, Aitchison ([1980](#)) proposed working in ratios of components. We focus on the Centered Log-Ratio (CLR) which treats the parts of the composition symmetrically and provides an informative covariance structure. The CLR transformation is defined for a D -part composition \mathbf{x} as:

$$y_i = \text{CLR}(x_i) = \log \left(\frac{x_i}{g(\mathbf{x})} \right), \quad (4)$$

Dominic LaRoche

this is page 4

where $g(\mathbf{x})$ is the geometric mean of \mathbf{x} . The $D \times D$ covariance matrix is then defined as:

$$\Gamma = [\text{cov}(y_i, y_j) : i, j = 1, \dots, D] \quad (5)$$

The CLR transformation is similar to the familiar Counts per Million (CPM) transformation (Law et al. 2014) defined as, $\log_2 \left(\frac{r_{gi} + 0.5}{t_i + 1} \times 10^6 \right)$, where r_{gi} is the number of sequence reads for each probe (g) and sample (i), (scaled to avoid zero counts), adjusted for the number of mapped reads (library count) for each sample t_i (scaled by a constant 1 to ensure the proportional read to library size ratio is greater than zero). The primary difference between the CLR and log(CPM) transformations is in the use of the geometric mean in the denominator of the CLR transformation. The use of the geometric mean results in subtracting the mean of the log transformed values from each log-transformed element thereby centering the vector of log-ratio transformed read counts. The difference appears minor but has important implications for the application of several common statistical methods.

Although the CLR transformation preserves the original dimension of the data, and gives equal treatment to every element of \mathbf{x} , the resulting covariance matrix, Γ , is singular. Therefore, care should be taken when using general multivariate methods on CLR transformed data. Aitchison (1986) proposed an alternative transformation, the additive log-ratio (ALR), which does not treat the components symmetrically but results in a non-singular covariance matrix. The ALR transformation is defined as,

$$y_i = \text{ALR}(x_i) = \log \left(\frac{x_i}{x_D} \right), \quad (6)$$

where x_D , the D^{th} component of x , can be any component.

As noted above, the compositional geometry must be accounted for when measuring the distance between two compositional vectors or finding the center of a group of compositions (Aitchison et al. 2000). Aitchison (1992) outlined several properties for any compositional difference metric which must be met: scale invariance, permutation invariance, perturbation invariance (similar to translation invariance for Euclidean distance), and subcompositional dominance (similar to subspace dominance of Euclidean distance). The scale invariance requirement is ignorable if the difference metric is applied to data on the same scale (which is generally not satisfied in raw RNA-seq data due to differences in read depth). The permutation invariance is generally satisfied by existing methods such as Euclidean distance (Martín-Fernández et al. 1998). However, the perturbation invariance and subcompositional dominance are not generally satisfied (Martín-Fernández et al. 1998).

Aitchison (1986; 1992) suggests using the sum of squares of all log-ratio differences. Billheimer, Guttorm, and Fagan (2001) use the geometry of compositions to define a norm which, along with the perturbation operator defined by Aitchison (1986), allow the interpretation of differences in compositions. Martin-Fernandez et al. (1998) showed that applying either Euclidean distance or Mahalanobis distance metric to CLR transformed data satisfies all the requirements of a compositional distance metric. Euclidean distance on CLR transformed compositions is referred to as Aitchison distance:

$$d_A(x_i, x_j) = \left[\sum_{k=1}^D \left(\log \left(\frac{x_{ik}}{g(x_i)} \right) - \log \left(\frac{x_{jk}}{g(x_j)} \right) \right)^2 \right]^{\frac{1}{2}} \quad (7)$$

Dominic LaRoche

this is page 5

or

$$d_A(x_i, x_j) = \left[\sum_{k=1}^D (clr(x_{ik}) - clr(x_{jk}))^2 \right]^{\frac{1}{2}}. \quad (8)$$

To avoid numerical difficulties arising from sequence targets with 0 reads, Martin-Fernandez et al. (2000) suggest an additive-multiplicative hybrid transformation. If zeros are present in the data We recommend using the Martin-Fernandez transformation with a threshold value of $\delta = \frac{0.55}{\text{Total Reads}}$ to account for differences in sequencing depth. The CLR transformation is then applied to the Martin-Fernandez transformed data which contains no zeros.

Up to this point we have referred to the total reads available per sequencing run, \mathbb{T} . However, it is more typical to work with the aligned reads in practice. The total aligned reads, T , is always a fraction of the total reads available for a sequencing run, \mathbb{T} . The fraction of the total reads aligned can be affected by multiple factors, including the choice of alignment algorithm, which we do not address here. We assume that T imposes the same constraints on the data as outlined above for \mathbb{T} and will refer exclusively to T hereafter.

3 Sample Quality Control

Problems with sample quality, library preparation, or sequencing may result in a low number of reads allocated to a given sample within a sequencing run. The Percent Pass Filter (% PF) metric provided on Illumina sequencers provides a subjective measure that can identify problems with sequencing that result in a low number of reads allocated to a sample. However, % PF will not necessarily catch problems associated with poor sample quality or problems with sample pre-processing since these processes may affect cluster generation, and not just cluster quality. This is particularly important for extraction-free RNA-Seq technologies, such as the HTG EdgeSeq^(tm), which allow for the use of smaller input amounts but lack the intermediate steps for checking sample quality. There is currently no objective way to evaluate sample quality based on the total number of reads attributed to a sample. We propose a method for objectively identifying problematic samples based on the total number of reads allocated to the sample.

For most experimental designs we expect the number of reads allocated to each sample in a sequencing run to arise from the same general data generating mechanism, namely the chemistry of the NGS-based measurement system, regardless of experimental condition. The objective is then to determine which samples arise from a different mechanism. Outlier detection is well suited for discovering observations that deviate so much from other observations that they are likely to have arisen from a different mechanism (Hawkins 1980). We base our method off Tukey's box-plots (Tukey 1977), which is a commonly used and robust method for detecting outliers (Ben-Gal 2009).

We expect the total number of reads allocated to each sample, t_i , to be equivalent notwithstanding random variation. For a given sequencing run with D samples we define the vector of total reads allocated to each sample as \mathbf{t} . Since the D -dimensional vector \mathbf{t} is a composition we have $\mathbf{t} \in \mathcal{S}^{D-1}$, the $D-1$ -dimensional simplex. As noted above, traditional statistical methods may not be appropriate for data in the simplex. Therefore, we map $\mathbf{t} \in \mathcal{S}^{D-1} \rightarrow \mathbf{x} = CLR(\mathbf{t}) \in \mathcal{R}^D$ using the Centered Log Ratio transformation 4. We then apply Tukey's method for detecting outliers to \mathbf{x} , which simply identifies those observations which lie outside 1.5 times the inter-quartile range.

Definition 1. x_i is a quality control sample failure if $x_i < \text{lower-quartile} - 1.5 \times \text{IQR}$ or $x_i > \text{upper-quartile} + 1.5 \times \text{IQR}$, where IQR is the interquartile range of \mathbf{x} .

We demonstrate the utility of our sample quality control measure using two sets of targeted RNA-

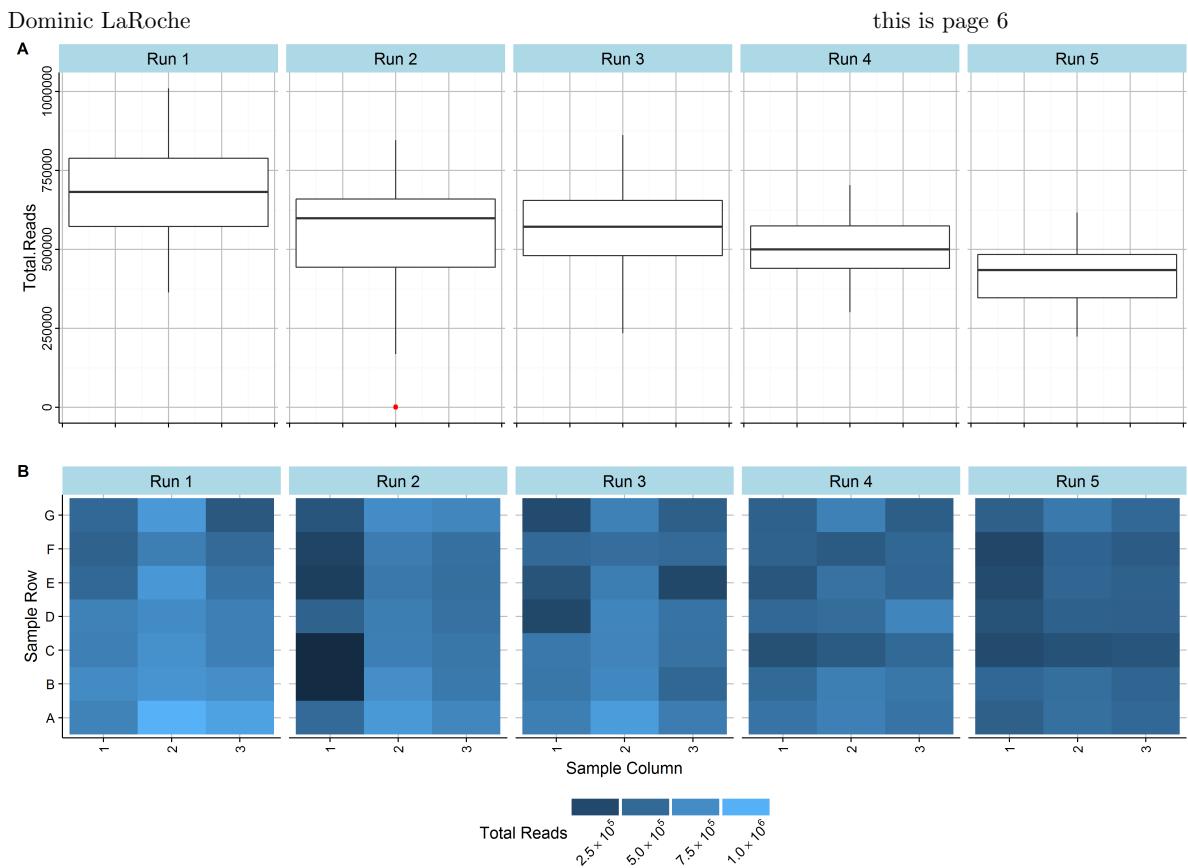


Figure 1: A) Distributions of total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. Only 1 sample is identified as a problematic sample. B) Heat-maps showing the relative totals for each sample within each run. The darker heat-maps for runs 4 and 5 reflect the generally lower number of total reads in those sequencing runs as compared to runs 1 and 2. This is caused by normal variation in the number of reads available in a sequencing run.

Seq data: 1) 120 mRNA technical replicate universal-RNA samples prepared with the HTG Edge-Seq Immuno-Oncology assay and sequenced in 5 different equally sized runs, and 2) 105 miRNA technical replicate samples of human plasma, FFPE tissue, and Brain RNA prepared with the HTG EdgeSeq Whole Transcriptome miRNA assay. These two data sets differ in the both the type of RNA (mRNA versus miRNA) and the number of sequence targets in each assay (558 versus 2,280 targets, for the mRNA and miRNA assays respectively). All samples were prepared for sequencing using the HTG EdgeSeq Processor and sequenced with an Illumina Mi-Seq sequencer.

We compare the utility of our method to evaluation of the un-transformed total counts. Figure 1 shows a boxplot and heat-map of the total number of reads allocated to each sample for each of 5 sequencing runs. Figure 2 shows the same data after CLR transformation. After transformation the poor samples become much more visually evident in the heat maps. Additionally, the ability to detect outlying values increases and the number of poor samples detected increases from 1 to 6.

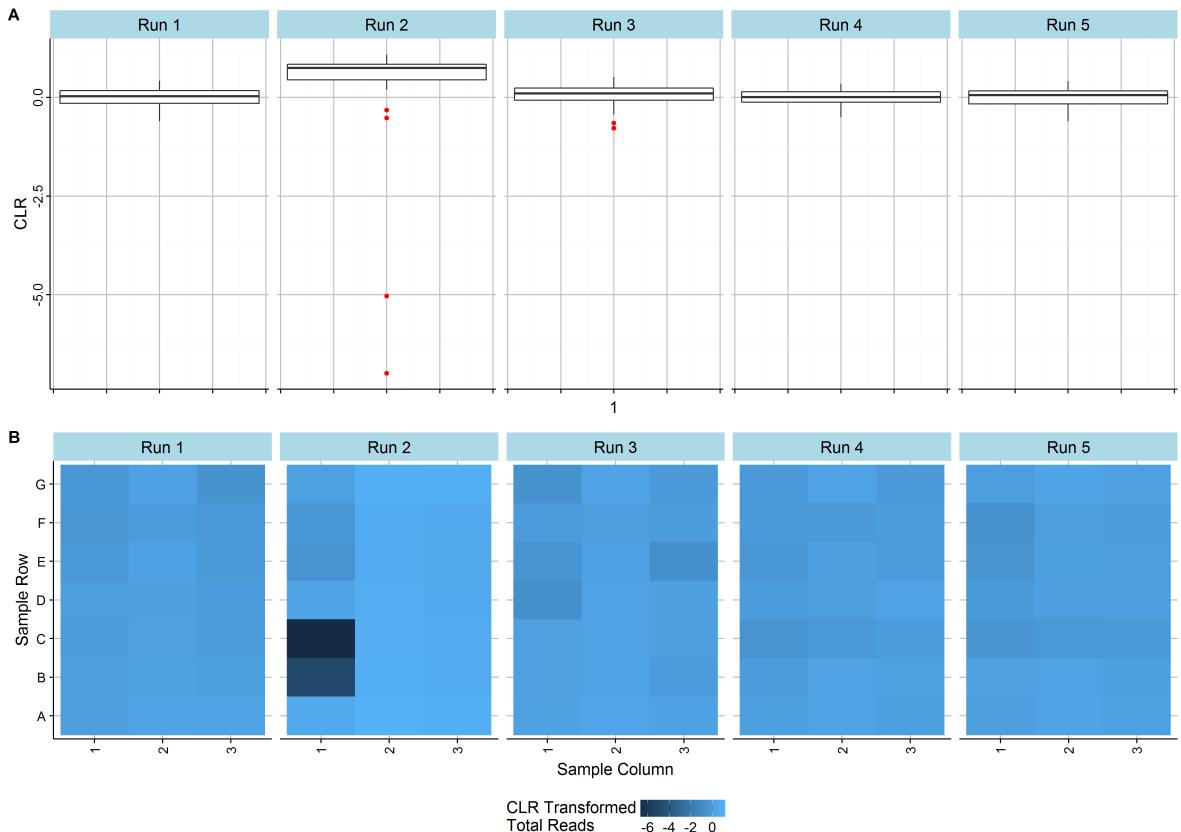


Figure 2: A) Distributions of CLR transformed total reads allocated to each sample in 5 runs on an Illumina Mi-Seq sequencer. After CLR transformation, 6 samples are identified as problematic. B) Heat-maps showing the relative CLR transformed totals for each sample within each run.

4 Testing for Compositional Invariance

Normalization and standardization methods for RNA-Seq generally assume that the total number of reads assigned to a sample does not affect the observed relative frequencies of probes within an assay. For example, implicit in the CPM transformation is the idea that if you re-scale the counts (by dividing by the total for each sample) then the resulting counts are comparable and any differences are due to underlying differences in expression. Other methods which apply a scaling factor to each sample, such as Trimmed-mean of M values (TMM) or Quantile normalization, also rely on this assumption. In the parlance of compositional data these methods assume *Compositional Invariance*, i.e. the underlying composition is statistically independent of the total size of the composition (the total counts for a sample, t).

Compositional invariance (CI) is an important property for RNA-Seq data which enables the comparison of samples with differing read depths. However, it is well documented that the quality of RNA-Seq depends on the read depth of the sequencing run with higher read-depths associated with higher quality data ([Tarazona et al. 2011](#); [Sims et al. 2014](#)). Read depth may affect the measurement of relative abundances for the target RNA sequences as some targets may receive proportionally more reads as the read depth increases. This would be a direct violation of CI and could lead seemingly differential expression between samples with different read depths, even after normalization. Another form of CI violation, that is perhaps more likely in RNA-Seq experiments, is the dependence between the variance of read counts and the read depth.

Aitchison ([1986](#)) outlined a simple model for testing compositional covariance using the ALR transformation,

$$[y_1 \dots y_d] = [1 \ t] \begin{bmatrix} \alpha_1 & \dots & \alpha_d \\ \beta_1 & \dots & \beta_d \end{bmatrix} + [e_1 \dots e_d], \quad (9)$$

where $y_1 \dots y_d$ are the d ALR transformed components, t is the vector of sample total aligned reads, $\alpha_1 \dots \alpha_d$ are the probe specific log-ratio intercepts, and $\beta_1 \dots \beta_d$ are the coefficients relating the the total aligned reads to the relative expression of the probe. A test for compositional invariance for the experiment then becomes a test of the null hypothesis, $H_o : \beta_1 = \dots = \beta_d = 0$. This test can be re-parameterized to test for dependence between the variance and total aligned reads as well.

Unfortunately, the small sample sizes and large number of probes typically associated with RNA-Seq experiments complicates the application of Aitchison's model. We propose an alternative visualization for detecting simultaneously detecting both violations of compositional invariance. We use the multivariate Aitchison distance (8) between all pairs of samples in a heat-map with the samples ordered by total aligned reads. If CI is violated we expect pairs samples with similar total aligned reads will have smaller scalar distances than those with large differences in total aligned reads. This will result in visual clustering around the 45 degree axis. If the variance depends on the total aligned reads, we expect the scalar distance between sample pairs to decrease with increasing read depth resulting in a visual gradient in the distance heat map.

We demonstrate this visualization with two sets of miRNA samples (Fig. 4) and two sets of mRNA samples(Fig. 4). The miRNA samples are composed of 40 technical replicates each of (A) plasma samples and (B) brain samples. In the miRNA data there is a clear gradient along the 45 degree axis for the plasma samples (Fig. 4.A). This indicates a dependence between the total aligned reads and the variance of the samples (as indicated by the increasing multivariate distance between replicates as the total aligned reads decreases). In contrast, there is no clear gradient in the brain samples

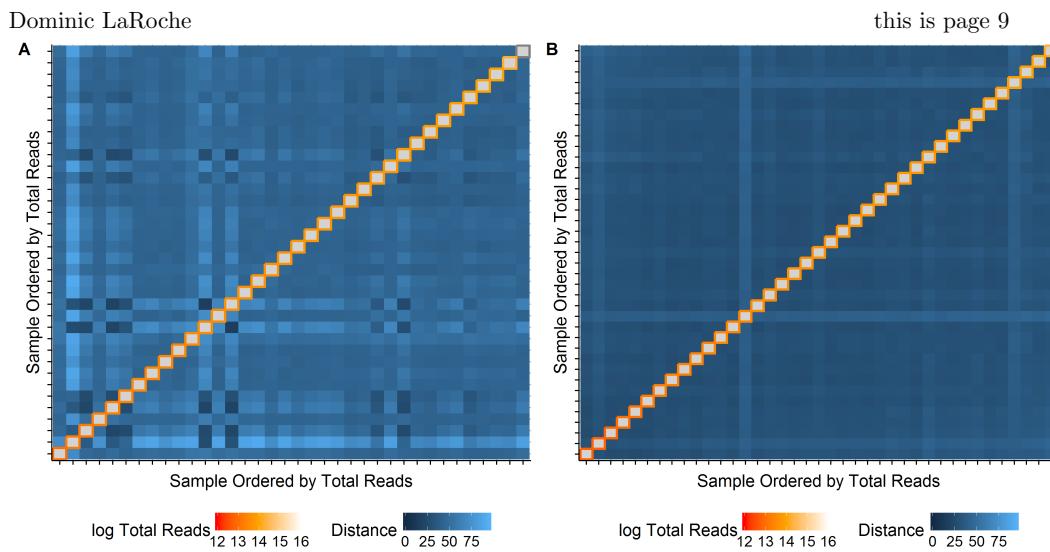


Figure 3: Two sets of miRNA samples with samples in (A.) showing a violation of compositional invariance and (B) showing compositional invariance.

(Fig. 4.B). The mRNA samples are composed of (A) 14 technical replicates of universal RNA and (B) 16 technical replicates of pancreas tissue. In the universal RNA there is a small cluster of samples with low total aligned reads which are more distant from samples with greater total aligned reads (Fig. 4.A). This indicates that the composition is dependent on the total aligned reads, a violation of compositional invariance for these samples. In contrast, the pancreas samples show no such pattern related to total aligned reads (Fig. 4.B).

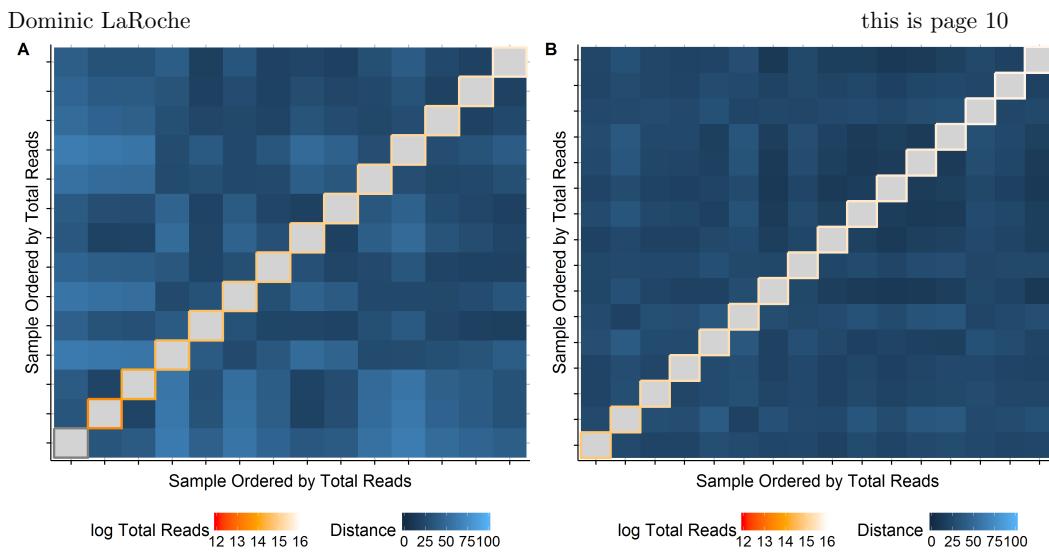


Figure 4: Two sets of mRNA samples with samples in (A.) showing a violation of compositional invariance and (B) showing compositional invariance.

5 Batch Effects and Normalization

Batch effects arising from differing laboratory conditions or operator differences have been identified as a problem in high-throughput measurement systems (Leek et al. 2010; Chen et al. 2011). Identifying and controlling for batch effects is a critical step in the transition of RNA-Seq from the lab to the clinic. Batch effects are typically identified with a hierarchical clustering (HC) method or principal components analysis (PCA). For both methods, the multivariate distance between the samples is visualized, either in a biplot for PCA or a dendrogram for HC, to check for the existence of clusters of samples related to batch. The compositional nature of RNA-Seq data has important implications for the detection of batch effects due to the incompatibility with standard measures of distance between compositions as noted above (Aitchison 1986; Martín-Fernández et al. 1998).

The next generation sequencing process results in arbitrary differences in scale among samples as some samples will receive more total reads than others. Principle components analysis is sensitive to differences in scale among the variables, failure to remove these difference can mask potential batch effects and leave unwanted technical variation in the data. Most normalization methods use a scaling factor calculated for each sample to re-scale the read count for each gene within the sample (Dillies et al. 2013). The CLR transformation can similarly be viewed as a scaling normalization (with the scale factor chosen as the inverse of the geometric mean $1/g(x)$). Unlike other normalization methods, the CLR transformation has the added benefit of being applied at the individual sample level, not experiment wise, and requires no assumptions about differential expression among samples, unlike other popular normalizations. This makes it particularly well suited for the clinic where there are generally no reference samples for normalization.

Aitchison demonstrated that the CLR transformation has several other useful properties in addition to re-scaling the data (Aitchison 1986), particularly with respect to PCA biplots (Aitchison and Greenacre 2002). Most notably for the detection of batch effects, the distance between any two points representing samples in the form-biplot approximates the Euclidean distance between the two samples or the Mahalanobis distance for covariance biplots. The CLR transformation retains the property that this distance is at least as great as the distance between any corresponding subset

Dominic LaRoche

this is page 11

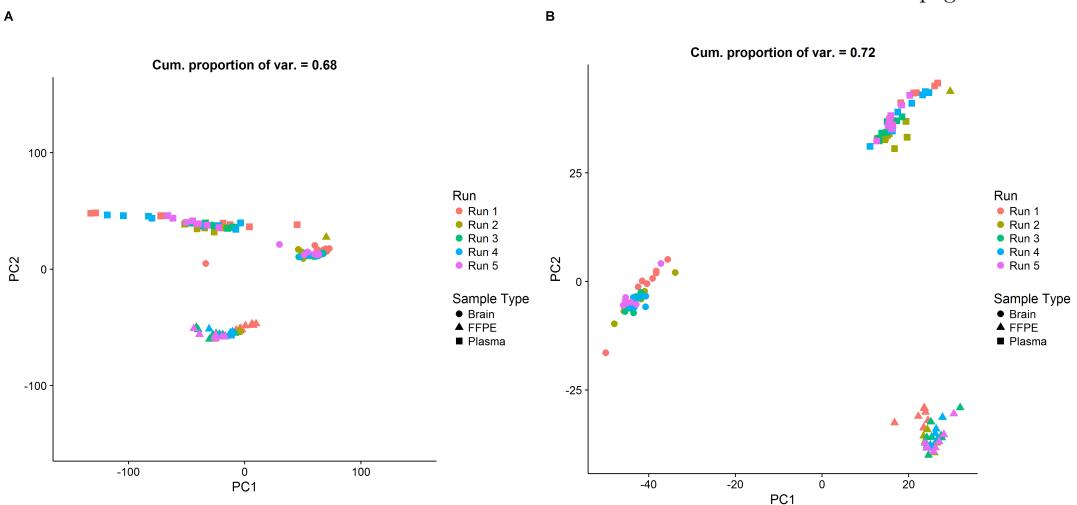


Figure 5: Principle component analysis of A) log-transformed and B) CLR-transformed read count data. The differences between sample types is much greater than the batch effects in both transformation. The CLR transformation results in tighter sample type clusters resulting from less variation along the first principle component.

of these two compositions (subspace dominance). Additionally, the euclidean distance between two CLR transformed samples is location invariant. Other scaling and normalization methods do not necessarily satisfy these properties and, therefore, batch effects may be masked or artificial.

We demonstrate the use of the compositional biplot to detect batch effects using technical replicates of three sample types: brain, plasma, and fresh frozen paraffin embedded (FFPE). Each sample is replicated 8 times in each of 5 sequencing runs for a total of 120 samples. Samples were prepared using the EdgeSeq Whole Transcriptome miRNA assay which measures 2,280 targets including including 11 control probes and 2,269 unique miRNA probes. All sequencing was performed on an Illumina Mi-seq^(tm) sequencer.

We create a second data set, by re-scaling the original data, to better illustrate the effects of changes in read depth on batch effect detection. To re-scale the samples from the original we multiply every read count in a given sample by a factor, ranging from 0.5 to 1.5, randomly generated from the uniform distribution. We then obtain a new data set in which the proportions between the read counts remains unchanged but the variance in the total number of reads among the samples is increased.

We perform a PCA on log-transformed and CLR transformed data. We then construct form-biplots of the first two principle components for each transformed data set (Fig. 5). The differences between the 3 samples types (brain, plasma, and FFPE) dominate the first two principle components for both data sets. However, the CLR transformed data provides tighter clusters, relative to the distance between the clusters, than the log-transformed raw data. There is also a single FFPE sample which is closer to the brain samples than the other samples. It is worth noting that this sample would have been removed using our proposed quality control metric.

Since the sample type differences overwhelm the potential batch effects we performed a second PCA on only the brain samples for both transformed data sets (Fig. 5). Both biplots exhibit clustering by batch but the CLR transformed data shows better separation between the batches,

Dominic LaRoche

this is page 12

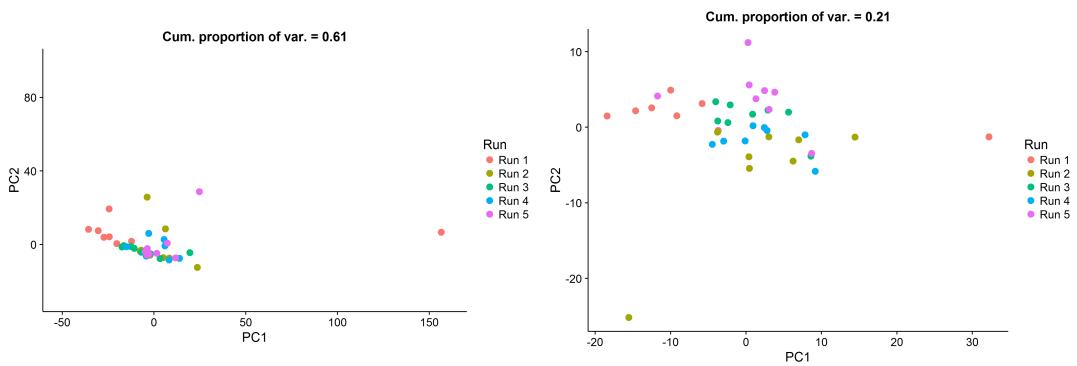
A**B**

Figure 6: Principle component analysis of only brain samples from A) log-transformed and B) CLR-transformed read count data. The batch effects are more easily identified in the CLR transformed data.

although batches are still overlapping.

Some of the batch effects detected in the log-transformed data may be attributable to the differences in total reads between batches. By randomly re-scaling each sample by a constant we are able to break the relationship between batch and the total reads in a sample. Figure ?? gives the biplots for log-transformed and CLR-transformed randomly re-scaled data. The sample type clusters in the log-transformed data become more diffuse while the CLR-transformed biplot remains unchanged. Most notably, the batch effects previously visible in the log-transformed brain samples become completely obscured in the randomly re-scaled data but remain unchanged in the CLR-transformed data (Fig. 5).

Dominic LaRoche

this is page 13

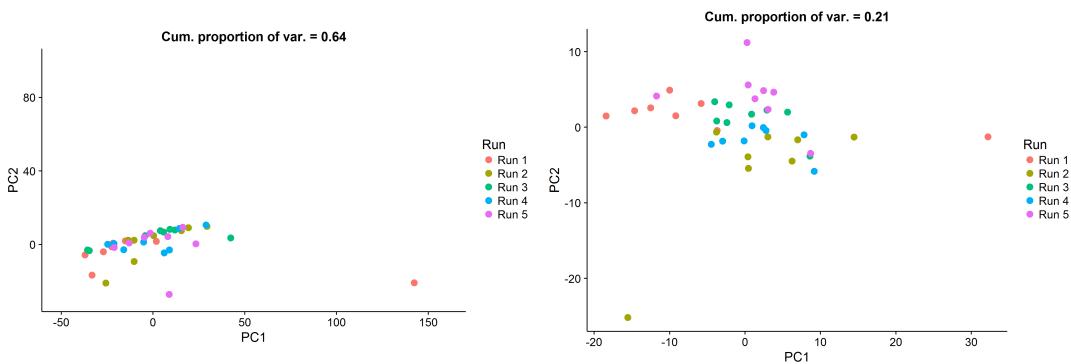
A**B**

Figure 7: Principle components analysis of the randomly re-scaled brain samples for A) log-transformed and B) CLR-transformed read count data. The batch effects visual in the log-transformed raw data disappear after random re-scaling whereas the batch effects remain identifiable in the CLR transformed data.

6 Discussion

Our sample quality control metric can identify problematic samples which arise from multiple failure modes, e.g. a low quality sample or a sequencing problem. However, it is conceivable that a sample might have an unusually low (or high) number of reads and still provide quality information. In certain experimental designs one might be able to further evaluate these samples with a PCA biplot on the CLR transformed data. In our PCA analysis we identified a FFPE sample which would have failed our quality control and was clearly very different from the other technical replicates. However, if this sample had remained quite similar to the other FFPE replicates this would have provided information that the sample may still be valuable. In this way, the quality control metric and PCA biplot can be used in tandem to provide additional information about the quality of a sample.

The compositional invariance visualization is a logical extension of the sample quality control metric since the assumption of the sample quality control is that the total number of aligned reads is related to the proportional allocation of reads within the sample. As noted above samples which violate the compositional invariance property may still contain valuable information. The identification compositional invariance violations allows the investigator to account for the dependency between the total aligned reads and the relative abundance of transcripts within the samples when modelling.

The principal components analysis biplot is a well known dimension reduction visualization. For the current data the dimension is reduced from 2,280 probes to 2 principle components. The utility of the data reduction, including the quality of the approximation of the multivariate distance between the samples, is proportional to the amount of variance explained by these two principle components. In our data the first two principle components explain between 72 and 21 percent of the variation in the data. The analysis with the lowest percent of variation explained by the first 2 components is of the CLR-transformed brain samples. Surprisingly, batch effects are still visible

Dominic LaRoche

this is page 14

in this plot, in which case they can be removed ([Luo et al. 2010](#)).

As RNA-Seq makes the transition from the research laboratory to the clinic there is a need for robust quality control metrics. The realization that RNA-Seq data are compositional opens the door to the existing body of theory and methods developed by John Atchison and others. We show that the properties of compositional data can be leveraged to develop new metrics and enhance existing methods.

- Aitchison, J. (1986, oct). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.
- Aitchison, J. (1992, may). On criteria for measures of compositional difference. *Mathematical Geology* 24(4), 365–379.
- Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology* 32(3), 271–275.
- Aitchison, J. and M. Greenacre (2002, oct). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51(4), 375–392.
- Aitchison, J. and S. Shen (1980, aug). Logistic-normal distributions: Some properties and uses. *Biometrika* 67(2), 261–272.
- Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biol* 11(10), R106.
- Ben-Gal, I. (2009). Outlier Detection. In *Data Mining and Knowledge Discovery Handbook*, pp. 117–130. Boston, MA: Springer US.
- Billheimer, D., P. Guttorp, and W. F. Fagan (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association* 96(456), 1205–1214.
- Chen, C., K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE* 6(2).
- Dillies, M. A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, N. S. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Lalo??, C. Le Gall, B. Scha??ffer, S. Le Crom, M. Guedj, and F. Jaffr??zic (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* 14(6), 671–683.
- Hawkins, D. M. (1980). *Identification of Outliers*. Dordrecht: Springer Netherlands.
- Law, C. W., Y. Chen, W. Shi, and G. K. Smyth (2014, jan). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* 15(2), R29.
- Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. a. Irizarry (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* 11(10), 733–739.
- Lovell, D., W. Müller, J. Taylor, A. Zwart, and C. Helliwell (2011). Proportions, Percentages, PPM: Do The Molecular Biosciences Treat Compositional Data Right? In *Compositional Data Analysis: Theory and Applications*, Number October, pp. 191–207. John Wiley & Sons, Ltd.
- Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS computational biology* 11(3), e1004075.
- Luo, J., M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zhao, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods, and J. Zhang (2010, aug). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The pharmacogenomics journal* 10(4), 278–91.
- Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology* 35(3), 253–278.

- Dominic LaRoche this is page 16
- Martín-Fernández, J. A., C. Barceló-Vidal, V. Pawlowsky-Glahn, A. Buccianti, G. Nardi, and R. Potenza (1998). Measures of difference for compositional data and hierarchical clustering methods. *Proceedings of IAMG* 98(1), 526–531.
- Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs : Pearson, K. : Free Download & Streaming : Internet Archive. *Proceedings of the Royal Society of London* 60, 489–498.
- Robinson, M. D. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* 11(3), R25.
- Robinson, M. D. and G. K. Smyth (2007). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9(2), 321–332.
- Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting (2014, jan). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15(2), 121–132.
- Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa (2011, dec). Differential expression in RNA-seq: a matter of depth. *Genome research* 21(12), 2213–23.
- Tukey, J. W. J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co.

Compositional data analysis of the stream sediment geochemical data in the Duolong mineral district, Tibet, China

X.C Liu¹, W.L. Wang¹, and Y.R. Pei¹

¹Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China; xcliu@cags.ac.cn

Abstract

Ten porphyry and epithermal Cu-Au deposits are recently found in the Duolong mineral district, northwest Tibet, China. There are still great prospecting potential of copper resources in Duolong. Previous work on the stream sediment geochemical data in this area did not consider the closure effect caused by the compositional data. The geochemical data in Duolong are derived from 3,217 samples and each sample records concentration values of 15 trace elements (Cu, Au, Pb, Zn, Cr, Ni, Mn, Ag, Sn, W, Mo, As, Sb, Bi, and Hg). In this contribution, we do principle component analysis of the 15 elements concentrations and separate their bimodal distributions using isometric logratio transformation and minimum message length and expectation maximization algorithm (MML-EL), respectively. We find that all the elements except Hg follow bimodal distributions. The right skew part of these bimodal distributions represents the high-average population. The low-average population is interpreted to represent the background distribution and the high-average population reflects the influences of multiple magmatic activities in this area. The concentrations of Hg follow a single log-normal distribution. The element associations in the first principle component are interpreted to be indicators for the Cu-Au mineralization potential in Duolong. The areas with high scores of the first component are consistent with most of the found Cu-Au deposits and the alteration zones exposed at surface. Four zones with high scores are suggested for further investigation on their mineral potential. Our compositional data analysis and bimodal separation provide useful information for understanding the geological and geochemical processes and aiding in further exploration in this area.

Kew words: compositional data, stream sediment data, bimodal distribution, Duolong, porphyry copper deposits.

1 Introduction

The Duolong mineral district is a newly discovered area with ten porphyry and epithermal Cu-Au ore deposits in the northwest Tibet, China (Li and others, 2007; Wang and Others, 2017). The current copper resources in this area have reached 16 million tonnages and the prospective copper resources are over 20 million tonnages (Li and others, 2016). The stream sediment geochemical data in this area provide important information for indentifying the anomaly of ore elements (Cu and Au) and narrowing the prospecting areas; however, the geochemical data are compositional data and their closure effect is ignored in existing exploration models and practice (e.g. Li and others, 2012; Wang and others, 2016). In this study, isometric log-ratio (ILR) transformation is applied to these geochemical data before doing principal component analysis. Our first aim is to extract geological and geochemical information from these geochemical data and aid in further mineral exploration in this area. The latest geological map (Fig. 1) is not broadly accepted because many important geological features like faults and intrusions are ill-exposed. Thus, our second aim is to aim to lithologic mapping in Duolong.

2 Geological Background

The Duolong mineral district is located on the south rim of the southern Qiangtang Terrane and the north part of Bangongco-Nujiang Suture zone (Fig. 1). Magmatic activities in this district are multiple and intensive. The magmatic rocks exposed in Duolong are granite porphyry, granodiorite porphyry, diorite, diorite porphyrite, basalt, and gabbro. All intermediate and felsic intrusions have similar emplacement ages of 116~128 Ma (Chen and others, 2013; Li and others, 2013, 2014, 2015; She and others, 2009; Wei and others, 2016), while the mafic dykes have a zircon U-Pb age of 126~127 Ma (Xu and others, 2017).

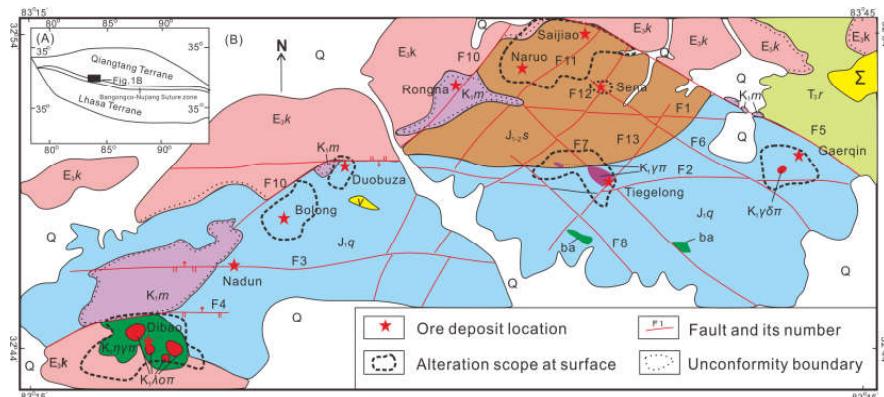


Figure 1: A simplified geological map of the Duolong mineral district, Tibet China, revised from Wang and others (2015). Q-Quaternary, E₃k-Upper Oligocene Kangtuo Formation, K₁m-Lower Cretaceous Meiriqieco Formation, J₁-s-Lower and Middle Jurassic Sewa Formation, J₁-q-Lower Jurassic Quse Formation, T₃r-Upper Triassic Riganpeicuo Formation, K₁γπ-Early Cretaceous monzonitic granite porphyry, K₁λσπ-Early Cretaceous granite porphyry, K₁γδπ-Early Cretaceous quartz porphyry, K₁γδπ-Early Cretaceous granodiorite porphyry, v-Gabbro, ba-Pillow basalt, Σ-Serpentinized olivinite.

Strata in the Duolong mineral district consist of the Upper Triassic Riganpeicuo Formation (T₃r), the Lower Jurassic Quse Formation (J₁q), the Middle Jurassic Sewa Formation (J₁s), the Lower Cretaceous Meiriqieco Formation (K₁m), the Upper Cretaceous Abushan Formation (K₂a), and the Upper Oligocene Kangtuo Formation (E₃k). The Riganpeicuo Formation (T₃r) is mainly composed of limestone in the northeast corner of Duolong. The Quse Formation (J₁q) is composed dark gray mudstone and feldspar-quartz siltstone. The Meiriqieco Formation (K₁m) is mainly composed of continental-face intermediate-basic volcanic rocks containing andesite, andesitic volcanic breccia, rhyolitic tuff, and tholeiite (Li and others, 2012).

Ten porphyry and epithermal Cu-Au deposits have been found in the last decade. The Cu-Au deposits are Dibao, Nadun, Bolong, Duobuzha, Rongna, Naruo, Sajiao, Sena, Tiegelong, and Gaerqin deposits (Li and others, 2012). These deposits are dated at 118~115 Ma (Li and others, 2011, 2013). Formation of these deposits is genetically related to the porphyritic granitoids emplaced below these deposits (Li and others, 2016a; Song and others, 2014; Zhu and others, 2015). The hydrothermal alteration is divided into potassic, intermediate argillic and propylitic alteration zones from the ore-bearing porphyry center outwards and

upwards (Li and others, 2016b). This is consistent with the typical alteration models from porphyry Cu-Au deposits (Sillitoe, 2000). Argillic and propylitic alteration is identified at surface (Wang and others, 2015). These exposed alteration zones at surface are direct indicators of ancient magmatic-hydrothermal systems at depth. The main ore minerals are Cu-bearing sulfides like chalcopyrite, bornite and pyrite,

3 Data and Methods

Currently used dataset are geological data and the stream sediment geochemical data. The geological data include the locations of 10 porphyry and epithermal Cu-Au deposits, fault traces, lithological units, and outcrops of magmatic rocks. The geochemical data at 1:50,000 scale with 0.5 km spatial resolution are composed of 3,217 samples and each sample records concentration values of 15 trace elements (Cu, Au, Pb, Zn, Cr, Ni, Mn, Ag, Sn, W, Mo, As, Sb, Bi, and Hg).

The statistical distribution of geochemical concentration data often contains at least two populations because of influences of multiple geological and geochemical processes (Carranza, 2009, P.68; Grunsky and Smee, 1999). Liu and others (2011) apply the minimum message length and expectation maximization algorithm (MML-EM) to separate the mixed distribution of geochemical data. This method is based on the minimum message length criterion and expectation-maximization algorithm (Figueiredo and Jain, 2002). The MML-EM algorithm has higher accuracies than the probability graphs in estimating the parameters of mixed distributions of element abundance (Liu and others, 2011). A MATLAB code of this algorithm is programmed and available on the developer's homepage. The MML-EM algorithm was employed to analyze the statistical distribution of the individual elements in Duolong.

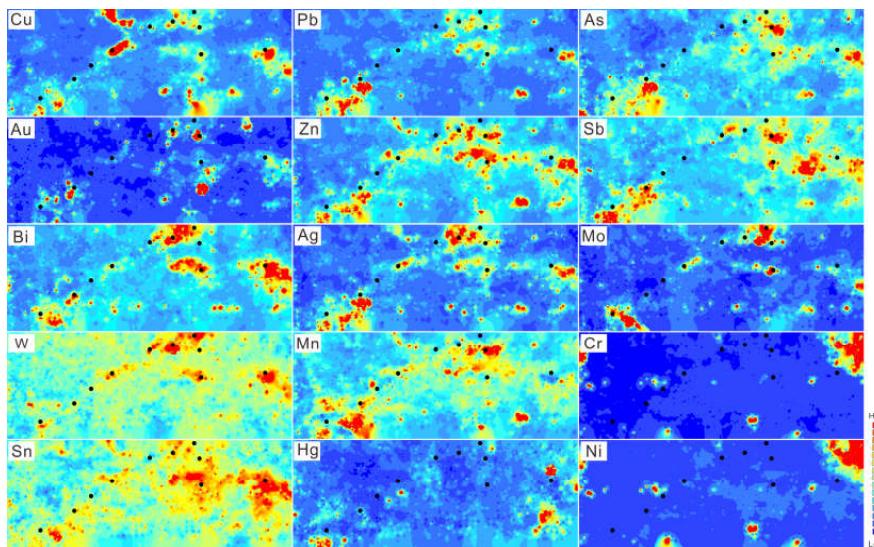


Figure 2: Plan maps of the stream sediment geochemical data in the Duolong District
The black solid circles mark the locations of the ore deposits. All the elements except Cr, Ni, Hg are concentrated mainly in the areas close to the found ore deposits. Cr and Ni have a similar spatial distribution.

Geochemical data are typical compositional data, but the closure effect among elements is often ignored in many cases (Buccianti and Grunsky, 2014). Statistical analysis directly applied to concentrations of major elements will lead to inappropriate interpretation because the compositional data cannot be correctly represented in Euclidean space and only carry relative information. The closure effect also exists in subsets of compositional data (Pawlowsky-Glahn and Buccianti, 2011). We used isometric log-ratio (ILR) transformation to overcome the closure effect (Egozcue and others, 2003). Then, principal component analysis (PCA) was used to reduce the dimensionality and identify the multi-element associations reflecting the geological and geochemical processes in Duolong. The principal component analysis was carried out using the correlation matrix of the data.

4 Results and Discussions

Plan maps of individual elements in Fig. 2 show that the ore elements (Cu, Au) and some others elements like As, Sb, Bi, Zn, and Ag are concentrated mainly in the areas close to the found ore deposits. These anomalies may have an association with the intermediate and felsic intrusions and later magmatic hydrothermal processes.

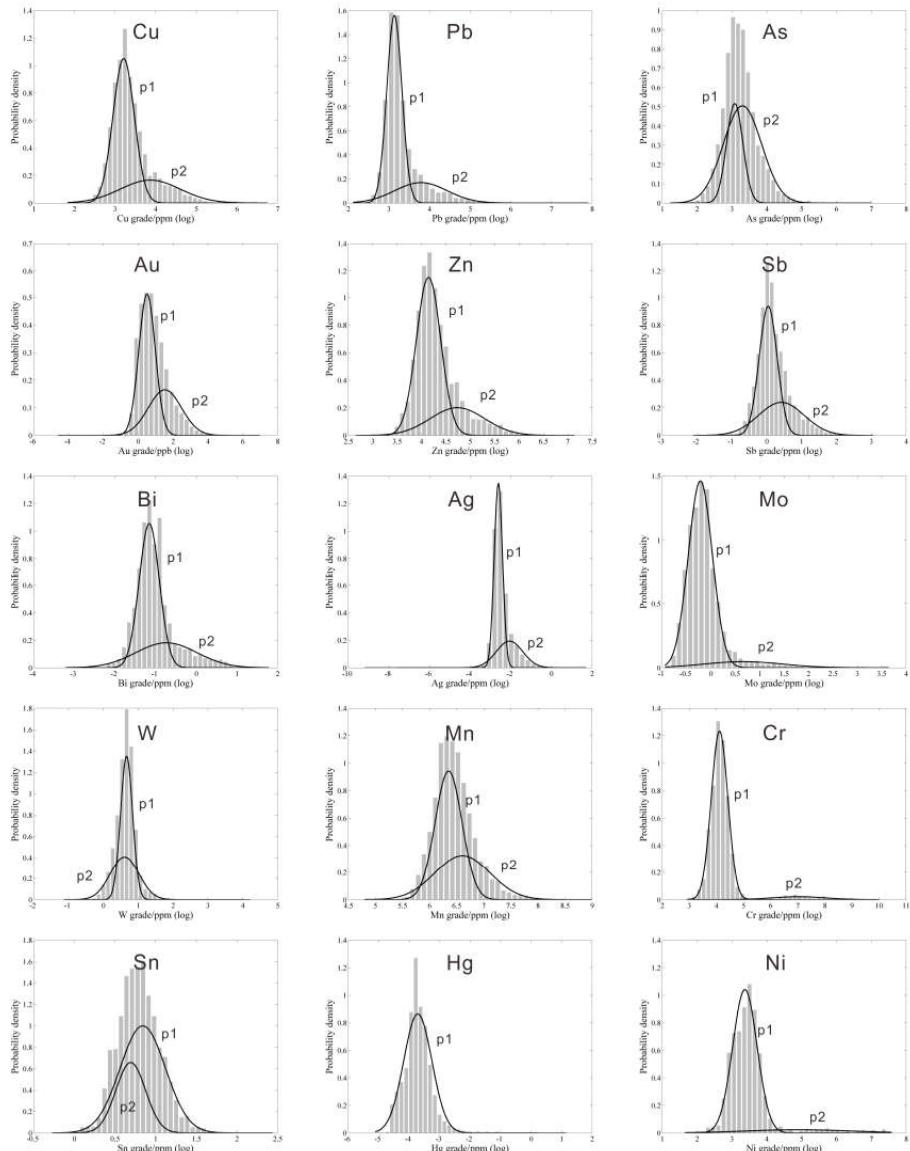


Figure 3: Separation of mixed distribution of geochemical concentration data in Duolong

The geochemical concentrations have been log-transformed before plotting the probability density distribution. The statistical distributions of all the elements except Hg are mixed with two log-normal populations (p1 and p2). The Hg concentrations follow a single log-normal distribution.

The spatial distribution of Cr is much similar to that of Ni, but these two elements have a significantly different spatial distribution from the other elements. Cr and Ni concentrations are very low in the most parts of Duolong and strong anomalies are mainly present in the northeast corner and a few small scattered areas. Limestone and serpentinitized olivinite outcrops in the northeast corner. Mafic and ultramafic rocks are often characterized by high concentrations of Cr and Ni (Oze and others, 2004);

therefore, high concentrations of Cr and Ni in the northeast corner are interpreted to be related to serpentinized olivinite and basalt. The anomalies of Cr and Ni in the other areas may be associated with some small mafic intrusions that are not identified at surface.

All the element data except Hg were partitioned into two log-normal populations (Fig. 3). The right-skew part of the bimodal distribution is actually a population with a higher average. The high-average populations of Cr and Ni are significantly separated from the low-average ones. The low-average populations of the bimodal elements except W and Sn are interpreted as the background distribution and the high-average ones may reflect the geochemical processes related to porphyry systems and mafic intrusions in Duolong. The two populations of W and Sn are more overlapped than other elements'. This means that the two populations are statistically close. This may reflect that W and Sn distributions are insignificantly influenced by magmatic activities in Duolong or the influence from magmatic activities has been altered by secondary processes. The statistical distribution of Hg concentration followed a single log-normal distribution.

We first compared the PCA results of log-based original data and those of ILR-based data. Rays of log-based PCA shows that all the elements except Cr are positive in the first component (Fig. 4), which is geochemically puzzling and actually reflects the closure effect. In contrast, the fifteen elements are scattered in the four quadrants in the biplot of ILR-based PCA. This indicates that the closure effect is overcome (Filzmoser and others, 2010).

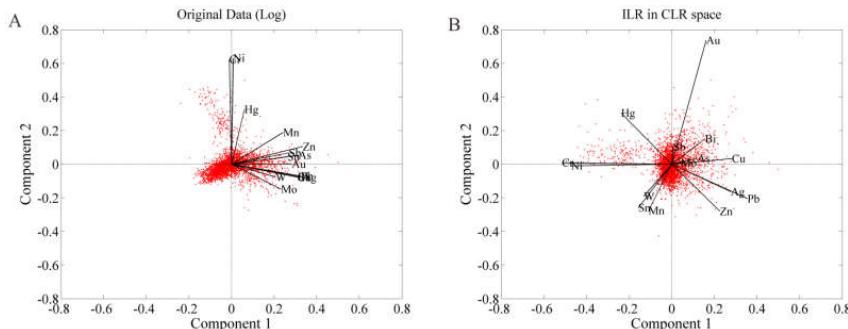


Figure 4: Biplots of principle component analysis of the original data (A) and ILR-transformed data (B)

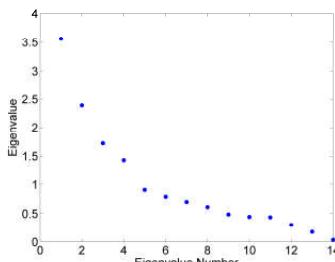


Figure 5: A plot of the eigenvalues in the ILR-transformed principle analysis

For the ILR-based PCA, the first component accounts for 25.4% of the variation of the data (Fig. 5). Cu, Au, Pb, Zn, Ag, Mo, As, Sb, and Bi have positive loadings in the first principle component. These elements are the chalcophile elements that readily form sulfides (White, 2013, p.261). Sulfides are the main metallic minerals at the porphyry and epithermal deposits in Duolong (Li and others, 2007). Strong anomalies of these seven elements are also close to the porphyry and epithermal deposits. Therefore, high values of these elements indicate a high mineral potential. The elements having negative loadings in the first principle component are Cr, Ni, Mn, W, Sn, and Hg. These six elements are lithophile or siderophile elements in the Goldschmidt's Classification (White, 2013, p.261). Cr and Ni are often associated with mafic and ultramafic rocks. This has been demonstrated in previous paragraphs. W, Sn, and Mn are the lithophile elements that have an affinity for oxygen (White, 2013, p.261). W and Sn are associated with S-type granites (Zhao and others, 2017), while porphyry Cu deposits are often genetically related to I-type granites (Sillitoe, 2010). The granites identified in Duolong favor for forming porphyry copper deposits.

Hg is a chalcophile element but often present in the distal parts of the primary halo of magmatic-hydrothermal deposits (Pirajno, 2009, p.363). Thus, the multi-element associations in the first principle component are interpreted to be indicators of the Cu-Au mineralization potential in Duolong. The scores of the first component can be used to map the mineral potential in Duolong.

We identified nine zones with high ILR-based PCA scores. The nine zones are numbered according to their correlation with the locations of ore deposits and the alteration scope exposed at surface. Nine of the ten ore deposits and six of the seven alteration zones exposed at surface are found in the first four zones.

Zone 1 contains three alteration zones at surface and five ore deposits (Rongna, Naruo, Sajjiao, Sena, and Tiegelong). The Rongna deposit is in the margin of this zone. Two ore deposits are located in Zone 2 and the Dibao deposit is buried below the alteration at surface. The Gaerqin deposit and the related alteration at surface are in the northwest part of Zone 3. It is common is that the elements having positive loadings in the first component have strong anomalies in these three zones.

The ILR-PCA scores in Zone 4 are not as high as those in the previous three zones. The Duobuza deposit and its exposed alteration are at the centre of Zone 4, but the Bolong deposit and its related alteration is outside this zone. It can be seen from Fig. 2 that the elements having positive loadings in the first component are not enriched in Zone 4. Thus, the reason why Zone 4 has high scores is that Cr, Ni, Mn, W, Sn, and Hg are depleted in this area. Further investigations are suggested to check why the surface above the Duobuza and Bolong deposits shows insignificant anomalies (or weak anomaly) of ore elements and other chalcophile elements. This may provide insightful information for further exploration in this area.

High PCA scores in Zone 5 are partly ascribed to high concentrations of Cu. Field work suggests that migration of Cu from the Rongna deposit due to weathering and surface drainage systems produce the Cu anomaly in this area.

Zone 6 has higher PCA scores than Zone 3 and Zone 4, but no ore deposits have been discovered yet. The high PCA scores in Zone 6 are interpreted be related to buried granite or porphyry deposits. A diorite is recently identified in this area, but its outcrop area is small (Xu and others, 2017). Given that porphyry and epithermal deposits in Duolong are genetically related to intermediate and felsic intrusions, we suggest further evaluation of its mineral potential. The last three zones are much smaller than the previous zones. We suggest further field work and evaluate their mineral potential.

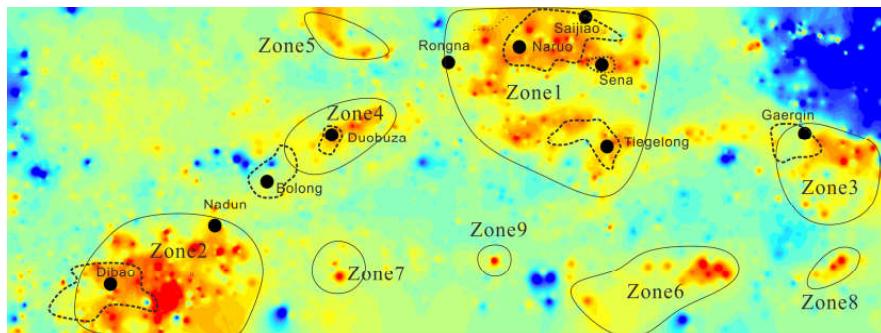


Figure 6: The score map of the first component in the ILR-based principle component analysis
Nine zones are marked with solid lines to represent the areas with a high mineral potential. The areas delineated by dotted lines are the alteration scope exposed at surface.

5 Conclusions

Previous exploration models ignore the closure effect of the stream sediment geochemical data in the Duolong district. In this contribution, we separate bimodal populations of these geochemical data and do principle component analysis in the composition data analysis framework. The main results are as follows:

- (1) The ore elements (Cu, Au) and other chalcophile elements (Pb, Zn, Ag, Mo, As, Sb, and Bi) are enriched in the areas close to the found Cu-Au deposits. High concentrations of Cr and Ni are associated with the exposed mafic rocks in Duolong.

- (2) The statistical distributions of all the elements except Hg contain two log-normal populations. The concentrations of Hg follow a single log-normal distribution. For the elements with a bimodal distribution, the right skew part of the statistical distribution represents the high-average population. The low-average population is interpreted to represent the background distribution and the high-average population reflects the influences of multiple magmatic activities in Duolong.
- (3) ILR-based principle component analysis suggests that the multi-element associations in the first principle component are indicators for the Cu-Au mineralization potential in Duolong. Nine zones are delineated based on the scores of the first component. The first four zones contain most of the found Cu-Au deposits and the alteration zones exposed at surface. Zone 5 is associated with the Rongna deposit. The remaining four zones are suggested for further investigation on their mineral potential. A small diorite is identified in Zone 6, therefore Zone 6 should be put a first priority.

Acknowledgements and appendices

The work is financially supported by National Natural Science Foundation of China (41402295), a Chinese Geological Survey project (DD20160026), the Chinese National Thousand Young Talents Plan to Wenlei Wang, and the State Key Program of National Natural Science of China (41430320). The authors appreciate Mr. Qin Wang from Chengdu University of Technology and Mr. Bin Lin from Institute of Mineral Resources for their assistance in the material collection and the fieldwork in Tibet.

References

- Carranza, E.J.M. (2009). *Geochemical Anomaly and Mineral Prospectivity Mapping in GIS*. Handbook of Exploration and Environmental Geochemistry, 11. Elsevier, Amsterdam, p.68.
- Chen, H.A., Zhu, X.P., Ma, D.F., Huang, H.X., Li, G.M., Li, Y.B., Li, Y.C., Wei, L.J., Liu, C.Q. (2013). Geochronology and geochemistry of the Bolong Porphyry Cu–Au Deposit, Tibet and its mineralizing significance. *Acta Geologica Sinica*, 87, 1593–1611 (in Chinese with English abstract).
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279-300.
- Filzmoser, P., Hron, K., & Reimann, C. (2010). The bivariate statistical analysis of environmental (compositional) data. *Science of The Total Environment*, 408(19), 4230-4238.
- Figueiredo, M.A.T., and Jain, A.K. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 24(3): 381-396.
- Grunsky, E.C. and Snee, B.W. (1999). The differentiation of soil types and mineralization from multi-element geochemistry using multivariate methods and digital topography. *Journal of Geochemical Exploration*, 67(1–3): 287-299.
- Li, G., Li, J., Qin, K., Zhang, T., & Xiao, B. (2007). High temperature, salinity and strong oxidation ore-forming fluid at Duobuza gold-rich porphyry copper deposit in the Bangonghu tectonic belt, Tibet: Evidence from fluid inclusions. *Acta Petrologica Sinica*, 23(5), 935-952.
- Li Y.B., Duo, J., and Zhong. W. T. (2012). An exploration model of the Duobuza porphyry Cu-Au deposit in Gaize Country, northern Tibet. *Geology and Exploration*, 48(2):0274-0287 (in Chinese with English abstract).
- Li, J.X., Qin, K.Z., Li, G.M., Xiao, B., Zhao, J.X., Chen, L. (2011). Magmatic-hydrothermal evolution of the Cretaceous Duolong gold-rich porphyry copper deposit in the Bangongco metallogenic belt, Tibet: Evidence from U-Pb and 40Ar/39Ar geochronology. *Journal of Asian Earth Sciences*, 41(6): 525-536.
- Li, J.X., Qin, K.Z., Li, G.M., Xiao, B., Zhao, J.X., Cao, M.J., Chen, L. (2013). Petrogenesis of ore-bearing porphyries from the Duolong porphyry Cu–Au deposit, central Tibet: evidence from U–Pb geochronology, petrochemistry and Sr–Nd–Hf–O isotope characteristics. *Lithos*, 160–161, 216–227.
- Li, J.X., Qin, K.Z., Li, G.M., Xia, B., Zhao, J.X., Cao, M.J., and Chen, L., (2014). Geochronology, geochemistry, and zircon Hf isotopic compositions of Mesozoic intermediate-felsic intrusions in central Tibet: petrogenetic and tectonic implications. *Lithos* 198–199, 77–91.
- Li, J.X., Qin, K., Li, G., Evans, N.J., Zhao, J., Cao, M., and Huang, F. (2016a). The Nadun Cu–Au mineralization, central Tibet: Root of a high sulfidation epithermal deposit. *Ore Geology Reviews*, 78: 371-387.
- Li, J.X. Qin, K.Z., Li, G.M., Xiao, B., Zhao, J.X., Chen, L. (2016b). Petrogenesis of Cretaceous igneous rocks from the Duolong porphyry Cu–Au deposit, central Tibet: evidence from zircon U–Pb geochronology, petrochemistry and Sr–Nd–Pb–Hf isotope characteristics. *Geological Journal*, 51(2): 285-307.

- Li, X.K., Li, C., Sun, Z.M., Wu, H. (2015). Geochronology and geochemistry of the diorite in Sajjiao Cu-Au Deposit, Tibet, and its mineralizing significance. *Geological Bulletin of China*, 34 (5), 908–918 (in Chinese with English abstract).
- Li, F., WANG, Y.H., Jiao, Y.J. (2016). Geophysical anomaly characteristics and prospecting direction of Duolong mining area. *Progress in Geophysics (in Chinese)*, 31(1) : 0217-0224.
- Liu, X.C., Hou, C., Shen, W., Zhang, D. (2011). MML-EM algorithm and its application on mixed distributions of geochemical data. *Earth Science*, 36(2): 355-359.
- Oze, C., Fendorf, S., Bird, D.K., Coleman, R.G. (2004). Chromium Geochemistry of Serpentine Soils. *International Geology Review*, 46(2): 97-126.
- Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis : theory and applications*: John Wiley & Sons, p.
- Pirajno, F., 2009. *Hydrothermal Processes and Mineral Systems*. Springer Netherlands, p.363.
- She, H.Q., Li, J.W., Ma, D.F., Li, G.M., Zhang, D.Q., Feng, C.Y., Qu, W.J., Pan, G.T. (2009). molybdenite Re-Os and SHRIMP zircon U-Pb dating of Duobuza porphyry copper deposit in Tibet and its geological implications. *Mineral Deposits*, 28 (6), 737–746 (in Chinese with English abstract).
- Sillitoe, R., 2000. Gold-rich porphyry deposits, descriptive and genetic models and their role in exploration and discovery. *Reviews in Economic Geology*, 13: 315-345.
- Sillitoe, R.H. (2010). Porphyry copper systems. *Economic Geology*, 105: 3-41.
- Song, Y., Tang, J. X., Qu, X., Wang, D., Xin, H., & Yang, C. (2014). Progress in the study of mineralization in the Bangongco-Nujiang metallogenic belt and some new recognition. *Advances in Earth Science*, 79, 795-809 (in Chinese with English abstract).
- Wang, D., Jiang, S., and Dong, F. (2016). Geological exploration of the Rongna porphyry copper deposit in the Duolong ore concentration area, northern Tibet. *Geology in China*, 2016, 43(5): 1599-1612(in Chinese with English abstract).
- Wang, Q., Tang, J., and Fang, X. (2015). Petrogenetic setting of andsites in Rongna ore block, Tiegelong Cu (Au - Ag) deposit, Duolong ore concentration area, Tibet: Evidence from zircon U-Pb LA-ICP-MS dating and petrogeochemistry of andsites. *Geology in China*, 42(5):1324-1336(in Chinese with English abstract).
- Wang, W.L., Cheng, Q.M., Tang, J.X., Pubuciren, Song, Y., Li, Y.B., Liu, .Z.B. (2017). Fractal/multifractal analysis in support of mineral exploration in the Duolong mineral district, Tibet, China. *Ore Geology Reviews*, in press.
- Wei Shaogang, Song Yang, Tang Juxing. (2016). Geochronology, geochemistry and petrogenesis of quartz diorite porphyrite from the Sena copper (gold) deposit, Tibet. *Geology in China*, 43(6): 1894-1912(in Chinese with English abstract).
- White, W.M., 2013. *Geochemistry*. John Wiley & Sons Inc, p.261.
- Xu, W., Li, C., Wang, M., Fan, J.J., Wu, H., Li, X. (2017). Subduction of a spreading ridge within the Bangong Co–Nujiang Tethys Ocean: Evidence from Early Cretaceous mafic dykes in the Duolong porphyry Cu–Au deposit, western Tibet. *Gondwana Research*, 41: 128-141.
- Zhao, W.W., Zhou, M.-F., Li, Y.H.M., Zhao, Z., Gao, J.F. (2017). Genetic types, mineralization styles, and geodynamic settings of Mesozoic tungsten deposits in South China. *Journal of Asian Earth Sciences*, 137: 109-140.
- Zhu, X., Li, G., Chen, H., Ma, D., Huang, H.. (2015). Zircon U-Pb, Molybdenite Re-Os and K-feldspar 40Ar/39Ar Dating of the Bolong Porphyry Cu–Au Deposit, Tibet, China. *Resource Geology*, 65(2): 122-135.

Compositional approach to the analysis of species abundance data

G.S. Monti¹, and S. Migliorati¹

¹ Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy
gianna.monti@unimib.it

Abstract

The investigation of species associations is a classical problem in ecology, in order to describe and predict environmental characteristics. In this contribution we apply the log-ratio approach to the analysis of species abundance data through the Aitchison geometry (Aitchison, 1986). The log-ratio transformations produce acceptable projections of the correlations among species in principal component space that will be used to analyze association among species. An application to a real dataset of the propose procedure is illustrated.

1 Introduction

In ecology several studies are focused on the variation in species compositions in several areas, in order to investigate the association among species, which is intrinsically connected with the ecological dynamics and changes. The niche theory (Hutchinson, 1957) shows the species distribution patterns and how the environmental conditions could influence the presence/absence of certain species in given sites of observation.

A species abundance data (sites by species) is a multivariate data table $\mathbf{Y} = [y_{ij}]$ of size $(n \times D)$ which collects the number of species ($j = 1, \dots, D$) found at certain sites ($i = 1, \dots, n$). The investigation of species association is fundamental to understand how different species respond to external factors such as climate changes or environmental pollution. Species associations arise when two or more species co-occur (presence-absence data) or are correlated (abundance data or community composition data) either more or less frequently than expected due to chance alone. Several statistical methods have been utilized to detect species associations, such as correlation analysis, analysis by contingency table, the use of cross-variograms and others (see Legendre and Legendre (1998, sec. 7.5 and 8.9) and Roxburgh and Chesson (1998)).

The original species abundance data \mathbf{Y} can be associated to the compositional matrix \mathbf{X} , whose row vectors \mathbf{x}_i are obtained normalizing, or closing, the corresponding \mathbf{y}_i vector, i.e. $\mathbf{x}_i = \mathcal{C}(\mathbf{y}_i) = (\sum_{j=1}^D y_{ij})^{-1} \mathbf{y}_i$, where $\mathcal{C}(\cdot)$ denotes the closure operation. Each component x_{ij} of the composition \mathbf{x}_i represents the proportion of the species j present in the i -th site. Each \mathbf{x}_i is thus a composition in the unit simplex $\mathcal{S}^D = \left\{ \mathbf{x}_i \mid x_{ij} > 0; \sum_{j=1}^D x_{ij} = 1 \right\}$, the suitable sample space for compositional data.

Here we briefly review some fundamental principles of the compositional data analysis which will be useful in the next section (see Aitchison (1986); Buccianti et al. (2006); Pawlowsky-Glahn and Buccianti (2011) for further details). The simplex $(\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$ has a $(D-1)$ -dimensional real Euclidean vector space structure, where \oplus and \odot are the perturbation and the powering operations and $\langle \cdot, \cdot \rangle_a$ represents the inner product. The geometry on the simplex is called *simplicial* or *Aitchison geometry* (Pawlowsky-Glahn and Egozcue, 2001). For $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, perturbation is defined as $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1 x_1^*, \dots, x_D x_D^*)$, and the powering operations is given by $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \dots, x_D^\alpha)$. Perturbation and powering operations have a role in the simplex analogous to that of sum and scalar product in real space. The inverse of \mathbf{x} is $\mathbf{x}^{-1} = \mathcal{C}(x_1^{-1}, \dots, x_D^{-1})$ and, by analogy with standard operations in real space, $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$.

The inner product $\langle \mathbf{x}, \mathbf{x}^* \rangle_a$, defined as $\langle \mathbf{x}, \mathbf{x}^* \rangle_a = \frac{1}{D} \sum_{r < s} \left(\log \frac{x_r}{x_s} \log \frac{x_r^*}{x_s^*} \right)$, induces a distance,

known as Aitchison distance (Aitchison et al., 2000)

$$d_a^2(\mathbf{x}, \mathbf{x}^*) = \frac{1}{D} \sum_{r < s} \left(\log \frac{x_r}{x_r^*} - \log \frac{x_s}{x_s^*} \right)^2 \quad (1)$$

and a norm

$$\|\mathbf{x}\|_a^2 = \frac{1}{D} \sum_{r < s} \left(\log \frac{x_r}{x_s} \right)^2.$$

In a compositional data, the essential information is contained in the ratio between its components. This principle led Aitchison (1986) to introduce the family of log-ratio transformations from \mathcal{S}^D to the real space, then he applied classical statistical analysis to the transformed observations. In particular, the centered log-ratio transformation of \mathbf{x} , $\text{clr}(\mathbf{x})$, is obtained by

$$\text{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_j}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right], \quad (2)$$

where $g(\mathbf{x}) = \left(\prod_{j=1}^D x_j \right)^{1/D}$. The clr transformation treats all components symmetrically by dividing by the geometric mean, making easier the interpretation of the resulting values of standard multivariate techniques.

Given a full composition $\mathbf{x} \in \mathcal{S}^D$, a subcomposition $\mathbf{s} \in \mathcal{S}^C$ with C parts is a composition obtained as the closure of a subvector of C parts of the original composition \mathbf{x} .

This contribution is structured as follows: in Section 2 we apply the log-ratio methodology in the analysis of species abundance data dealing with two principal issues: firstly we discuss the transformations and dissimilarity measures for species compositional abundance data that can be used in a cluster analysis and secondly we consider principal component analysis of species abundance data in order to look for species association. In Section 3 the proposed approach is applied to a real world dataset.

2 Methodology

2.1 Transformations and dissimilarity measures for species compositional data

The analysis of species abundance data requires a suitable transformation to apply any statistical analysis which could lead to useless results if they are directly applied to the original data.

As we have seen previously, compositional data are frequent in ecology, for example when one wants to investigate the distribution of different species belonging to a biological community in a give site. We quickly review some common distances used in ecology to analyze dissimilarity between species.

- Euclidean distance (d_e)

$$d_e(\mathbf{y}_i, \mathbf{y}_s) = \sqrt{\sum_{j=1}^D (y_{ij} - y_{is})^2},$$

- Chord distance (d_c) is the Euclidean distance computed after dividing each value by the norm of the row vector, also known as cosine distance

$$d_c(\mathbf{y}_i, \mathbf{y}_s) = \sqrt{\sum_{j=1}^D \left(\frac{y_{ij}}{\|\mathbf{y}_i\|} - \frac{y_{is}}{\|\mathbf{y}_s\|} \right)^2},$$

- Distance between species profiles (d_{sp}) is the Euclidean distances between the compositional species data

$$d_{sp}(\mathbf{y}_i, \mathbf{y}_s) = \sqrt{\sum_{j=1}^D (\mathcal{C}(\mathbf{y}_i) - \mathcal{C}(\mathbf{y}_s))^2} = \sqrt{\sum_{j=1}^D (\mathbf{x}_i - \mathbf{x}_s)^2},$$

- Hellinger distance is based on the square root of ratios between components of the composition:

$$d_H(\mathbf{y}_i, \mathbf{y}_s) = \sqrt{\sum_{j=1}^D (\sqrt{\mathcal{C}(\mathbf{y}_i)} - \sqrt{\mathcal{C}(\mathbf{y}_s)})^2} = \sqrt{\sum_{j=1}^D (\sqrt{\mathbf{x}_i} - \sqrt{\mathbf{x}_s})^2},$$

- Kullback-Leibler divergence, which is generally used as a dissimilarity measure between two probability distributions, could be used also to compare two vectors of normalised counts, i.e. two compositions

$$d_{KL}(\mathbf{y}_i, \mathbf{y}_s) = \sum_{j=1}^D x_{ij} (\log x_{ij} - \log x_{sj})$$

As we deal with species compositional data, it seems to be more appropriate to introduce the Aitchison distance between two compositions $\mathbf{x}_i, \mathbf{x}_s \in \mathcal{S}^D$ (see equation 1). d_a depends only on the ratios of components and can be computed more efficiently by computing the Euclidean distance across clr-transformed vectors:

$$d_a(\mathbf{x}_i, \mathbf{x}_s) = d_e(\text{clr}(\mathbf{x}_i), \text{clr}(\mathbf{x}_s)),$$

the clr transformation is in fact symmetric and isometric.

A measure of dissimilarity d for compositional data should respect some requirements:

- scale invariance: $d(k_1 \mathbf{x}_i, k_2 \mathbf{x}_s) = d(\mathbf{x}_i, \mathbf{x}_s)$, $\forall k_1, k_2 \in \mathbb{R}_+$ and $\forall \mathbf{x}_i, \mathbf{x}_s \in \mathcal{S}^D$
- subcompositional dominance: $d(\mathbf{s}_i, \mathbf{s}_r) \leq d(\mathbf{x}_i, \mathbf{x}_r)$ for any subcompositions $\mathbf{s}_i, \mathbf{s}_r \in \mathcal{S}^C$ obtained respectively from $\mathbf{x}_i, \mathbf{x}_r \in \mathcal{S}^D$
- perturbation invariance: $d(\mathbf{x}_i \oplus \mathbf{p}, \mathbf{x}_s \oplus \mathbf{p}) = d(\mathbf{x}_i, \mathbf{x}_s)$, $\forall \mathbf{x}_i, \mathbf{x}_s, \mathbf{p} \in \mathcal{S}^D$

The scale invariance property is satisfied by those distances involving ratios between components, but subcompositional dominance and perturbation invariance are fulfilled only by the Aitchison measure. For a detailed discussion of these and other properties see Billheimer et al. (2001); Pawlowsky-Glahn and Egozcue (2001). For these reasons it seems appropriate to use the Aitchison distance in the analysis of species compositional data, in particular for clustering method used to investigate association among species.

2.2 Analysis of association among species

The proposed log-centered data transformation (equation 2) may be useful in multivariate analysis for species abundance data such as ordination methods like principal component analysis (PCA) or K-means clustering, which aggregates the sampling sites into homogeneous clusters obtained by minimizing a suitable measure of dissimilarity.

As suggested by Legendre (2005) the procedure for the identification of associated species, given a compositional species abundance table, can be synthesized as follows: firstly it is necessary to conduct an overall test of concordance using all species (Legendre, 2005), than, if the test is significant, one has to look for groups of correlated species via PCA (Jolliffe, 1986). The conventional PCA,

after standardizing the compositional row vectors of the matrix \mathbf{X} , produces a linear combination of the original species. Then the relative positions of the species in the scatterplot of the first two principal components can be interpreted in terms of their correlations (Johnson and Wichern, 2007). Another way to look for correlated species could be to apply K-means clustering to find groups of species. After that, a concordance analysis should be performed in each group to identify the species that are significantly associated. In this contest we suggest to use bayesian principal component analysis (BPCA) (Bishop, 1999) which allows us to incorporate prior knowledge about associations between species into the estimation problem. In BPCA it is possible to assess the stability of a PCA, especially when only small sample sizes are available.

In the conventional PCA the components are orthogonal linear transformations of the original variables which account for the variance in decreasing proportions. More precisely, let \mathbf{S} be the sample covariance matrix of the original data \mathbf{X} , with eigenvectors \mathbf{u}_j and corresponding eigenvalues λ_j ($j = 1, \dots, D$). Let q be the largest eigenvalues, with $q < d$, then the principal components are a reduced representation of the data set as $\mathbf{X}^* = \mathbf{U}^T(\mathbf{X} - \bar{\mathbf{x}})$ where $\mathbf{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_q\}$ and $\bar{\mathbf{x}}$ is the sample mean. A further PCA-based method allowing to study correlated species is probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999) which reformulated PCA as the maximum likelihood solution of a specific latent variable model. In the PPCA, Tipping and Bishop (1999) have re-defined the observed variable \mathbf{x} as a linear transformation of \mathbf{x}^* - with prior distribution $N(\mathbf{0}, \sigma^2 \mathbf{I}_q)$ - with additive Gaussian noise $\mathbf{x} = \mathbf{W}\mathbf{x}^* + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ where \mathbf{W} is a $D \times q$ matrix of unknown parameters, $\boldsymbol{\mu}$, is a D -dimensional vector and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_D)$. It follows that the marginal distribution of the observed variable is $N(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}_D)$. An estimate of the parameters $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ can be obtained maximizing the log-likelihood function. In the BPCA Bishop (1999) introduces a normal prior on \mathbf{W} and determined $\boldsymbol{\mu}$ and σ^2 by maximum likelihood. For further details see Bishop (1999).

Conventional PCA and PPCA do not suggest in a incontrovertible way the number of principal components to be retained. Generally the first few principal components are retained that explain a good percentage of the total variance. BPCA leads to an automatic selection of the appropriate model dimensionality, furthermore the inferences from a bayesian analysis are more informative than the conventional ones.

3 Analysis of a real dataset

To validate the proposed approach in the analysis of species association we use a species abundance dataset related to spongivore surveys that were conducted on coral reefs at 69 sites from 12 countries across the Tropical Northwestern Atlantic (Caribbean) marine province from 2008 to 2012 (Pawlik and Loh, 2017). The final dataset has 69 rows (sites) and 14 columns (species).

The 14 species are not concordant with one another (Kendall's W coefficient of concordance is equal to 0.2076, Friedman's chi-square= 197.595 with p-value < 0.001). After standardizing the clr-transformed species vectors to means of 0 and variances of 1, we performed the BPCA fitting the multivariate normal distribution described in the previous section. As the posterior distribution can not be calculated analytically, we performed an MCMC algorithm - 3 MCMC chains, each with 10,000 iterations (first 2,000 discarded) with thinning rate equal to 8 -. All statistical computations were performed using the R software (Jan Smycka and Keil, 2014; R Core Team, 2017).

Figure 1 summarizes the posterior distributions of eigenvalues and percentages of explained variance. We can see that the boxplots are pretty narrow indicating stability and robustness of the results. The first two components explain more than 50% of the total variance. The biplot could provide a meaningful representation of the structure of species and sites. Figure 2 shows BPCA biplots for 5%, 50% and 97.5% quantiles of the posterior distributions. Biplots highlight essentially two groups of species.

Also a Ward's hierarchical agglomerative clustering computed on the spearman correlation matrix

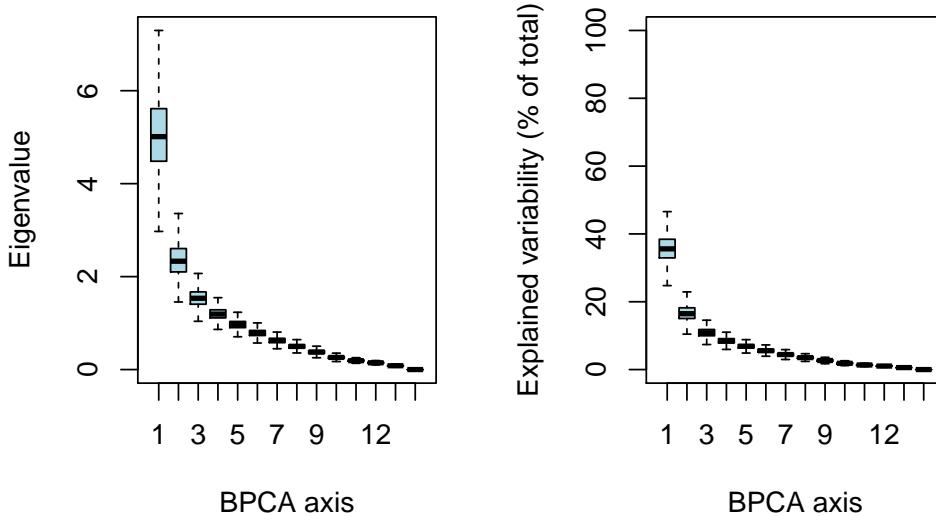


Figure 1: Boxplots of posterior distributions of eigenvalues and percentages of explained variance.

- which is interpreted as similarity indices among the species - of the clr-transformed species vectors leads to the same two groups.

Figures 3, 4 and 5 show the posterior distributions of the BPCA loadings (dotted grey lines represent posterior quantiles, 5%, 50% and 95%) which give an indication of the influence of the species on the first two bayesian principal components.

4 Discussion

In this contribution we propose to apply the log-ratio approach in the analysis of species abundance data, as a practical alternative to the usual methods applied by ecologists. Furthermore we suggest Bayesian PCA, starting from a clr-transformed species vectors, to reduce the dimensionality of the data set and to look for association among groups of species.

REFERENCES

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London, Chapman & Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
- Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology* 32(3), 271–275.
- Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association* 96, 1205–1214.

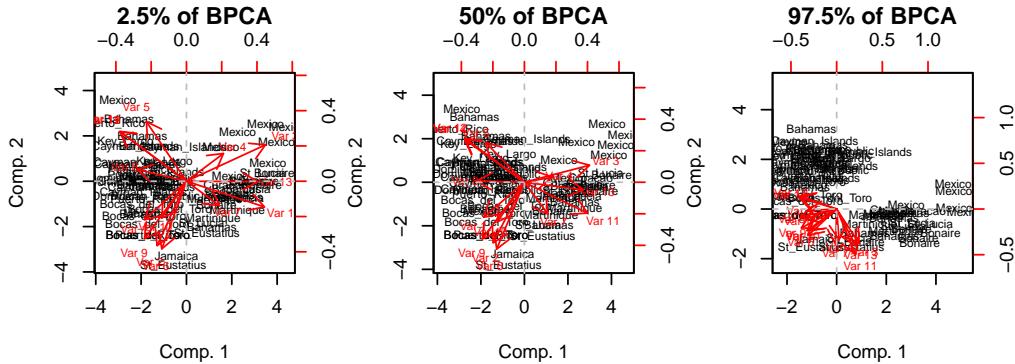


Figure 2: BPCA biplot for 5%, 50% and 97.5% quantiles of the posterior distributions. Each biplot shows the sponge species vectors projected in the space of the first two BPCA axes.

Bishop, C. M. (1999). Bayesian pca. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, Cambridge, MA, USA, pp. 382–388. MIT Press.

Buccianti, A., G. Mateu-Figueras, and V. e. Pawlowsky-Glahn (Eds.) (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Volume 264 of *Special Publications*. Geological Society, London.

Hutchinson, G. (1957). Concluding remarks. In: *Concluding Remarks: Cold Spring Harbor Symposium on Quantitative Biology*, 415–427.

Jan Smycka, J. and P. Keil (2014). *bPCA: Bayesian PCA Package*.

Johnson, R. and D. Wichern (2007). *Applied multivariate statistical analysis* (Sixth edition ed.). Prentice-Hall: London.

Jolliffe, I. (1986). *Principal component analysis*. Springer Verlag, New York.

Legendre, P. (2005). Species associations: the kendall coefficient of concordance revisited. *Journal of Agricultural, Biological and Environmental Statistics* 10(2), 226–245.

Legendre, P. and L. Legendre (1998). *Numerical Ecology* (2nd English ed.). Amsterdam: Elsevier Science BV.

Pawlak, J. and T. Loh (2017). *Spongivorous species abundance at survey sites on Caribbean coral reefs, 2008-2012 (Sponge Chem Ecology project)*. Biological and Chemical Oceanography Data Management Office (BCO-DMO). Dataset version 2017-03-01.

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

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

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roxburgh, S. and P. Chesson (1998). A new method for detecting species associations with spatially autocorrelated data. *Ecology* 79, 2180–2192.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3), 611–622.

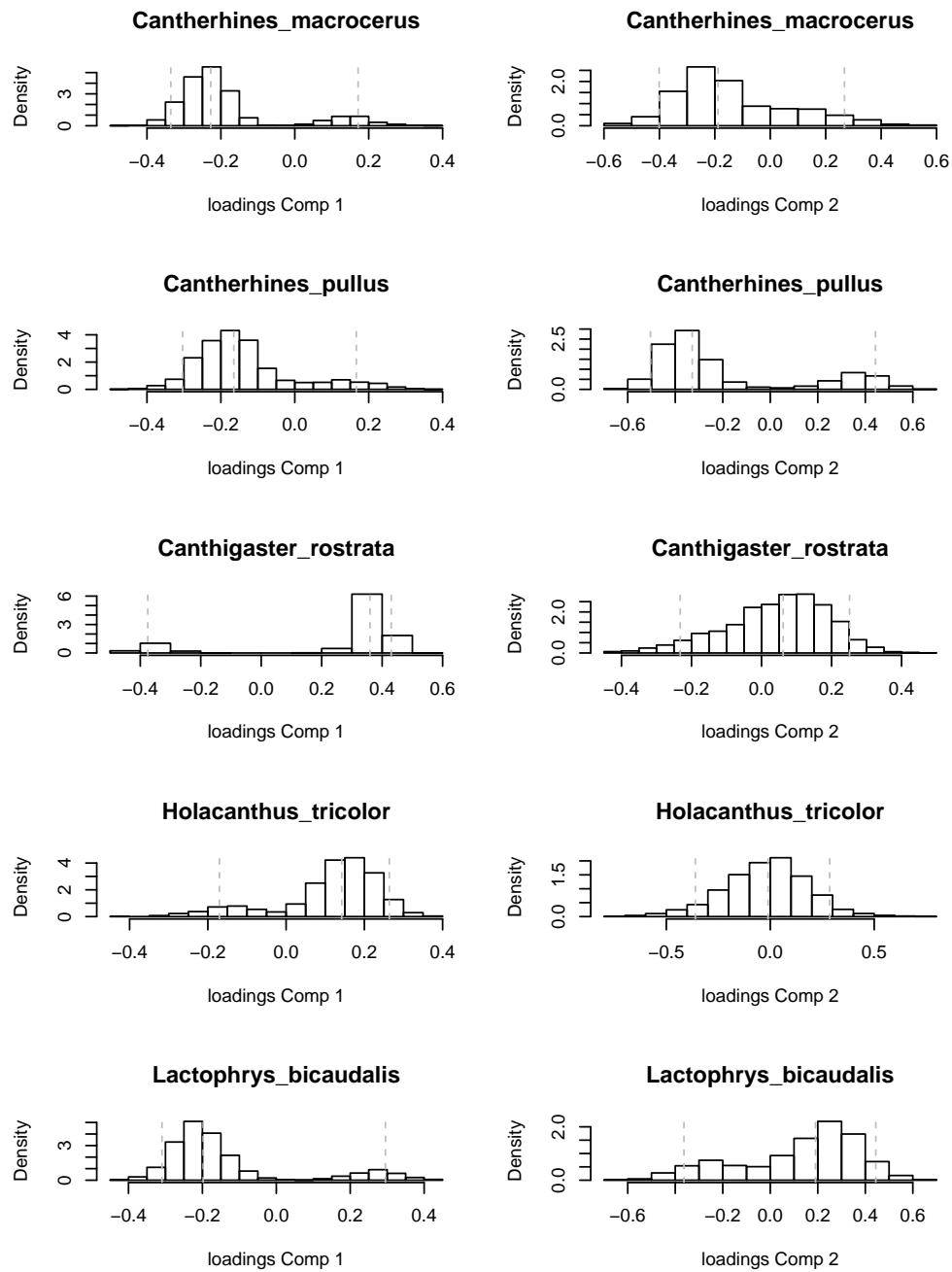


Figure 3: Posterior distributions of the BPCA loadings related to five species.

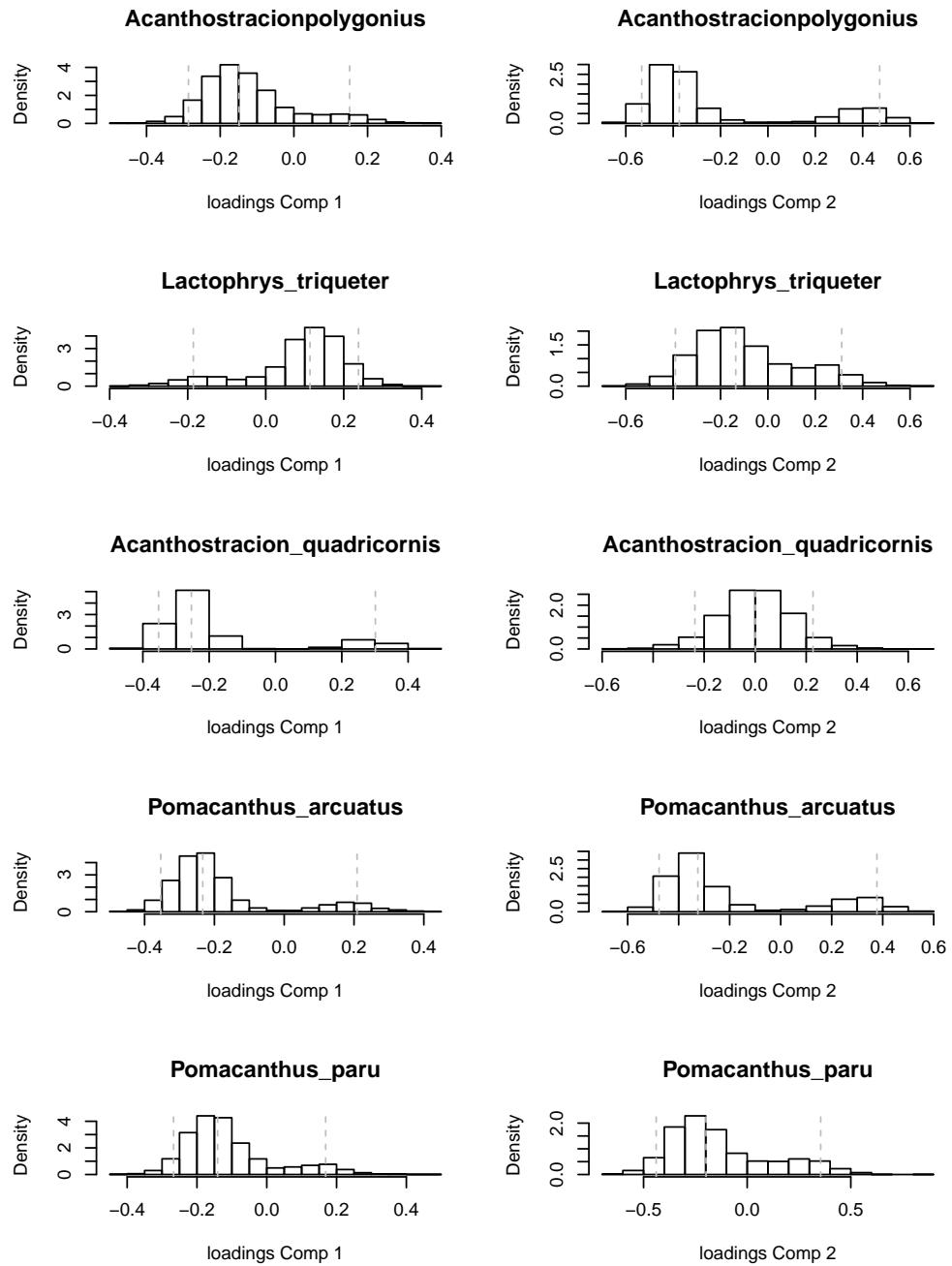


Figure 4: Posterior distributions of the BPCA loadings related to five species.

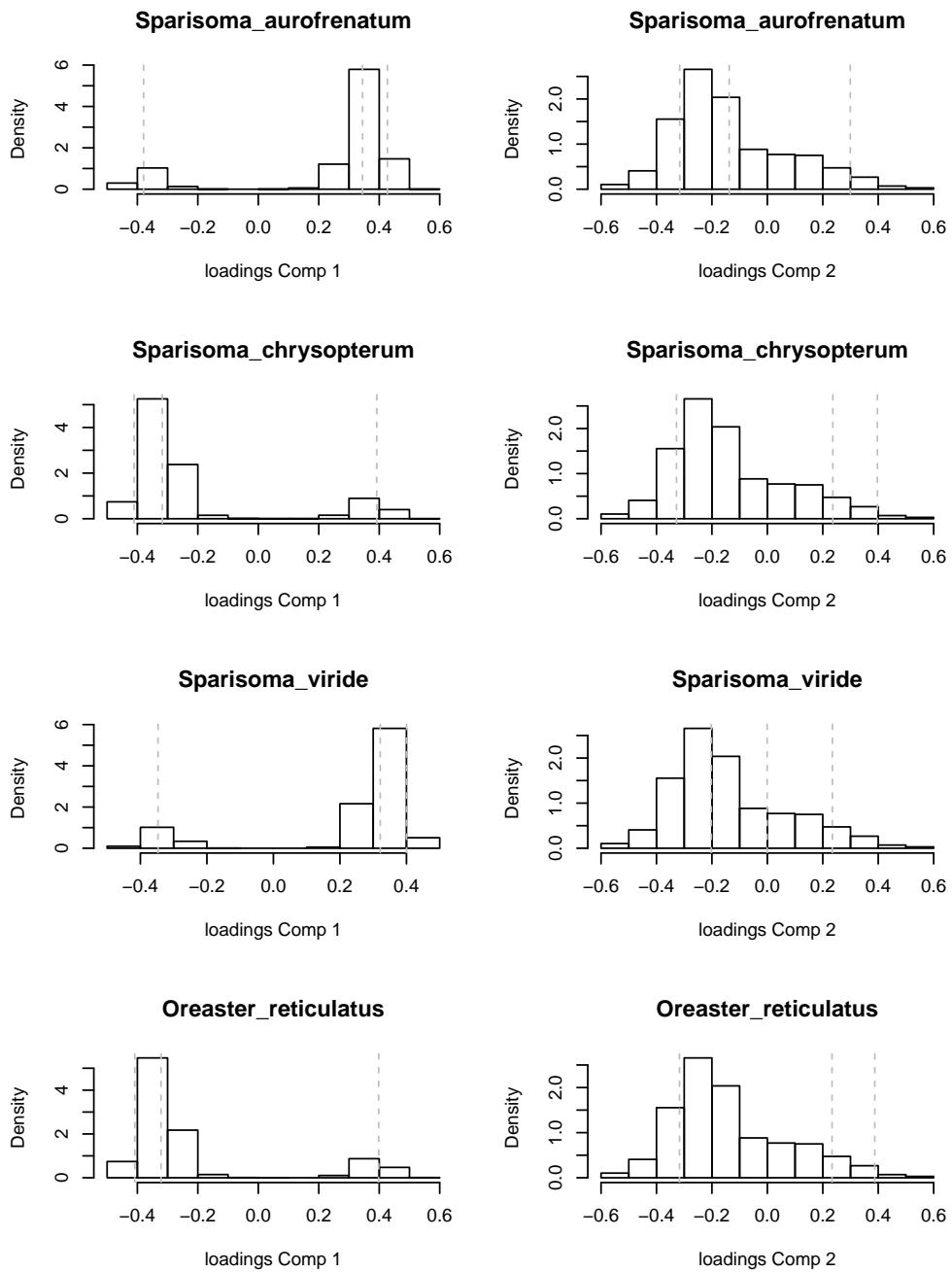


Figure 5: Posterior distributions of the BPCA loadings related to four species.

Modified Kolmogorov-Smirnov Test of Goodness of Fit

G.S. Monti¹, G. Mateu-Figueras²,

M. I. Ortego³, V. Pawlowsky-Glahn² and J. J. Egozcue³

¹Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy

gianna.monti@unimib.it

²Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain

³Department of Civil and Environmental Engineering, Technical University of Catalonia-BarcelonaTECH, Spain

Abstract

A modified version of the Kolmogorov-Smirnov (KS) test is presented as a tool to assess whether a specified, although arbitrary, probability model is unsuitable to describe the underlying distribution of a set of observations. The KS test computes distances between points of the sample cumulative distribution function and the hypothetical one as absolute differences between them, and then considering the supreme distance as test statistics. The modification here proposed consists of computing the mentioned distances as Aitchison distances of the probabilities as two part compositions.

In this contribution, we investigate by simulation the asymptotic distribution of the proposed test statistic, checking the appropriateness of the Gumbel distribution. The properties of the asymptotic distribution are studied for samples coming from generic distributions such as uniform, normal, lognormal, gamma, beta and exponential with different values of the parameters. A brief Monte Carlo investigation is made of the type I error and power of the test.

1 Introduction

The main purpose of this paper is to develop a goodness of fit test to assess the appropriateness of a certain theoretical distribution to the empirical one given a sample. We propose a modified version of the Kolmogorov-Smirnov test, which considers the largest absolute difference between two cumulative distribution functions (CDFs) as a dissimilarity. Section 2 presents the modified KS statistic which we propose in this paper. Section 3 deals with a Monte Carlo simulation study in order to investigate the asymptotic distribution of the proposed statistic and also to investigate the type I error and power of the test. Section 4 reports some comments on our proposal, which is just a first attempt to provide a log-ratio approach to a goodness of fit test, and suggests possible relationships between the sample size and the form of the test statistics.

2 The modified Kolmogorov-Smirnov statistic

Consider an independent sample, denoted $\mathbf{x} = (x_1, \dots, x_i, \dots, x_n)$, coming from a continuous random variable X . Let the hypothetical CDF be $F(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of F . We formulate the hypothesis

$$H_0 : X \sim F(\cdot|\boldsymbol{\theta}),$$

against the alternative that the random variable does not follow the claimed distribution.

H_0 can be tested using the well-known Kolmogorov-Smirnov (KS) statistic introduced by Kolmogorov (1933), which is a tool to assess whether a specified probability model is suitable to describe the underlying distribution of a set of observations. The expression of the KS statistic is

$$D_{KS} = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\},$$

is the empirical distribution function (EDF) of the sample and counts the proportion of the sample points less than or equal to x , and where $\mathbf{1}\{A\}$ is the indicator of event A. In the context of tests of fit the Kolmogorov-Smirnov statistic can be formulated as follows. Suppose that $F(x)$ is a continuous distribution, to be tested as the parent distribution of a given random sample X_1, \dots, X_n . Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the order statistics ($i = 1, \dots, n$), and consider the largest difference at the points where the EDF is greater than $F(x)$ and the largest difference at the points where the EDF is smaller than $F(x)$ as

$$\begin{aligned} D_{KS}^+ &= \max_{i=1,\dots,n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\}, \\ D_{KS}^- &= \max_{i=1,\dots,n} \left\{ F(X_{(i)}) - \frac{(i-1)}{n} \right\}, \end{aligned} \quad (1)$$

then, the Kolmogorov-Smirnov statistic is

$$D_{KS} = \max \{D_{KS}^+, D_{KS}^-\}. \quad (2)$$

The distribution of this statistic is known, even for finite samples (Birnbaum, 1952; Darling, 1957), and tables are available (Owen, 1962; D'Agostino and Stephens, 1986).

The modification of the KS test statistics in Equation (2) proposed here consists in replacing the absolute difference between the sample and the hypothetical CDF, which is a distance between real numbers, by a suitable difference for probabilities. Probabilities, like for instance i/n and $F(x_{(i)})$, can be considered as two part compositions, like for instance $(i/n, 1-i/n)$ and $(F(x_{(i)}), 1-F(x_{(i)}))$. In this case, a natural way of measuring the distance between probabilities is adopting the Aitchison distance (Aitchison, 1983; Aitchison et al., 2001). For 2-part compositions the Aitchison square distance between $\mathbf{p}_1 = (p_1, 1-p_1)$ and $\mathbf{p}_2 = (p_2, 1-p_2)$ is

$$d_a^2(\mathbf{p}_1, \mathbf{p}_2) = \left(\frac{1}{\sqrt{2}} \ln \frac{p_1}{1-p_1} - \frac{1}{\sqrt{2}} \ln \frac{p_2}{1-p_2} \right)^2,$$

which is the square difference between the logit transforms of p_1 and p_2 up to the factors $1/\sqrt{2}$. Therefore, the Aitchison distance between two probabilities, $d_a^2(p_1, p_2)$, can be identified with $d_a^2(\mathbf{p}_1, \mathbf{p}_2)$. Under this perspective, we propose to consider

$$\begin{aligned} D_a^+ &= \max_{i=1,\dots,n-1} \left\{ d_a \left(\frac{i}{n}, F(X_{(i)}) \right) \right\}, \\ D_a^- &= \max_{i=2,\dots,n} \left\{ d_a \left(F(X_{(i)}), \frac{(i-1)}{n} \right) \right\}, \end{aligned} \quad (3)$$

and the modified KS statistic is

$$D_a = \max \{D_a^+, D_a^-\}. \quad (4)$$

Note that the ranges of the index i in Equations (3) have been modified with respect to Equation (1), thus excluding infinite distances. In fact, a probability equal to 0 or equal to 1 is always at an infinite distance of other probabilities considered as compositions.

An important property of D_a as a test statistics is that it is invariant under a reversion of the orientation of the axis of the data. This means that the CDFs $F(x|\theta)$, i/n , $(i-1)/n$ can be substituted by $1-F(x|\theta)$, $1-i/n$, $1-(i-1)/n$ respectively in Equation (3) and the value of D_a does not change. This property is not fulfilled by the KS test statistics D_{KS} .

In order to complete a practical test, the distribution of the statistic in Equation (4) needs to be studied. However, the statistic (4) is the maximum of several distances. As a consequence,

the asymptotic distribution of D_a is a generalized extreme value distribution (GEVD) (Embrechts et al., 1997). GEVD applies even in the case in which there is a weak dependence between the variables from which the maximum is computed. The appropriate type of GEVD is determined by the behaviour of the upper tail of the distances. In the present case, as the support of the distances is not bounded, the Weibull type of GEVD is excluded, and as the decay of the upper tails is exponential, the asymptotic distribution of the maximum is the Gumbel distribution (GEVD with $\xi = 0$) (Appendix A).

Supported by a large number of Monte Carlo simulations (not shown here), we have observed that the D_a statistic follows reasonably well a GEVD for maxima, that is, D_A asymptotically follows a Gumbel distribution (Gumbel, 1954) and its parameters approximately depend on the sample size, as shown in Section 3.

3 Simulation results

In order to investigate the parameters of the asymptotic distribution of the D_a statistic, we have conducted a Monte Carlo study. The Monte Carlo (MC) procedure is as follows. A case consists of a single maximum likelihood estimation of the parameters of the Gumbel distribution. This case is obtained from $m = 500$ simulated samples coming from a given distribution with fixed sample size and parameters, i.e. we consider only the all-parameters-known case. For each case, the distribution model and the sample size are randomly selected. This is repeated for 1,000 different cases, thus obtaining 1,000 estimations of the scale and shape parameters of the Gumbel distribution fitted to D_a . To obtain robust results, in these simulations a 5% trimmed D_a statistic was used.

The considered reference models were: normal and lognormal distribution with different mean and scale parameters, uniform distribution with several supports, exponential distribution with several rates, and gamma and beta distribution with great variety in the parameters. The sample sizes were randomly selected ranging from $n = 5$ up to $n = 15,000$.

Two regression models, linear and quadratic, of the 1,000 MC estimates of the Gumbel parameters, μ (location) and σ (scale), were estimated against the log-size of the sample. The results are displayed in Figure 1 and in Tables 1 and 2. Using the F-test to compare both models, the

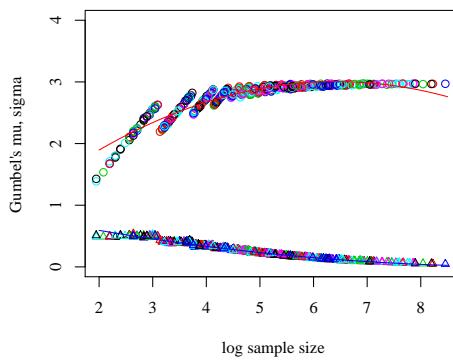


Figure 1: MC results for the Gumbel parameters. Horizontal alignment for scale parameter. Tilted alignment for location parameter. Colours indicate different distribution models. Red and blue lines represent the estimated quadratic models.

quadratic appears to be better than the linear one.

Table 1: Regression output for models (1) : $\mu = \beta_{01} + \beta_{11} \ln(n) + \varepsilon$ and (2) : $\mu = \beta_{01} + \beta_{11} \ln(n) + \beta_{21}(\ln(n))^2 + \varepsilon$.

<i>Dependent variable:</i>		
Coefficients	(1)	(2)
β_1	0.159*** (0.006)	0.729*** (0.019)
β_2		-0.057*** (0.002)
β_0	1.989*** (0.027)	0.664*** (0.047)
Observations	500	500
R ²	0.628	0.867
Adjusted R ²	0.627	0.866
Residual Std. Error	0.163 (df = 498)	0.098 (df = 497)
F Statistic	840.711*** (df = 1; 498)	1,619.445*** (df = 2; 497)
<i>Note:</i>		
*p<0.1; **p<0.05; *** p<0.01		

Table 2: Regression output for models (1) : $\sigma = \beta_{02} + \beta_{12} \ln(n) + \varepsilon$ and (2) : $\sigma = \beta_{02} + \beta_{12} \ln(n) + \beta_{22}(\ln(n))^2 + \varepsilon$.

<i>Dependent variable:</i>		
Coefficients	(1)	(2)
β_1	-0.092*** (0.001)	-0.185*** (0.004)
β_2		0.009*** (0.0004)
β_0	0.708*** (0.005)	0.926*** (0.009)
Observations	500	500
R ²	0.947	0.977
Adjusted R ²	0.947	0.977
Residual Std. Error	0.029 (df = 498)	0.019 (df = 497)
F Statistic	8,947.071*** (df = 1; 498)	10,484.430*** (df = 2; 497)
<i>Note:</i>		
*p<0.1; **p<0.05; *** p<0.01		

A brief Monte Carlo investigation was made on the size (type I error) and on the power of the test. 5,000 samples of size $n = 30, 50, 100$ were drawn from each of several distributions. The probability of rejection using the modified Kolmogorov-Smirnov test (Tables 3 and 4) was determined. Results

in Table 3 show a low conservative test in the sense that the actual significance level would be much greater than that given by the table, especially when the sample size is small.

Table 3: Probability of rejecting the null hypothesis using D_a (trim 5%) statistic with different sample sizes n . The numbers are the result of Monte Carlo simulations with 5,000 samples for each distribution.

Underlying distribution	Critical Level α	$n=30$	$n = 50$	$n = 100$
$N(0, 1)$	0.05	0.0268	0.0242	0.0310
$N(0, 1)$	0.1	0.0644	0.0664	0.0964
$Unif(1, 2)$	0.05	0.0244	0.0232	0.0362
$Unif(1, 2)$	0.1	0.0706	0.0704	0.1024
$Gamma(3, 5)$	0.05	0.0236	0.0276	0.0392
$Gamma(3, 5)$	0.1	0.0674	0.0730	0.0996
$Beta(2, 3)$	0.05	0.0258	0.0272	0.0382
$Beta(2, 3)$	0.1	0.0686	0.0732	0.0974

Table 4: Probability of rejecting hypothesis of Standard Normal distribution using D_a (trim 5%) statistic with different sample sizes n . The numbers are the result of Monte Carlo simulations with 5,000 samples for each distribution.

Underlying distribution	Critical Level α	$n=30$	$n = 50$	$n = 100$
$N(0, 4)$	0.05	0.9232	0.9876	1.0000
	0.1	0.9652	0.9962	1.0000
Student's t , 3 d.f.	0.05	0.4034	0.5360	0.8124
	0.1	0.5274	0.6628	0.8992
Exponential, rate=1	0.05	0.6122	0.8002	0.9792
	0.1	0.7276	0.8898	0.9922
Gamma, shape=rate=1	0.05	0.6160	0.8018	0.9788
	0.1	0.7398	0.8908	0.9936

4 Discussion

In this contribution we have proposed a modified version of the KS statistic. Although the test can be very useful in univariate statistics, the use in bivariate situations may be important, particularly to test goodness of fit for copulas. However, our proposal is just a first tentative to provide a log-ratio approach to a goodness of fit test, and to suggest possible relationships between the sample size and the form of the test statistics. Further studies are required to arrive at any definitive conclusions.

Acknowledgements

Research partially financially supported by the Italian Ministry of University and Research, FAR (Fondi di Ateneo per la Ricerca) 2015. The authors also gratefully acknowledge support by the Spanish Ministry of Education and Science under project ‘CODA-RETOS’ (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)) and by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project ‘COSDA’ (Ref. 2014SGR551).

REFERENCES

- Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70(1), 57–65.
- Aitchison, J., C. Barceló-Vidal, A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to letter to the editor by S. Rehder and U. Zier on Logratio analysis and compositional distance. *Mathematical Geology* 33(7), 849–860.
- Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association* 47(259), 425–441.
- Castillo, E. (1988). *Extreme Value Theory in Engineering*. Statistical Modeling and Decision Science. San Diego, Ca. (USA): Academic Press.
- Castillo, E., A. Hadi, N. Balakrishnan, and J. Sarabia (2004). *Extreme value and related models with Applications in Engineering and Science*. London, GB: Wiley. 384 p.
- D'Agostino, R. and M. Stephens (1986). *Goodness-of-fit Techniques*. Statistics, textbooks and monographs. New York (USA): Marcel Dekker, INC.
- Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics* 28(4), 823–838.
- Embrechts, P., C. Klöppelberg, and T. Mikosch (1997). *Modelling extremal values*. Springer Verlag, Berlin.
- Gumbel, E. (1954). *Statistical theory of extreme values and some practical applications*, Volume 33. U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series. (1st ed.).
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* 4, 83–91.
- Kotz, S. and S. Nadarajah (2000). *Extreme value distributions. Theory and applications*. London, GB: Imperial College Press. 185 p.
- Owen, D. (1962). *A Handbook of Statistical Tables*. Addison-Wesley, Reading, Mass.

A The Gumbel Distribution

The material in this appendix is well known and can be found in Embrechts et al. (1997) or in Castillo (1988), among others. The generalized extreme value distribution (GEVD) has the expression (Von Mises-Jenkinson formula; Embrechts et al. (1997); Castillo et al. (2004); Kotz and Nadarajah (2000))

$$F_Z(z|\mu, \sigma, \xi) = \exp \left[- \left(1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right)^{-1/\xi} \right], \quad 1 + \frac{\xi}{\sigma}(z - \mu) > 0, \quad (5)$$

where μ is a location parameter, σ is a scale parameter and ξ is a shape parameter. Parameters μ and ξ have support on the whole real line, and σ is positive. The values of the shape parameter ξ define the three families of asymptotic distribution: Weibull for $\xi < 0$, Fréchet for $\xi > 0$ and Gumbel in the limit case $\xi = 0$.

In particular, if $\xi = 0$ (Gumbel distribution), the expression (5) has the limit form

$$F_Z(z|\mu, \sigma, \xi = 0) = \exp \left[- \exp \left(- \frac{z - \mu}{\sigma} \right) \right], \quad z \in \mathbb{R}. \quad (6)$$

The corresponding probability density function is

$$f_Z(z|\mu, \sigma, \xi = 0) = \frac{1}{\sigma} \exp\left(-\frac{z-\mu}{\sigma}\right) \exp\left(-\exp\left(-\frac{z-\mu}{\sigma}\right)\right).$$

The mean and variance of the GEVD-Gumbel distribution are, respectively,

$$\text{E}(Z) = \mu + \sigma^2 \gamma, \quad \text{Var}(Z) = \frac{\pi^2}{6} \sigma^2,$$

where γ is the Euler-Mascheroni constant, $\gamma = -\int_0^\infty e^{-t} \ln t dt$. The inverse CDF sampling technique could be used to generate a random sample from a Gumbel distribution: if $U \sim \text{Unif}(0, 1)$ then $Y = F^{-1}(U) = -\ln(-\ln U)$ has the standard Gumbel distribution (with $\mu = 0$ and $\sigma = 1$). Given this result, the calculation of critical values of the test probability distribution is easy to compute.

Interpreting the impact of explanatory variables in compositional models

J. Morais^{1,2}, C. Thomas-Agnan¹, and M. Simioni³

¹Toulouse School of Economics, University of Toulouse 1 Capitole, France; *joanna.morais@live.fr*

²BVA, Boulogne-Billancourt, France

³INRA UMR 1110 MOISA, Montpellier, France

Abstract

Regression models have been developed for the case where the dependent variable is a vector of shares. Some of them, from the marketing literature, are easy to interpret but they are quite simple and can only be complexified at the expense of a large number of parameters. Other models, compositional regression models, are based on the simplicial geometry and use a log-ratio transformation of shares. They are flexible in terms of explanatory variables, but their interpretation is not straightforward, due to the link between shares. This paper combines both literatures in order to obtain a performing market-share model allowing to get relevant interpretations, which can be used for decision making in practical cases.

For example, we are interested in modeling the impact of media investments on automobile manufacturers sales. In order to take into account the competition, we model the brands market-shares as a function of brands media investments. We furthermore focus on compositional models where some explanatory variables are compositional. Two specifications are possible: in Model A, a unique coefficient is associated to each compositional explanatory variable, whereas in Model B a compositional explanatory variable is associated to component-specific and cross-effect coefficients.

Model A and Model B are estimated for our application in the B segment of the French automobile market, from 2003 to 2015. In order to enhance the interpretability of these models, we present different impact measures (marginal effects, elasticities, odds ratios) and we show that elasticities are particularly useful to isolate the impact of an explanatory variable on a particular share. We prove that elasticities can be equivalently computed from the transformed model and from the initial model. Direct and cross effects of media investments are computed for both models. Model B shows interesting non-symmetric synergies between brands.

Key words: Elasticity, odds ratio, marginal effect, compositional model, market-shares model, media investments impact.

1 Introduction

We are interested in modeling the impact of media investments on automobile manufacturer sales. We consider that the sales volume in a particular segment of the automobile market is mainly determined by the demand through the socio-economic and regulatory context. Thus, each brand tries to have “the largest share of the cake” using marketing tools, like price and media investments. The impact of media investments of brand j on its own sales cannot be assessed without taking into account the competition. Thus, we want to model the impact of media investments on market-shares, taking into account the marketing actions of competitors, directly (cross-effects) or indirectly.

In the existing literature, we found different types of models to model shares (Morais et al. (2016) for a comparison). Some of them, from the marketing or econometric literature, are perfectly adapted to model market-shares and to interpret direct and cross impacts of media investments, but the proposed models are quite simple or can only be complexified at the expense of a very large number of parameters. Other models adapted to share data are proposed, which are called compositional regression models and are based on the simplicial geometry. These mathematical models are very flexible in terms of explanatory variables and complexity (alternative-specific and cross-effect parameters), but their interpretation is not straightforward. This paper combines both literatures in order to obtain a performing market-share model allowing to get relevant and appropriate interpretations, which can be used for example to help decision making of automobile manufacturers concerning their media investments.

Here we focus on compositional models which are coming from the so called Compositional Data Analysis (CODA) literature (Pawlowsky-Glahn et al. (2015)). A composition of D components is a vector of D shares, lying in a space called the simplex, and then respecting the following constraints: components are positive and summing up to one. Compositional models are “transformation” models in the sense that they use a log-ratio transformation of shares. Transformation models have several advantages compared to other share models: they are easy to estimate (usually by OLS on coordinates) and flexible in terms of explanatory variables (they can be compositional or classical variables, with or without component-specific parameters). More specifically, we focus on models where a compositional dependent variable is explained by some compositional explanatory variables. We make a difference between two specifications of this model: in Model A, a unique coefficient is associated to each compositional explanatory variable (Wang et al. (2013)), whereas in Model B a compositional explanatory variable is associated to component-specific and cross-effect coefficients (Chen et al. (2016)).

In compositional models, the interpretation of parameters is not straightforward as all shares are linked by the summing up to one constraint. They are usually interpreted in terms of marginal effects on the transformed shares. In this paper we propose several interpretations directly linked to the shares, in terms of marginal effects, elasticities and odds ratios, in order to enhance the interpretability of these models. We show that marginal effects on shares are not well adapted to interpret these models because they depend a lot on the considered observation. Elasticities are useful to isolate the impact of an explanatory variable on a particular share as they correspond to the relative variation of a component to the relative variation of an explanatory variable, *ceteris paribus*. We show that they can be computed from the transformed model or equivalently from the model in the simplex. Other types of elasticities and odds ratios can be computed for ratios of shares, which are observation independent but they can be complicated to use in practice.

Model A and Model B are applied to an automobile market data set, where the aim is to explain the brands market-shares in a segment with brands media investments. The two models are interpreted using marginal effects, elasticities and odds ratios, and they are compared in terms of (out-of-sample) goodness-of-fit using quality measures adapted for share data.

This paper is organized as follows: the second section presents the two types of compositional models; the third section explains how to interpret them; the fourth section presents the results

of the estimation of the models for the French automobile market along with interpretations and quality measures. Finally, the last section concludes on the findings and on further directions to be investigated.

2 Compositional regression models

2.1 Definition and notations

By definition shares are compositional data: a composition is a vector of D parts of some whole which carries relative information. D -compositions lie in a space called the simplex \mathcal{S}^D .

$$\mathcal{S}^D = \left\{ \mathbf{s} = (s_1, s_2, \dots, s_D)' : s_j > 0, j = 1, \dots, D; \sum_{j=1}^D s_j = 1 \right\}$$

Compositions are subject to the following constraints: the components are positive and sum up to 1. Because of these constraints, classical regression models cannot be used directly.

The following operations are used in the simplex (Van Den Boogaart and Tolosana-Delgado (2013) for example):

- $\mathcal{C}()$ denotes the closure operation which transforms volumes into shares: $\mathcal{C}(\tilde{x}_1, \dots, \tilde{x}_D)' = \left(\frac{\tilde{x}_1}{\sum_{j=1}^D \tilde{x}_j}, \dots, \frac{\tilde{x}_D}{\sum_{j=1}^D \tilde{x}_j} \right)' = (x_1, \dots, x_D)'$ where \tilde{x} denotes the volume and x denotes the share of a variable.
- \oplus is the *perturbation operation*, corresponding to the addition operation in the simplex: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \dots, x_D y_D)'$ with $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$
- \odot is the *power transformation*, corresponding to the multiplication operation in the simplex: $\lambda \odot \mathbf{x} = \mathcal{C}(x_1^\lambda, \dots, x_D^\lambda)'$ with $\lambda \in \mathbb{R}, \mathbf{x} \in \mathcal{S}^D$
- \square is the *compositional matrix product*, corresponding to the matrix product in the simplex: $\mathbf{B} \square \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^D x_j^{b_{1j}}, \dots, \prod_{j=1}^D x_j^{b_{Dj}}\right)'$ with $\mathbf{B} \in \mathbb{R}_{D \times D}, \mathbf{x} \in \mathcal{S}^D$

2.2 Log-transformation approach

Compositional data analysis is based on the log-ratio transformation of compositions in order to obtain coordinates which can be represented in a \mathbb{R}^{D-1} Euclidean space¹. Then, classical methods suited for data in the Euclidean space, like linear regression models, can be used on coordinates. Below, terms with a “*” refer to transformed elements (in coordinates), whereas terms without “*” refer to elements in the simplex (compositions).

Several transformations are developed in the CODA literature (Pawlowsky-Glahn et al. (2015) for example). The ILR (isometric log-ratio) transformation is preferred for compositional regression models. It consists in a projection of components on an orthonormal basis of \mathcal{S}^D in order to obtain $D - 1$ orthonormal coordinates. Considering the transformation matrix $\mathbf{V}_{D \times (D-1)}$, ILR coordinates are defined as:

$$ilr(\mathbf{s}) = \mathbf{V}' \log(\mathbf{s}) = \mathbf{s}^* = (s_1^*, \dots, s_{D-1}^*)'$$

Its inverse transformation is given by: $\mathbf{S} = ilr^{-1}(\mathbf{S}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{S}^*))'$.

¹Or in \mathbb{R}^D in the case of the CLR transformation.

After inverse transformation, results of a compositional analysis are the same regardless of the chosen transformation. However, ILR is preferred for compositional regression models.

2.3 Two types of compositional models

In this section, we consider two types of models adapted to a compositional dependent variable explained by compositional explanatory variable (and potentially classical variables). The difference between the two models is about the specification of the relationship between compositional explanatory and dependent variables: in contrast with Model B, Model A does not allow for component-specific and cross effect parameters associated to a compositional explanatory variable \mathbf{X} . In this paper, we add the possibility to use classical variables Z as explanatory variables. There is no difference between Models A and B with regard to classical variables: component-specific parameters are specified. For simplicity, models are presented with a single explanatory variable of each type (compositional X and classical Z), but of course several ones can be used like in the examples presented in Section 4.

2.3.1 Model A: Compositional dependent and explanatory variables without component-specific and cross-effect parameters

Model A is presented by Wang et al. (2013). In Model A, a compositional explanatory variable is associated to a unique parameter $b \in \mathbb{R}$ (Table 1, Eq. (1)). Thus, cross-effects² are not modeled directly, but indirectly through the shares closure. Indeed, we show in Morais et al. (2016) that Model A in Equation (1) can be written in attraction form like in Equation (3). This equation contains a closure, and we can see that a change of X_l will have an indirect impact on S_j through the denominator. Moreover, the attraction form of Model A enables to see that Model A respects the IIA (independence from irrelevant alternative) property. This property means that the ratio of shares of two alternatives j and l , S_j/S_l , does not depend on characteristics of other alternatives $m \neq j, l$. Note that Equation (3) can be expressed either in terms of shares X_j or in terms of volumes \bar{X}_j thanks to the closure operation. If a classical explanatory variable Z is used in Model A, it is associated to a composition of parameters \mathbf{c}^3 .

The ILR transformation is used in order to estimate Model A [Eq. (5)]. Assuming that the transformed error terms are normal (implying that the non-transformed compositional error terms are “normal in the simplex”), we can use OLS to estimate the model.

An important feature of Model A is that compositional explanatory variables \mathbf{X} have to be of the same dimension that the compositional dependent variable \mathbf{S} , such that $\mathbf{S}, \mathbf{X} \in \mathcal{S}^D$. This model is adapted when compositions \mathbf{X} and \mathbf{S} refer to two variables associated to the same components in the same order, for example \mathbf{S} can be the composition of brands market-shares and \mathbf{X} the composition of brand media investments (where brands are in the same order in \mathbf{S} and \mathbf{X}) (see Section 4), or \mathbf{S} can be the composition of GDP from three sectors and \mathbf{X} the composition of labor force of these three sectors. Otherwise, this model makes no sense. Then, Equation (5) is estimated using $(D - 1) \times T$ observations (the number of ILR coordinates $D - 1$ times the number of observations T). Actually, this model specification is close to the specification of multinomial or market-share models (see Morais et al. (2016) for a comparison).

²We denote by cross-effect the effect of a variation of X_l on S_j , where $l \neq j$.

³It can be surprising to see that in the attraction form of Model A, the variable Z is powering the intercept c_j , but this corresponds to the term $Z_t \odot \mathbf{c}$.

2.3.2 Model B: Compositional dependent and explanatory variables with component-specific and cross-effect parameters

Model B is used by Van Den Boogaart and Tolosana-Delgado (2013) and Chen et al. (2016) for example. Using exactly the same dependent and explanatory variables as Model A [Eq. (2)], it allows each component X_l of \mathbf{X} to have a specific impact on each component S_j of \mathbf{S} . This is particularly visible in the attraction form of Model B [Eq. (4)]: instead of having a unique parameter $b \in \mathbb{R}$ associated to \mathbf{X} , we have a matrix of parameters $\mathbf{B} \in \mathbb{R}^{D_S \times D_X}$. If $D_S = D_X$ and \mathbf{S} and \mathbf{X} refer to the same components in the same order, then \mathbf{B} is a square matrix with direct effect on the diagonal and cross-effects outside of the diagonal. There is no difference between Model A and Model B for the specification of the intercept and classical explanatory variables. The same remark than for Model A can be done concerning the attraction form of Model B: Equation (4) can be expressed either in terms of shares X_j or in terms of volumes \check{X}_j thanks to the closure operation.

As in Model A, in order to estimate Model B, we transform it using the ILR transformation [Eq. (6)]. But here, $D_S - 1$ equations are estimated separately (one for each coordinate of \mathbf{S}) with T observations each. The complexity of Model B is reflected by a large number of parameters. This can be an issue if the number of observations T is too small.

Note that in Model B, $\mathbf{X} \in \mathcal{S}^{D_X}$ and $\mathbf{S} \in \mathcal{S}^{D_S}$ can have different dimensions. For example, \mathbf{S} can be the composition of GDP from three sectors and \mathbf{X} the composition of labor force for six occupation categories. In our application, $D_S = D_X$: \mathbf{S} is the composition of brands market-shares and \mathbf{X} is the composition of brand media investments (see Section 4).

One can show that Model A is a particular case of Model B where $D_S = D_X$ and where B^* is a diagonal matrix with $b^* = b$ on the diagonal and 0 otherwise, that is where only the j^{th} ILR coordinates of compositional explanatory variables are relevant to explain the j^{th} ILR coordinates of the dependent variable (see the Appendix A.1 for demonstration in the case of $D = 3$).

Table 1: Two kinds of models for compositional dependent and explanatory variables

	Model A	Model B
In compositions	$\mathbf{S}_t = \mathbf{a} \oplus b \odot \mathbf{X}_t \oplus Z_t \odot \mathbf{c} \oplus \epsilon$ (1)	$\mathbf{S}_t = \mathbf{a} \oplus \mathbf{B} \square \mathbf{X}_t \oplus Z_t \odot \mathbf{c} \oplus \epsilon$ (2)
In attraction form	$S_{jt} = \frac{a_j X_{jt}^b c_j^{Z_t} \epsilon_{jt}}{\sum_{m=1}^D a_m X_{mt}^b c_m^{Z_t} \epsilon_{mt}}$ (3)	$S_{jt} = \frac{a_j \prod_{l=1}^D X_{lt}^{b_{jl}} c_j^{Z_t} \epsilon_{jt}}{\sum_{m=1}^D a_m \prod_{l=1}^D X_{lt}^{b_{ml}} c_m^{Z_t} \epsilon_{mt}}$ (4)
In coordinates	$\mathbf{S}_t^* = \mathbf{a}^* + \mathbf{X}_t^* \cdot b + \mathbf{c}^* Z_t + \epsilon_t^*$ (5)	$\mathbf{S}_t^* = \mathbf{a}^* + \mathbf{X}_t^* \cdot \mathbf{B}_k^* + \mathbf{c}^* Z_t + \epsilon_t^*$ (6)
Component-specific parameters for X	No	Yes
Cross-effects for X	No	Yes
Dimension	D for \mathbf{S} and \mathbf{X}	D_S for \mathbf{S} ; D_X for \mathbf{X}
Nb. parameters	$(D - 1)(1 + K_Z) + K_X$	$(D_S - 1)(1 + K_Z + \sum_{k=1}^{K_X} (D_k - 1))$

\mathbf{X}_t : compositional explanatory variable; Z_t : classical explanatory variable.

D_S : number of components of \mathbf{S} ; D_X or D_k : number of components of \mathbf{X}_k .

$\mathbf{S}, \mathbf{a}, \mathbf{b}, \mathbf{X}, \epsilon \in \mathcal{S}^D$; $b, X \in \mathbb{R}$; $\mathbf{B} \in \mathbb{R}^{D_S \times D_X}$; $\mathbf{S}^*, \mathbf{a}^*, \mathbf{b}^*, \mathbf{B}^*, \mathbf{X}^*, \epsilon^*$: ILR coordinates.

ϵ : normal in the simplex distributed error terms ; ϵ^* : normal distributed error terms.

K_X and K_Z : number of compositional and classical explanatory variables ($K_X = K_Z = 1$ in the table).

\mathbb{E}^\oplus : expected value in the simplex.

3 Interpretation of compositional models

As the estimation of compositional models is performed in the coordinate space, the interpretation of the fitted parameters is difficult because parameters are linked to the log-ratio transformation of shares, not directly to the shares. It is possible to derive the coefficients in the simplex associated to shares using the inverse transformation, but their interpretation is not straightforward either.

We are going to show that relative impacts, like elasticities or odds ratios, are more natural (as is the case of the classical logistic model) than marginal effects, to interpret impacts on shares.

Table 2 compares the different measures of impact assessment of explanatory variables (compositional and classical) in Model A and Model B, which are detailed below. Note that it is not possible to measure the impact of the share of X_{lt} , but only of the corresponding volume of \check{X}_{lt} . Indeed, a share cannot increase *ceteris paribus* because it implies a change in other shares. However, we can consider a change in the volume of \check{X}_{lt} , with all other volumes $\check{X}_{mt}, m \neq l$ fixed.

3.1 Marginal effect of a component

In classical linear models, coefficients are usually interpreted in terms of marginal effects: if the explanatory variable increases by one, then the dependent variable increases by the value of the coefficient. In the case of compositional models, we prove in this paper that it is possible to compute marginal effects, but it is not straightforward. The marginal effect of the component \check{X}_{lt} (in volume) on the dependent share S_{jt} is defined as:

$$me(\mathbb{E}^{\oplus} S_{jt}, \check{X}_{lt}) = \frac{\partial \mathbb{E}^{\oplus} S_{jt}}{\partial \check{X}_{lt}} \quad (7)$$

where $\mathbb{E}^{\oplus} S_{jt}$ is the “expected value in the simplex” of S_{jt} (Morais et al. (2016)), such that $\mathbb{E}^{\oplus} S_{jt} = \frac{a_j X_{jt}^b c_j^{Z_t}}{\sum_{m=1}^D a_m X_{jt}^b c_m^{Z_t}}$ for Model A and $\mathbb{E}^{\oplus} S_{jt} = \frac{a_j \prod_{l=1}^P X_{lt}^{b_{jl}} c_j^{Z_t}}{\sum_{m=1}^D a_m \prod_{l=1}^P X_{lt}^{b_{ml}} c_m^{Z_t}}$ for Model B.

For Model B, we show that marginal effects can be computed as follows:

$$me(\mathbb{E}^{\oplus} S_{jt}, \check{X}_{lt}) = \frac{\partial \mathbb{E}^{\oplus} S_{jt}}{\partial \log \mathbb{E}^{\oplus} S_{jt}} \frac{\partial \log \mathbb{E}^{\oplus} S_{jt}}{\partial \log \check{X}_{lt}} \frac{\partial \log \check{X}_{lt}}{\partial \check{X}_{lt}} = \left(b_{jl} - \sum_{m=1}^D S_{mt} b_{ml} \right) \frac{\mathbb{E}^{\oplus} S_{jt}}{\check{X}_{lt}} \quad (8)$$

If ME_{D_S, D_X} is the matrix containing all marginal effects, we then have:

$$ME(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t) = [\mathbf{S}_{jt}] \mathbf{W}_t \mathbf{B} \odot \begin{bmatrix} \frac{1}{\check{\mathbf{X}}_{lt}} \end{bmatrix} = [\mathbf{S}_{jt}] \odot \mathbf{W}_t \mathbf{V} \mathbf{B}^* \mathbf{V}' \odot \begin{bmatrix} \frac{1}{\check{\mathbf{X}}_{lt}} \end{bmatrix} \quad (9)$$

where \odot denotes the Hadamard product here (term by term product)⁴, $[\mathbf{S}_{jt}]$ is a $D_S \times D_S$ matrix with S_{jt} on the j^{th} row, $\begin{bmatrix} \frac{1}{\check{\mathbf{X}}_{lt}} \end{bmatrix}$ is a $D_X \times D_X$ matrix with \check{X}_{lt} on the l^{th} column, \mathbf{B}^* and \mathbf{B} denote the parameters in the transformed space and in the simplex, and \mathbf{W}_t is a $D_S \times D_S$ matrix composed of diagonal terms equal to $1 - \mathbb{E}^{\oplus} S_j$ and non-diagonal terms in column j equal to $-\mathbb{E}^{\oplus} S_j$. Similar results can be found for Model A in Table 2, where \mathbf{B} is replaced by b .

This marginal effect matrix can also be computed using ILR coordinates and Jacobian matrices instead of using the attraction form of the model (Appendix A.2).

3.2 Elasticity of a dependent share relative to a component

The marginal effect $me(\mathbb{E}^{\oplus} S_{jt}, \check{X}_{lt})$ depends on all shares S_{mt} and on volumes \check{X}_{lt} . Thus, it can vary a lot across observations, and therefore it is not a good measure to summarize the impact of a

⁴Note that \odot in bold denotes the Hadamard product whereas \odot denotes the power transformation.

component \check{X}_{lt} on a share S_{jt} . We are going to show that elasticities are more natural to interpret compositional models.

The first elasticity we may want to compute is the elasticity of the share S_{jt} relative to the volume of \check{X}_{lt} . It corresponds to the relative variation of S_{jt} induced by a relative variation of 1% of \check{X}_{lt} :

$$e_{jlt} = e(\mathbb{E}^\oplus S_{jt}, \check{X}_{lt}) = \frac{\frac{\partial \mathbb{E}^\oplus S_{jt}}{\partial \check{X}_{lt}}}{\frac{\check{X}_{lt}}{S_{jt}}} = \frac{\partial \log \mathbb{E}^\oplus S_{jt}}{\partial \log \check{X}_{lt}} \quad (10)$$

These elasticities are easy to compute from the attraction form of $\mathbb{E}^\oplus S_{jt}$, in a similar way than marginal effects [Eq. (8)]. They can also be expressed in a matrix form $E(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t)$ (results are in Table 2). The relationship between marginal effects and elasticities is as follows:

$$ME(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) = [\mathbf{S}_{jt}] \odot E(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) \odot [\mathbf{1}/\check{\mathbf{X}}_{lt}]$$

These elasticities allow to isolate the impact of one \check{X} 's component on one S 's component which is very useful. $e(\mathbb{E}^\oplus S_{jt}, \check{X}_{lt})$ depends on observations but only through the S_{mt} , not through \check{X}_{lt} . Then, if shares are not varying too much, as it is the case in our example (see Section 4), they can be a good measure of impact.

As for marginal effects, the elasticity matrix can also be computed from ILR coordinates (Appendix A.2).

Note that for a small relative change of \check{X}_{lt} equal to $h = \frac{\Delta \check{X}_{lt}}{\check{X}_{lt}}$, a first order Taylor approximation of the share denoted S'_{jt} is:

$$S'_{jt} = S_{jt}(1 + he_{jlt}) \quad (11)$$

We can verify that, for a small h , the S'_{mt} do belong to the simplex (they are summing up to one because $\sum_{m=1}^D \mathbb{E}^\oplus S_{mt} e_{jlt} = 0$, see proof in the Appendix A.3).

Moreover, we can link these elasticities to simplicial derivatives⁵ (i.e. derivatives in the simplex). Indeed, the simplicial derivative of the composition \mathbf{S} with respect to the log of a particular component \check{X}_l is defined as follows:

$$e_{lt}^\oplus = \frac{\partial^\oplus \mathbb{E}^\oplus \mathbf{S}_t}{\partial^\oplus \log \check{X}_{lt}} = \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^\oplus \mathbf{S}_t}{\partial \log \check{X}_{lt}}\right)\right) = \mathcal{C}(\exp(e_{1lt}), \dots, \exp(e_{Dlt})) \quad (12)$$

For a small relative change of \check{X}_l equal to $h = \frac{\Delta \check{X}_{lt}}{\check{X}_{lt}}$, another first order Taylor approximation of share denoted \mathbf{S}_t'' is⁶:

$$\mathbf{S}_t'' = \mathbf{S}_t \oplus h \odot e_{lt}^\oplus = \mathcal{C}(S_{1t} \exp(h e_{1lt}), \dots, S_{Dt} \exp(h e_{Dlt})) \quad (13)$$

Note that when $h \rightarrow 0$, $\exp(h e_{jlt}) \simeq 1 + h e_{jlt}$, so that:

$$\mathbf{S}_t'' \simeq \mathcal{C}(S_{1t}(1 + h e_{1lt}), \dots, S_{Dt}(1 + h e_{Dlt})) = \mathcal{C}(S'_{1t}, \dots, S'_{Dt}) = (S'_{1t}, \dots, S'_{Dt}) \quad (14)$$

where S'_{jt} are computed in Equation (11) and \mathbf{S}_t'' in Equation (13). The last equality of Equation (14) is justified by the fact that $\sum_{m=1}^D \mathbb{E}^\oplus S_{mt} = 1$ and $\sum_{m=1}^D \mathbb{E}^\oplus S_{mt} e_{jlt} = 0$.

3.3 Elasticity and odds ratio of a ratio of dependent shares relative to a component

In order to avoid being observation dependent, other measures can be computed for interpreting Models A and B. However, they are concerning ratios of shares, not directly a single share. Then, they can be complicated to interpret in practical cases.

⁵See Equation (9.9), p.183, in Pawlowsky-Glahn et al. (2015).

⁶See Equation (12.13), p.168, in Pawlowsky-Glahn and Buccianti (2011).

Elasticity of a ratio of dependent shares As compositional data analysis is based on a log ratio approach, elasticities of ratios are easy to compute. We can be interested in the elasticity of a ratio of shares (or volumes) $\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}$ relative to an infinitesimal change in the volume of \check{X}_{lt} .

$$e(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}) = \frac{\partial \log(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't})}{\partial \log \check{X}_{lt}} \quad (15)$$

We see in Table 2 that the result is constant across observations because it only depends on parameters. Note here that Model A respects the IIA (Independence from Irrelevant Alternatives) property, meaning that the ratio of two shares $\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}$ only depends on the corresponding components j and j' of $\check{\mathbf{X}}$. Then, $e(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}) = 0$ if $l \neq j, j'$. Moreover, the elasticity of the ratio between the share j and the share j' relative to a change in \check{X}_{jt} is the same for all considered shares j' . This is a lack of flexibility of Model A, because it implies that an increase of \check{X}_{jt} will reduce proportionally all other shares. Model B does not satisfy the IIA property, and then this model is able to take into account possible synergies between brands.

Odds ratio of a ratio of dependent shares Another type of interpretation which can be used for shares is the odds ratio. The advantage of this measure is that it is a measure of impact of a discrete change, as opposed to infinitesimal change, of \check{X}_l (\check{X}_l is increased by $\Delta \times 100\%$ between situations $t = t1$ and $t = t2$) on the ratio $\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}$. The empirical odds ratio for a couple of shares $\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}$ relative to \check{X}_{lt} is given by:

$$OR(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}, \Delta) = \frac{(\mathbb{E}^{\oplus} S_{j,t2}/\mathbb{E}^{\oplus} S_{j',t2})|\check{X}_{l,t2}}{(\mathbb{E}^{\oplus} S_{j,t1}/\mathbb{E}^{\oplus} S_{j',t1})|\check{X}_{l,t1}} \quad (16)$$

where $\check{X}_{l,t2} = (1 + \Delta)\check{X}_{l,t1}$ and $\Delta \geq 0$.

Remark: $e(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt})$ and $OR(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}, \Delta)$ are more or less measuring the same thing differently, if Δ is small:

$$\begin{aligned} e(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}) &\simeq \frac{(\mathbb{E}^{\oplus} S_{jt2}/\mathbb{E}^{\oplus} S_{j't2}) - (\mathbb{E}^{\oplus} S_{jt1}/\mathbb{E}^{\oplus} S_{j't1})}{(\mathbb{E}^{\oplus} S_{jt1}/\mathbb{E}^{\oplus} S_{j't1})} / \frac{\check{X}_{lt2} - \check{X}_{lt1}}{\check{X}_{lt1}} \\ &\simeq \frac{OR(\mathbb{E}^{\oplus} S_{jt}/\mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}, \Delta) - 1}{(\check{X}_{lt2} - \check{X}_{lt1})/(\check{X}_{lt1})} \end{aligned}$$

3.4 Elasticity of a particular ratio of dependent shares relative to a particular ratio of components

Usually, compositional models are interpreted directly on coordinates. Thus, it is advised to choose an appropriate ILR transformation in order to have ILR coordinates which make sense for the considered application, using sequential binary partition for example (Hron et al. (2012)). But, previously the interpretation was made in terms of marginal effects on ILR coordinates, that is marginal effects on a particular log ratio of shares. We show here that we can go a step further and make an interpretation in terms of elasticity for the ratio of shares directly.

Chen et al. (2016) interpret in the case of Model B the impact of the ratio $X_l/g(X_{-l}) = \check{X}_l/g(\check{X}_{-l})$ on the ratio $\mathbb{E}^{\oplus} S_j/g(\mathbb{E}^{\oplus} S_{-j}) = \mathbb{E}^{\oplus} \check{S}_j/g(\mathbb{E}^{\oplus} \check{S}_{-j})$ (ratios on shares or volumes are equivalent), which is the ratio of a particular share (or volume) S_j over the geometric average of other shares (or volumes). The adapted ILR transformation is the following:

$$ilr(\mathbf{X})_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i}{(\prod_{j=1+i}^D x_j)^{1/(D-i)}}, \quad i = 1, \dots, D-1$$

With this transformation, the first expected coordinate of \mathbf{S} in Model A, is equal to:

$$\mathbb{E}ilr(\mathbf{S})_1 = \sqrt{\frac{D-1}{D}} \log \frac{\mathbb{E}^\oplus S_{1t}}{g(\mathbb{E}^\oplus S_{-1t})} = a_1^* + b^* \sqrt{\frac{D-1}{D}} \log \frac{\check{X}_{1t}}{g(\check{X}_{-1t})} + c_1^* Z_t$$

In Model B, the first expected coordinate of \mathbf{S} is equal to:

$$\mathbb{E}ilr(\mathbf{S})_1 = \sqrt{\frac{D_S-1}{D_S}} \log \frac{\mathbb{E}^\oplus S_{1t}}{g(\mathbb{E}^\oplus S_{-1t})} = a_1^* + b_{11}^{*(j,l)} \sqrt{\frac{D_X-1}{D_X}} \log \frac{\check{X}_{1t}}{g(\check{X}_{-1t})} + b_{12}^{*(j,l)} \sqrt{\frac{D_X-2}{D_X-1}} \log \frac{\check{X}_{2t}}{g(\check{X}_{-1-2t})} + \dots$$

In order to interpret their model, Chen et al. (2016) compute the marginal effect of $ilr(X)_1^{(l)}$ on $ilr(S)_1^{(j)}$:

$$me(\mathbb{E}ilr(S)_1^{(j)}, ilr(\check{X})_1^{(l)}) = \frac{\partial \sqrt{\frac{D_S-1}{D_S}} \log(\mathbb{E}^\oplus S_{jt}/g(\mathbb{E}^\oplus S_{-jt}))}{\partial \sqrt{\frac{D_X-1}{D_X}} \log(\check{X}_{lt}/g(\check{X}_{-lt}))} = b_{11}^{*(j,l)}$$

such that an increase of one unit of $ilr(\check{X})_1^{(l)}$ implies an increase of $b_{11}^{*(j,l)}$ units of $\mathbb{E}ilr(S)_1^{(j)}$ ⁷.

Note that this is only true if $\sqrt{\frac{D_X-1}{D_X}} \log(\check{X}_{lt}/g(\check{X}_{-lt}))$ moves because \check{X}_{1t} moves while other \check{X}_{jt} remain constant. Otherwise, other ILR coordinates in the right part of the equation are moving and the marginal effect should take it into account. However, for Model A, we do not have this problem because other ILR coordinates of \mathbf{X} are not used.

We show that this is equivalent to compute the following elasticity (multiplying by a factor if $D_S \neq D_X$):

$$e \left(\frac{\mathbb{E}^\oplus S_{jt}}{g(\mathbb{E}^\oplus S_{-jt})}, \check{X}_{lt} \right) = \frac{\partial \log(\mathbb{E}^\oplus S_{jt}/g(\mathbb{E}^\oplus S_{-jt}))}{\partial \log \check{X}_{lt}} = \sqrt{\frac{(D_X-1)/D_X}{(D_S-1)/D_S}} b_{11}^{*(j,l)}$$

Thus, instead of saying that when $ilr(\check{X})_1^{(l)}$ increases by 1 unit, $\mathbb{E}ilr(S)_1^{(j)}$ increases by $b_{11}^{*(j,l)}$ units, one can say that when \check{X}_{lt} increases by 1%, $\mathbb{E}^\oplus S_{jt}/g(\mathbb{E}^\oplus S_{-jt})$ increases by $b_{11}^{*(j,l)}\%$ (in the case where $D_S = D_X$). Note that this $b_{11}^{*(j,l)}$ will be different for each permutation (i.e. each couple j, l). Chen et al. (2016) show how one can determine in one step the first coefficient of $B^{*(j,l)}$, the $b_{11}^{*(j,l)}$ which is used to compute the above elasticity, for all possible permutations without fitting several times the model.

3.5 Elasticities and odds ratios relative to a classical variable

The same kind of interpretations can be done for classical variables Z , as presented in Table 2, except for the elasticity including the geometrical mean.

Indeed, this would allow to measure the marginal effect (not the elasticity) of Z_t over $\sqrt{\frac{D_S-1}{D_S}} \log \frac{S_{1t}}{g(S_{-1t})}$. This marginal effect would be equal to c_1^* for Model A and Model B, but this kind of interpretation is not useful to understand the impact of Z on the final shares. Thus, we do not show this measure in Table 2.

Note that in practice, elasticities and other measures depending on $\mathbb{E}^\oplus S_{jt}$ are estimated using the observed shares S_{jt} , not the fitted shares \widehat{S}_{jt} .

⁷ $ilr(S)_1^{(j)}$ denotes the first ILR coordinate of \mathbf{S} where S_j is in the first position; $ilr(\check{X})_1^{(l)}$ denotes the first ILR coordinate of $\check{\mathbf{X}}$ where \check{X}_l is in the first position.

Table 2: Measures of impact assessment for Model A and Model B

Var	Measure	Effect	Model A	Model B
X	$me(S_{jt}, \check{X}_{lt})$	Direct	$b(1 - S_{jt}) \frac{S_{jt}}{\check{X}_{lt}}$	$(b_{jl} - \sum_{m=1}^D S_{mt} b_{ml}) \frac{S_{jt}}{\check{X}_{lt}}$
		Indirect	$(-bS_{lt}) \frac{S_{jt}}{\check{X}_{lt}}$	
	$ME(\mathbf{S}_t, \check{\mathbf{X}}_t)$	Matrix	$[\mathbf{S}_{jt}] \odot \mathbf{W}_t b \odot [1/\check{\mathbf{X}}_{lt}]$	$[\mathbf{S}_{jt}] \odot \mathbf{W}_t \mathbf{B} \odot [1/\check{\mathbf{X}}_{lt}]$
	$e(S_{jt}, \check{X}_{lt})$	Direct	$b(1 - S_{jt})$	$(b_{jl} - \sum_{m=1}^D S_{mt} b_{ml})$
		Indirect	$-bS_{lt}$	
	$E(\mathbf{S}_t, \check{\mathbf{X}}_t)$	Matrix	$\mathbf{W}_t b$	$\mathbf{W}_t \mathbf{B}$
	$e\left(\frac{S_{jt}}{S_{j't}}, \check{X}_{lt}\right)$	Direct	b	$(b_{jl} - b_{j'l})$
		Indirect	0	
	$OR\left(\frac{S_{jt}}{S_{j't}}, \check{X}_{lt}, \Delta\right)$	Direct	$(1 + \Delta)^b$	$(1 + \Delta)^{(b_{jl} - b_{j'l})}$
		Indirect	0	
Z	$e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{X}_{lt}\right)$	Direct	b	$b_{11}^{*(j,l)} \sqrt{\frac{D_X - 1}{D_X}} / \sqrt{\frac{D_S - 1}{D_S}}$
		Indirect	0	
	$me(S_{jt}, Z_t)$		$(\log c_j - \sum_{m=1}^D S_{mt} \log c_m) S_{jt}$	
	$ME(\mathbf{S}_t, Z_t)$	Vector	$[\mathbf{S}_{jt}] \odot \mathbf{W}_t \log \mathbf{c}$	
	$e(S_{jt}, Z_t)$		$(\log c_j - \sum_{m=1}^D S_{mt} \log c_m) Z_t$	
	$E(\mathbf{S}_t, Z_t)$	Vector	$\mathbf{W}_t \log \mathbf{c} \cdot Z_t$	
	$e\left(\frac{S_{jt}}{S_{j't}}, Z_t\right)$		$\log(c_j/c_{j'}) Z_t$	
	$OR\left(\frac{S_{jt}}{S_{j't}}, Z_t, \Delta\right)$		$(c_j/c_{j'})^{\Delta Z_t}$	

In this table, $\mathbb{E}^\oplus S_{jt}$ is denoted by S_{jt} to shorten notations, and \odot denotes the Hadamard product.

Moreover, these measures are estimated using observed shares S_{jt} in practice, not fitted shares.

Direct effect when $l = j$; indirect effect when $l \neq j$.

\mathbf{W}_t contains $1 - S_{it}$ on the diagonal and $-S_{it}$ otherwise.

4 Impact of media investments on brands market-shares

In Europe, the automobile market is usually segmented in 5 segments, from A to E, according to the size of the vehicle chassis. Within each segment, one can suppose that consumers intending to buy new cars make their choice between brands⁸ according to the price and the “image” of the brand. The image of the brand is supposed to reflect the notion of quality and reliability of the brand. Car manufacturers spend millions of euros in media investments to enhance their image, giving rise to the following question: do the media investments have an impact on brands market-shares⁹?

In order to answer this question in the present paper, we model brands market-shares of the B segment of the French automobile market¹⁰ as a function of brand media investments (in TV, radio, press, outdoor, internet and cinema), of brand average catalogue price and of a scrapping incentive dummy variable. In a further work, we consider modeling other segments, and differentiate media investments according to channels.

In this paper, three brands are highlighted (Renault, Peugeot, Citroen, the leaders of the B segment) while other brands of the B segment are aggregated in a category “Others” (Fig. 1). The media investments are the sum of TV, radio, press, outdoor, internet and cinema investments in euros by brands for their vehicles in the B segment (Fig. 1). They do not include advertising budget for the brand itself. Actually we use the media investments of one, two and three months before the purchase time (at time $t - 1, t - 2, t - 3$) as explanatory variables. The average brand price (average of catalogue prices weighted by corresponding sales at the vehicle level) is also used as an explanatory variable (Fig. 1). It does not include potential promotions made in the car

⁸Inside a segment, a brand generally supplies only one main vehicle. Thus, we can consider that the alternatives for a consumer inside a particular segment coincide with the available brands in this segment.

⁹We decide to ask the question in terms of market-shares instead of in terms of sales volumes because one can suppose that at time t , brands have to share a market for which the size is mainly determined by the demand.

¹⁰The B segment is the most important segment in terms of sales in France (around 40% of new passenger car sales).

dealership at the time of purchase. Even if they do not vary a lot across time, prices are used to position brands within the segment. We also control for scrapping incentive periods. The corresponding dummy variable is a “classical” variable (not compositional) and varies across time only, not across brands.

Model A and Model B can be considered in this framework: Model A considers that the effect of media investments and price are the same for all brands whereas Model B implies cross-effects and brand-specific impacts of media investments and price on market-shares.

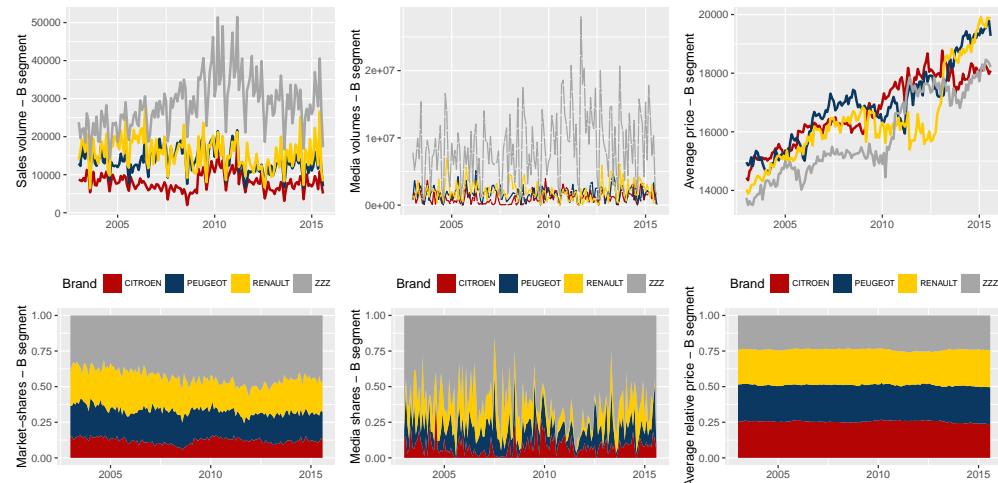


Figure 1: Sales, media and average price of brands, in volume and in share, in the B segment

This section presents the results of this application. We interpret the two models A and B in terms of elasticities and odds ratios of shares, and we compare them in terms of goodness-of-fit measures.

4.1 Non brand-specific impact of media investments (Model A)

Model In the case where it is assumed that brand media investments and brand prices have the same effect for all brands, the following equations correspond to the model in the simplex and the attraction formulation of the model:

$$\begin{aligned} \mathbf{S}_t &= \mathbf{a} \bigoplus_{\tau=1}^3 b_\tau \odot \mathbf{M}_{t-\tau} + b_P \odot \mathbf{P}_t + SI_t \odot \mathbf{c} + \boldsymbol{\varepsilon}_t \\ \Leftrightarrow S_{jt} &= \frac{a_j \cdot \prod_{\tau=1}^3 M_{t-\tau,j}^{b_\tau} \cdot P_{t,j}^{b_P} \cdot c_j^{SI} \cdot \varepsilon_{jt}}{\sum_{m=1}^4 a_m \cdot \prod_{\tau=1}^3 M_{t-\tau,m}^{b_\tau} \cdot P_{t,m}^{b_P} \cdot c_m^{SI} \cdot \varepsilon_{mt}} \end{aligned}$$

where $\mathbf{S}, \mathbf{M}_{t-\tau}, \mathbf{P} \in \mathcal{S}^4$ are the compositions of brand sales, of brand media investments at time $t-1, t-2$ and $t-3$, and of brand prices. $b_\tau, b_P \in \mathbb{R}$ are the parameters associated to compositional explanatory variables and $\mathbf{c} \in \mathcal{S}^4$ is a composition of parameters associated to the dummy variable SI (scrapping incentive).

The ILR transformed version of the model is:

$$\begin{aligned} \mathbf{S}_t^* &= \mathbf{a}^* + \sum_{\tau=1}^3 b_\tau \mathbf{M}_{t-\tau}^* + b_P \mathbf{P}_t^* + \mathbf{c}^* S I_t + \boldsymbol{\varepsilon}_t^* \\ \Leftrightarrow S_{jt}^* &= a_j^* + \sum_{\tau=1}^3 b_\tau^* M_{j,t-\tau}^* + b_P^* P_{jt}^* + c_j^* S I_t + \varepsilon_{jt}^* \quad \text{for } j = 1, 2, 3 \end{aligned}$$

where $\boldsymbol{\varepsilon}^*$ is supposed to be a Gaussian distributed error term. The balance matrix used for the ILR transformation is the default matrix in the R software:

$$V_{ILR,4} = \begin{bmatrix} -\sqrt{1/2} & -\sqrt{1/6} & -\sqrt{1/12} \\ \sqrt{1/2} & -\sqrt{1/6} & -\sqrt{1/12} \\ 0 & \sqrt{2/3} & -\sqrt{1/12} \\ 0 & 0 & \sqrt{3/4} \end{bmatrix} \quad (17)$$

Results All explanatory variables are significant at 0.1% according to the analysis of variance (ANOVA). Figure 2 compares observed and fitted shares. It confirms that the model succeeds in fitting the main trends of brands market-shares. However, the model underestimates the market-share of “Others” at the beginning of the period, and overestimates it at the end.

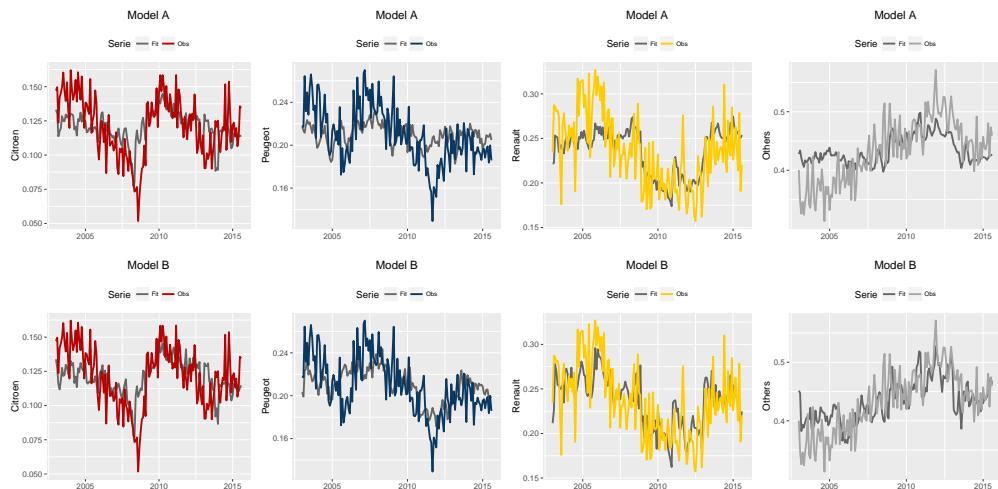


Figure 2: Observed (color) and predicted (grey) brands market-shares

The parameters estimated on the ILR transformed model are presented in Table 3. The corresponding parameters for the model in the simplex are in Table 4. We remark that the coefficient associated to the price is positive, which can be surprising, but price here is correlated with the image of quality of the brand, which is very important for the customer who buy a durable and expensive good like a car.

4.2 Brand-specific impact of media investments (Model B)

Model Now, let us look at a different specification of the model (dependent and explanatory variables are the same as in Model A) where brand-specific coefficients are assumed and cross-

Table 3: Estimated parameters on ILR coordinates - Model A

	Estimate	Std. Error	t value	Pr(> t)
a_1^*	0.3439	0.0151	22.84	0.0000***
a_2^*	0.3363	0.0159	21.19	0.0000***
a_3^*	0.6620	0.0263	25.14	0.0000***
b_1	0.0267	0.0071	3.79	0.0002***
b_2	0.0241	0.0062	3.90	0.0001***
b_3	0.0264	0.0062	4.26	0.0000***
b_P	1.2217	0.2313	5.28	0.0000***
c_1^*	-0.0241	0.0338	-0.71	0.4758
c_2^*	-0.1690	0.0334	-5.05	0.0000***
c_3^*	0.1292	0.0336	3.84	0.0001***
Nb param.	10			
Signif. codes:	0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’, 0.1 ‘ ’, 1			

Table 4: Estimated parameters in the simplex - Model A

	S_1 (Citroen)	S_2 (Peugeot)	S_3 (Renault)	S_4 (Others)
(Intercept)	0.1300	0.2114	0.2502	0.4084
M_{t-1}		0.0267		
M_{t-2}		0.0241		
M_{t-3}		0.0264		
P_t		1.2217		
SI	0.2610	0.2523	0.2086	0.2780

effects are directly modeled. It corresponds to the following model:

$$\begin{aligned} \mathbf{S}_t &= \mathbf{a} \bigoplus_{\tau=1}^3 \mathbf{B}_\tau \square \mathbf{M}_{t-\tau} \oplus \mathbf{B}_P \square \mathbf{P}_t \oplus SI_t \odot \mathbf{c} \oplus \boldsymbol{\varepsilon}_t \\ \Leftrightarrow S_{jt} &= \frac{a_j \cdot \prod_{\tau=1}^3 \prod_{l=1}^4 M_{t-\tau,l}^{b_{\tau,jl}} \cdot \prod_{l=1}^4 P_{t,l}^{b_{P,jl}} \cdot c_j^{SI} \cdot \varepsilon_{jt}}{\sum_{m=1}^4 a_m \cdot \prod_{\tau=1}^3 \prod_{l=1}^4 M_{t-\tau,l}^{b_{\tau,ml}} \cdot \prod_{l=1}^4 P_{t,l}^{b_{P,ml}} \cdot c_m^{SI} \cdot \varepsilon_{mt}} \end{aligned}$$

where $\mathbf{B}_\tau, \mathbf{B}_P \in \mathbb{R}^{D \times D}$ are the matrices of parameters associated to compositional explanatory variables.

The corresponding ILR transformed model is:

$$\begin{aligned} \mathbf{S}_t^* &= \mathbf{a}^* + \sum_{\tau=1}^3 \mathbf{B}_\tau^* \mathbf{M}_{t-\tau}^* + \mathbf{B}_P^* \mathbf{P}_t^* + \mathbf{c}^* SI_t + \boldsymbol{\varepsilon}_t^* \\ \Leftrightarrow S_{jt}^* &= a_j^* + \sum_{\tau=1}^3 \sum_{l=1}^3 b_{\tau,jl}^* M_{t-\tau,l}^* + \sum_{l=1}^3 b_{P,jl}^* P_{t,l}^* + c_j^* SI_t + \varepsilon_{jt}^* \quad \text{for } j = 1, 2, 3 \end{aligned}$$

where $\boldsymbol{\varepsilon}^*$ is supposed to be a Gaussian distributed error term. The same balance matrix $V_{ILR,4}$ is used.

Results All variables of the model are significant at 0.1% according to the ANOVA, except the price which is significant at 1%. According to Figure 2, Model B seems to fit better than Model A (see Section 4.3 for associated quality measures). The estimated parameters of the models are given in Table 5 and Table 6.

Table 5: Estimated parameters on ILR coordinates - Model B

	S_1^* (Peu. vs Cit.)	S_2^* (Reu. vs Cit.,Peu.)	S_3^* (Oth. vs Cit.,Peu.,Reu.)
(Intercept)	0.3686***	0.3637***	0.6940***
$M_{t-1,1}^*$	0.0193.	-0.0052	0.0081
$M_{t-1,2}^*$	0.0162	0.0319*	-0.0245
$M_{t-1,3}^*$	-0.0069	0.0009	0.0279
$M_{t-2,1}^*$	0.0208.	-0.0093	0.0205.
$M_{t-2,2}^*$	0.0151	0.0361**	-0.0259.
$M_{t-2,3}^*$	-0.0197	-0.0338	0.0278
$M_{t-3,1}^*$	0.0289**	-0.0115	0.0278*
$M_{t-3,2}^*$	0.0104	0.0206*	-0.0274.
$M_{t-3,3}^*$	-0.0114	0.0064	0.0323.
P_1^*	0.8854.	-0.5981	1.9138***
P_2^*	0.0151	0.2615	0.6509
P_3^*	-0.6442	-0.3729	2.4717***
SI^*	-0.0394	-0.2088***	0.2070***
Adjusted R2	0.3353	0.3255	0.3269
Nb param.	42		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 6: Estimated parameters of M_{t-1} in the simplex - Model B

	S_1 (Citroen)	S_2 (Peugeot)	S_3 (Renault)	S_4 (Others)
$M_{t-1,1}$	0.0179	-0.0079	-0.0067	-0.0032
$M_{t-1,2}$	-0.0016	0.0111	-0.0161	0.0066
$M_{t-1,3}$	-0.0132	0.0084	0.0292	-0.0243
$M_{t-1,4}$	-0.0030	-0.0115	-0.0064	0.0209

4.3 Interpretation of models A and B

Marginal effect of media investments We calculate the marginal effects of media investments at time $t-1$ on market-shares at time t . The average marginal effects are reported in Table 7. They are quite consistent between Model A and Model B, with positive direct marginal effects and negative cross marginal effects. However, these measures are not really adapted to summarize an impact as they fluctuate a lot across time, as we can see in Figure 3 (marginal effects can be larger than 6e-08 but we voluntarily cropped the graph). The marginal effects of Citroen media investments are especially very high when these investments are very low, for example between 2007 and 2009.

Table 7: Average marginal effects of media investments \check{M}_{t-1} on market-shares

$me(S_{jt}, \check{M}_{t,t-1})$	Model A				Model B			
	$\check{M}_{C,t-1}$	$\check{M}_{P,t-1}$	$\check{M}_{R,t-1}$	$\check{M}_{Z,t-1}$	$\check{M}_{C,t-1}$	$\check{M}_{P,t-1}$	$\check{M}_{R,t-1}$	$\check{M}_{Z,t-1}$
$SCitroen,t$	1.93e-05	-1.65e-09	-2.13e-09	-3.01e-10	1.68e-05	-7.20e-10	-2.82e-09	-2.00e-10
$SPeugeot,t$	-4.58e-06	1.14e-08	-3.09e-09	-5.30e-10	-7.67e-06	5.51e-09	7.72e-09	-7.52e-10
$SRenault,t$	-4.88e-06	-3.64e-09	1.35e-08	-5.96e-10	-6.43e-06	-1.14e-08	2.23e-08	-5.71e-10
$SOthers,t$	-9.89e-06	-6.10e-09	-8.24e-09	1.43e-09	-2.66e-06	6.60e-09	-2.72e-08	1.52e-09

C: Citroen; P: Peugeot; R: Renault; Z: Others.

Figures in bold: direct elasticities.

Elasticity of the share S_j relative to X_t For Model A, cross elasticities are necessarily negative and direct elasticities are necessarily positive if the parameter b is positive. Moreover, cross-elasticities of market-shares S_j with respect to a particular media budget $M_{l,t-1}$ are equal

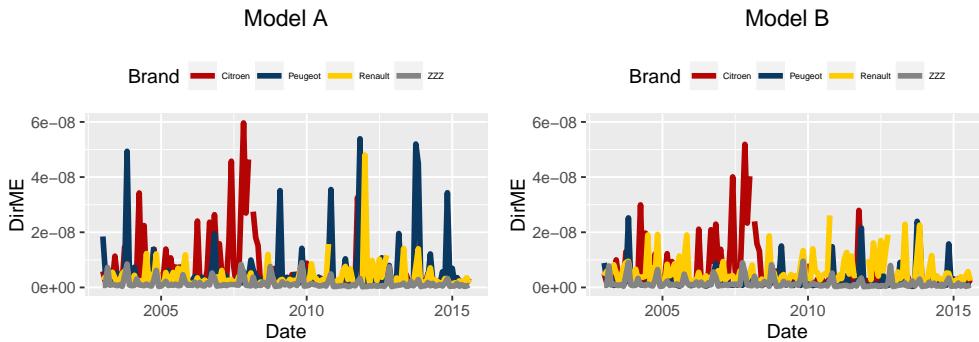


Figure 3: Direct marginal effects of $M_{j,t-1}$ on S_{jt} across time

for any brand $j \neq l$. This is a lack of flexibility of Model A compared to Model B: it does not allow positive interaction between brands, and it considers that if a brand increases its media investments of 1% it will affect in the same way all competitors market-shares S_j (they will all decrease by $b\%$).

Let us consider a situation where the market shares of Citroen, Peugeot, Renault and Others in the B segment are respectively 10%, 25%, 25% and 40%. According to Table 8, if Renault increases its media investments M_{t-1} about 1%, the average elasticity of Model A on the studied period suggests that its market-share should increase by 0.0204% to reach 25.005% and that competitors market-shares should decrease by 0.0204% to reach respectively 9.998%, 24.995% and 39.992%¹¹.

In Model B, when brand-specific effects and cross-effects are taken into account, the direct elasticity of Renault market-share in the B segment relative to its corresponding media investments is much higher than other brands (0.0327), contrary to Peugeot which has the lowest (0.0099). Note that positive cross-effects (synergies) are possible in Model B: for example when Renault invests more in media, it tends to help its own market-share a lot, but also to raise a little bit the share of Peugeot, and to have a negative impact on Citroen and Others. Then, after closure and depending on the considered values of S_j , an increase in Renault media investments in the B segment can increase or decrease the Peugeot market-share.

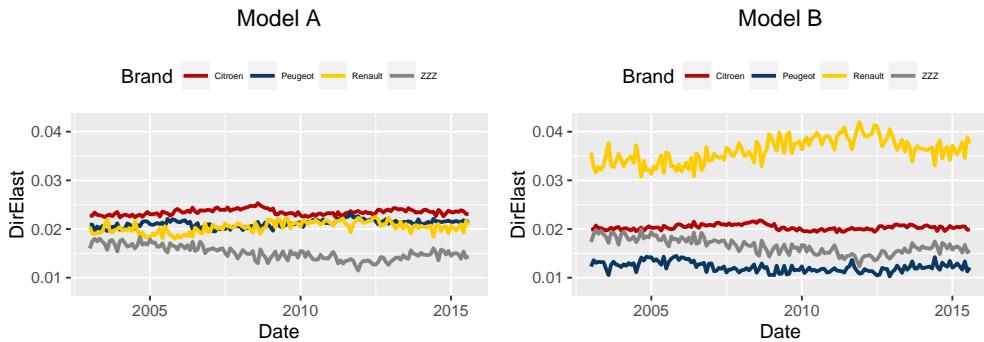
Taking the same example as previously, according to Model B, if Renault increases its media investments M_{t-1} of about 1%, the average elasticity on the studied period suggests that its market-share should increase by 0.0327% to reach 25.008% and that competitors market-shares should respectively decrease by 0.0097%, increase by 0.0119% and decrease by 0.0208% to reach respectively 9.999%, 25.003% and 39.992.

As shown in Figure 4, the estimated direct elasticities are quite stable across time. However, as elasticities in Model A are computed using the same parameter b for all brands, they are closer to each other than in Model B where they are computed using different parameters b_{jl} . The direct elasticity of Renault is larger than those of other brands during the whole studied period.

Elasticity of the ratio $\frac{S_j}{S_{j'}}$ relative to \dot{X}_l (Table 10 in the Appendix A.4)

In Model A, the elasticity of a ratio $S_j/S_{j'}$ relative to \dot{X}_j is equal to 0.0267, whereas in Model B it can be smaller or larger according to the considered brands: the largest elasticity is for S_R/S_Z

¹¹NB: here we take an example for an arbitrary share of 25% using the average elasticity. However, the only way to ensure that the sum of the modified shares $\sum_{m=1}^D S'_{mt}$ is equal to 1 is to use the corresponding elasticities calculated at the same time t , not the average elasticities.

Figure 4: Direct elasticity of S_{jt} relative to $M_{j,t-1}$ across timeTable 8: Average elasticity of market-shares relative to media investments \check{M}_{t-1}

$e(S_{jt}, \check{M}_{l,t-1})$	Model A				Model B			
	$\check{M}_{C,t-1}$	$\check{M}_{P,t-1}$	$\check{M}_{R,t-1}$	$\check{M}_{Z,t-1}$	$\check{M}_{C,t-1}$	$\check{M}_{P,t-1}$	$\check{M}_{R,t-1}$	$\check{M}_{Z,t-1}$
$S_{Citroen,t}$	0.0235	-0.0056	-0.0063	-0.0116	0.0204	-0.0028	-0.0097	-0.0078
$S_{Peugeot,t}$	-0.0032	0.0211	-0.0063	-0.0116	-0.0054	0.0099	0.0119	-0.0163
$S_{Renault,t}$	-0.0032	-0.0056	0.0204	-0.0116	-0.0043	-0.0173	0.0327	-0.0111
$S_{Others,t}$	-0.0032	-0.0056	-0.0063	0.0151	-0.0008	0.0054	-0.0208	0.0161

C: Citroen; P: Peugeot; R: Renault; Z: Others.

Figures in bold: direct elasticities.

relative to \check{X}_R which is equal to 0.0535. In general, ratios between the market-share of Renault and another brand are quite positively sensitive to media investments of Renault. For example, if the ratio S_R/S_Z is equal to $25/40 = 0.6250$ and Renault increases by 1% its media investments, then the ratio will increase to 0.6253. Let us remind that this measure does not depend on the considered period. This evolution is consistent with the fact that the market-share of Renault is very positively elastic and the market-share of “Others” is very negatively elastic to Renault media investments, as seen in Table 8.

Odds ratio of $\frac{S_j}{S_{j'}}$ to a change of \check{X}_l (Table 11 in the Appendix A.4)

As expected, this measure is consistent with the previous one. In Model A, the odds ratio of any couple of brand market-shares $S_j/S_{j'}$ to a change of 10% of $\check{M}_{j,t-1}$ is equal to 1.0025, whereas it can reach 1.0054 in Model B for the ratio S_R/S_Z for a change of 10% in $\check{M}_{R,t-1}$. It means that if the ratio of market-shares of Renault over Others is equal to $25/40 = 0.6250$ and Renault decides to increase its media budget by 10%, then this ratio will increase to 0.6266 according to Model A and to 0.6284 according to Model B.

Elasticity of $\frac{S_j}{g(S_{-j})}$ relative to \check{X}_l (Table 12 in the Appendix A.4)

As in Model A, no matter which transformation is used, the parameter b_1 will be the same, then we obtain that $e\left(\frac{S_{jt}}{g(S_{-jt})}, \frac{M_{j,t-1}}{g(M_{-j,t-1})}\right) = e\left(\frac{S_{jt}}{S_{j't}}, M_{j,t-1}\right) = e\left(\frac{S_{jt}}{S_{j't}}, \frac{M_{j,t-1}}{M_{j',t-1}}\right)$. Moreover, these elasticities are consistent with previous impact measures, and the largest one concerns the ratio $\frac{S_R}{g(S_{-R})}$ relatively to the ratio $\frac{M_R}{g(M_{-R})}$, which is equal to 0.0389%. For example, let us consider a situation where the market-shares are the following: $(S_C, S_P, S_R, S_Z)' = (13, 22, 25, 40)'$, inducing that $\frac{S_R}{g(S_{-R})} = 1.1095$. Then, if Renault increases its media investments by 1% of the geometric average of other brands media investments, we can expect its market-share to move from 110.95% to 110.99% of the geometric average market-share of others.

4.4 Complexity and goodness-of-fit

We have seen that Model A and Model B can be used for the same type of application. Model B is more complex than Model A because it allows to have component-specific parameters for each explanatory variables along with cross-effects parameters. The number of parameters to fit of Model B can be a serious limitation when the number of components D and the number of explanatory compositions K increase. For example, in our application Model A involves 10 parameters whereas Model B involves 42.

However, Model B is also more flexible than Model A in the sense that it allows to have positive synergies (positive interactions) between some shares, whereas cross elasticities of Model A are necessarily negative¹². For example, we see in Table 8 that when media investments of Citroen increase, it tends to benefit also to “Others”, and when media investments of Renault increase, it tends to benefit to Peugeot.

Is the complexity of Model B useful to explain brands market-shares of the B segment? To answer this question, let us look at cross-validated quality measures¹³ (Table 9). Quality measures agree that Model B is much better than Model A to fit brands market-shares of the B segment of the French automobile market.

Table 9: Quality measures - Model A and Model B

	R_T^2	R_A^2	KL_C	$RMSE$
Model A	0.3039	0.2578	0.0386	0.0324
Model B	0.4532	0.2816	0.0399	0.0318

5 Conclusion

The focus of this paper is to present two types of compositional models for the case when the dependent variable and some of the explanatory variables are compositions, and to interpret them. A composition is a vector of shares called components (for example the brands market-shares in a given market), which are positive numbers and sum up to one. Compositional models are transformation models: they use a log-ratio transformation to transform components into coordinates in order to enhance the estimation. The difference between Model A and Model B is due to the model specification: in Model A, a single global coefficient is associated to an explanatory composition, whereas in Model B we assume that each component of the explanatory composition has a specific impact on each component of the dependent variable. Thus, in Model B, cross-effects between components are explicitly specified and can be positive, whereas in Model A they are implicit and negative by construction. Consequently, Model B is more flexible but also much more complex than Model A, and the number of parameters to fit can be a serious limitation to use it.

This paper presents a set of possible measures, mutually consistent, to interpret parameters of these two models: marginal effects, elasticities and odds ratios. The elasticity of a component relative to an explanatory variable is the relative variation of this component to a relative variation of the explanatory variable, ceteris paribus. This type of measure is totally adapted to enhance the interpretability of these models. However, this measure is observation dependent and we have to make sure that it is stable across observations to use it. Marginal effects are not well adapted to interpret this kind of models because they depend a lot on the considered observation. The other types of measures presented have the advantage to be observation independent, but they are more difficult to interpret in practical cases because they involve ratios.

¹²As long as the direct elasticity is positive (the cross elasticity is of opposite sign of the direct elasticity by construction).

¹³The out-of-sample computation process and the quality measures used are the same than in Morais et al. (2016).

The two models are applied to the B segment of the French automobile market, for the purpose of measuring the impact of brand media investments on brands market-shares. Model B fits our data better than Model A according to several quality measures. In Model B, Renault is the brand which has the largest direct elasticity to media investments. The model shows interesting non-symmetric synergies between brands.

In a further work, it would be interesting to mix Model A and Model B in order to chose to put more or less flexibility on each explanatory variable. As compositions are observed across time, the potential autocorrelation of error terms has to be considered. Moreover, from a marketing point of view, it would be interesting to measure the impact of each channel (TV, radio, press, outdoor, internet, cinema) separately.

Acknowledgements

We thank BVA and the Marketing Direction of Renault for sharing valuable data with us, and for their support during the model specification and interpretation. This work was supported by the market research agency BVA and the French national research agency ANRT.

References

- Chen, J., X. Zhang, and S. Li (2016). Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics*, 1–16.
- Hron, K., P. Filzmoser, and K. Thompson (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics* 39(5), 1115–1128.
- Morais, J., C. Thomas-Agnan, and M. Simioni (2016, 12). A tour of regression models for explaining shares. *TSE Working Paper* (16-742).
- Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.
- Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.
- Van Den Boogaart, K. G. and R. Tolosana-Delgado (2013). *Analysing Compositional Data with R*. Springer.
- Wang, H., L. Shangguan, J. Wu, and R. Guan (2013). Multiple linear regression modeling for compositional data. *Neurocomputing* 122, 490–500.

A Appendix

A.1 Model A is a particular case of Model B

Let consider a Model B where $D_S = D_X = 3$, where the matrix of coefficients in the transformed space is equal to $\mathbf{B}^* = \begin{bmatrix} b^* & 0 \\ 0 & b^* \end{bmatrix}$, and where $\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & 0 \end{bmatrix}$. Then, $\mathbf{B} = \mathbf{VB}^*\mathbf{V}' = \frac{1}{3}b^* \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$ such that the matrix \mathbf{B} does verify the rows sum and columns sum equal to 0 requirement.

We can check that in this case we have $\mathbf{B} \square \mathbf{X} = b \odot \mathbf{X}$:

$$\begin{aligned} \mathbf{B} \square \mathbf{X} &= \mathcal{C}(X_1^{\frac{2}{3}b} X_2^{-\frac{1}{3}b} X_3^{-\frac{1}{3}b}, X_1^{-\frac{1}{3}b} X_2^{\frac{2}{3}b} X_3^{-\frac{1}{3}b}, X_1^{-\frac{1}{3}b} X_2^{-\frac{1}{3}b} X_3^{\frac{2}{3}b})' \\ &= \mathcal{C}(X_1^b (X_1 X_2 X_3)^{-\frac{1}{3}b}, X_2^b (X_1 X_2 X_3)^{-\frac{1}{3}b}, X_3^b (X_1 X_2 X_3)^{-\frac{1}{3}b})' \\ &= \mathcal{C}(X_1^b, X_2^b, X_3^b)' = b \odot \mathbf{X} \end{aligned}$$

Then, in this particular case, the Model B specification is equivalent to the Model A specification.

A.2 Marginal effect calculus

We are going to demonstrate how to compute marginal effects of the volume \check{X}_{it} on the dependent shares S_{jt} , and elasticities of S_{jt} relative to \check{X}_{it} , using the transformed and the non-transformed models. The demonstration is made for Model B, with $D = 3$ components and an

ILR transformation defined by the transformation matrix $\mathbf{V} = \begin{bmatrix} \sqrt{\frac{2}{3}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$. Let us remind that

$\mathbf{X}^* = ilr(\mathbf{X}) = \mathbf{V}' \log(\mathbf{X})$, and $\mathbf{X} = ilr^{-1}(\mathbf{X}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{X}^*))$.

We define the following transformations:

$$\begin{aligned} T : (\check{X}_1, \check{X}_2, \check{X}_3)' &\rightarrow (\check{X}_1^*, \check{X}_2^*)' \\ F : (\check{X}_1^*, \check{X}_2^*)' &\rightarrow (\mathbb{E}S_1^*, \mathbb{E}S_2^*)' = (a_1^* + b_{11}^* \check{X}_1^* + b_{12}^* \check{X}_2^*, a_2^* + b_{21}^* \check{X}_1^* + b_{22}^* \check{X}_2^*)' \\ T^{-1} : (\mathbb{E}S_1^*, \mathbb{E}S_2^*)' &\rightarrow (\mathbb{E}^\oplus S_1, \mathbb{E}^\oplus S_2, \mathbb{E}^\oplus S_3)' \end{aligned}$$

We are going to use the following property of Jacobian matrices: $J = J_{T^{-1}} J_F J_T$, implying that:

$$ME(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) = \left[\frac{\partial \mathbb{E}^\oplus S_{it}}{\partial \check{X}_{jt}} \right]_{D,D} = \left[\frac{\partial \mathbb{E}^\oplus S_{it}}{\partial \mathbb{E}S_{jt}^*} \right]_{D,D-1} \left[\frac{\partial \mathbb{E}S_{it}^*}{\partial \check{X}_{jt}^*} \right]_{D-1,D-1} \left[\frac{\partial \check{X}_{it}^*}{\partial \check{X}_{jt}} \right]_{D-1,D}$$

and

$$E(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) = \left[\frac{\partial \log \mathbb{E}^\oplus S_{it}}{\partial \log \check{X}_{jt}} \right]_{D,D} = \left[\frac{1}{\mathbf{S}_{it}} \right] \odot \left[\frac{\partial \mathbb{E}^\oplus S_{it}}{\partial \mathbb{E}S_{jt}^*} \right]_{D,D-1} \left[\frac{\partial \mathbb{E}S_{it}^*}{\partial \check{X}_{jt}^*} \right]_{D-1,D-1} \left[\frac{\partial \check{X}_{it}^*}{\partial \check{X}_{jt}} \right]_{D-1,D} \odot [\mathbf{X}_{jt}]$$

where \odot denotes the Hadamard product here (term by term product)¹⁴, $\left[\frac{1}{\mathbf{S}_{it}} \right]$ is a $D \times D - 1$ matrix with $1/S_{it}$ on the i^{th} row and $[\mathbf{X}_{jt}]$ is a $D - 1, D$ matrix with X_{jt} on the j^{th} column.

¹⁴Note that \odot in bold denote the Hadamard product whereas \odot denote the power transformation.

The Jacobian of the model in coordinates J_F

$$J_F = \begin{bmatrix} \frac{\partial \mathbb{E}S_1^*}{\partial \tilde{X}_1^*} & \frac{\partial \mathbb{E}S_1^*}{\partial \tilde{X}_2^*} \\ \frac{\partial \mathbb{E}S_2^*}{\partial \tilde{X}_1^*} & \frac{\partial \mathbb{E}S_2^*}{\partial \tilde{X}_2^*} \end{bmatrix} = \begin{bmatrix} b_{11}^* & b_{12}^* \\ b_{21}^* & b_{22}^* \end{bmatrix} = \mathbf{B}^*$$

The Jacobian of the transformation J_T The ILR transformation is defined by:

$$(\tilde{X}_1^*, \tilde{X}_2^*)' = T(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3)' = \left(\sqrt{\frac{2}{3}} \log \tilde{X}_1 - \frac{1}{\sqrt{6}} \log \tilde{X}_2 - \frac{1}{\sqrt{6}} \log \tilde{X}_3, \frac{1}{\sqrt{2}} \log \tilde{X}_2 - \frac{1}{\sqrt{2}} \log \tilde{X}_3 \right)'$$

$$\text{Then, } J_T = \begin{bmatrix} \frac{\partial \tilde{X}_1^*}{\partial \tilde{X}_1} & \frac{\partial \tilde{X}_1^*}{\partial \tilde{X}_2} & \frac{\partial \tilde{X}_1^*}{\partial \tilde{X}_3} \\ \frac{\partial \tilde{X}_2^*}{\partial \tilde{X}_1} & \frac{\partial \tilde{X}_2^*}{\partial \tilde{X}_2} & \frac{\partial \tilde{X}_2^*}{\partial \tilde{X}_3} \\ \frac{\partial \tilde{X}_3^*}{\partial \tilde{X}_1} & \frac{\partial \tilde{X}_3^*}{\partial \tilde{X}_2} & \frac{\partial \tilde{X}_3^*}{\partial \tilde{X}_3} \end{bmatrix} = \mathbf{V}' \odot \begin{bmatrix} \frac{1}{\tilde{X}_1} \\ \frac{1}{\tilde{X}_2} \\ \frac{1}{\tilde{X}_3} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{2}{3}} \frac{1}{\tilde{X}_1} & -\frac{1}{\sqrt{6}} \frac{1}{\tilde{X}_2} & -\frac{1}{\sqrt{6}} \frac{1}{\tilde{X}_3} \\ 0 & \frac{1}{\sqrt{2}} \frac{1}{\tilde{X}_2} & -\frac{1}{\sqrt{2}} \frac{1}{\tilde{X}_3} \end{bmatrix}$$

where $\begin{bmatrix} \frac{1}{\tilde{X}_j} \end{bmatrix}$ is a $D-1, D$ matrix with $1/X_j$ on the j^{th} column.

The Jacobian of the inverse transformation $J_{T^{-1}}$

$$\begin{aligned} (\mathbb{E}^\oplus S_1, \mathbb{E}^\oplus S_2, \mathbb{E}^\oplus S_3)' &= T^{-1}(\mathbb{E}S_1^*, \mathbb{E}S_2^*)' = \mathcal{C}(\exp(\mathbf{V} \cdot \mathbb{E}\mathbf{S}^*))' \\ &= \mathcal{C} \left(\exp(\mathbb{E}S_1^*)^{\sqrt{\frac{2}{3}}}; \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}} \exp(\mathbb{E}S_2^*)^{\frac{1}{\sqrt{2}}}; \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}} \exp(\mathbb{E}S_2^*)^{-\frac{1}{\sqrt{2}}} \right)' \\ &= \left(\frac{u_1}{DEN}; \frac{u_2}{DEN}; \frac{u_3}{DEN} \right) \end{aligned}$$

where

$$\begin{aligned} u_1 &= \exp(\mathbb{E}S_1^*)^{\sqrt{\frac{2}{3}}} \\ u_2 &= \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}} \exp(\mathbb{E}S_2^*)^{\frac{1}{\sqrt{2}}} \\ u_3 &= \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}} \exp(\mathbb{E}S_2^*)^{-\frac{1}{\sqrt{2}}} \\ DEN &= u_1 + u_2 + u_3 \end{aligned}$$

In order to compute the matrix $J_{T^{-1}} = \begin{bmatrix} \frac{\partial \mathbb{E}^\oplus S_1}{\partial \mathbb{E}S_1^*} & \frac{\partial \mathbb{E}^\oplus S_1}{\partial \mathbb{E}S_2^*} \\ \frac{\partial \mathbb{E}^\oplus S_2}{\partial \mathbb{E}S_1^*} & \frac{\partial \mathbb{E}^\oplus S_2}{\partial \mathbb{E}S_2^*} \\ \frac{\partial \mathbb{E}^\oplus S_3}{\partial \mathbb{E}S_1^*} & \frac{\partial \mathbb{E}^\oplus S_3}{\partial \mathbb{E}S_2^*} \end{bmatrix}$, we need to compute the derivatives of the numerators of $\mathbb{E}^\oplus \mathbf{S}$: $\mathbf{u} = (u_1, u_2, u_3)'$ with respect to $\mathbb{E}\mathbf{S}^*$.

$$\left(\frac{\partial \mathbf{u}}{\partial \mathbb{E}\mathbf{S}^*} \right) = \mathbf{V} \odot \mathbf{u} = \begin{bmatrix} \frac{\partial u_1}{\partial \mathbb{E}S_1^*} = \sqrt{\frac{2}{3}} u_1 & \frac{\partial u_1}{\partial \mathbb{E}S_2^*} = 0 \\ \frac{\partial u_2}{\partial \mathbb{E}S_1^*} = -\frac{1}{\sqrt{6}} u_2 & \frac{\partial u_2}{\partial \mathbb{E}S_2^*} = \frac{1}{\sqrt{2}} u_2 \\ \frac{\partial u_3}{\partial \mathbb{E}S_1^*} = -\frac{1}{\sqrt{6}} u_3 & \frac{\partial u_3}{\partial \mathbb{E}S_2^*} = -\frac{1}{\sqrt{2}} u_3 \end{bmatrix}$$

Now we can compute the elements of $J_{T^{-1}}$. For example, the first element of this matrix is:

$$\frac{\partial \mathbb{E}^\oplus S_1}{\partial \mathbb{E}S_1^*} = \frac{DEN \sqrt{\frac{2}{3}} u_1 - u_1 [\sqrt{\frac{2}{3}} u_1 - \frac{1}{\sqrt{6}} u_2 - \frac{1}{\sqrt{6}} u_3]}{DEN^2} = \frac{\frac{3}{\sqrt{6}} u_1 (u_2 + u_3)}{DEN^2} = \frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 (1 - \mathbb{E}^\oplus S_1)$$

using the fact that $u_1/DEN = \mathbb{E}^\oplus S_1$ and $u_2 + u_3 = DEN - u_1$.

Similar computations give the results for the whole matrix:

$$\begin{aligned} J_{T^{-1}} &= \begin{bmatrix} \frac{\partial \mathbb{E}^\oplus S_1}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^\oplus S_1}{\partial \mathbb{E} S_2^*} \\ \frac{\partial \mathbb{E}^\oplus S_2}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^\oplus S_2}{\partial \mathbb{E} S_2^*} \\ \frac{\partial \mathbb{E}^\oplus S_3}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^\oplus S_3}{\partial \mathbb{E} S_2^*} \end{bmatrix} = \begin{bmatrix} \frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1(1 - \mathbb{E}^\oplus S_1) & \frac{1}{\sqrt{2}} \mathbb{E}^\oplus S_1(\mathbb{E}^\oplus S_3 - \mathbb{E}^\oplus S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 \mathbb{E}^\oplus S_2 & \frac{1}{\sqrt{2}} \mathbb{E}^\oplus S_2(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 \mathbb{E}^\oplus S_3 & -\frac{1}{\sqrt{2}} \mathbb{E}^\oplus S_3(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_2) \end{bmatrix} \\ &= [\mathbf{S}_{it}] \odot \begin{bmatrix} \frac{3}{\sqrt{6}}(1 - \mathbb{E}^\oplus S_1) & \frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_3 - \mathbb{E}^\oplus S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 & \frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 & -\frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_2) \end{bmatrix} = [\mathbf{S}_{it}] \odot \mathbf{W}^* \end{aligned}$$

The Jacobian of the model in the simplex J

$$\begin{aligned} J = J_{T^{-1}} J_F J_T &= \begin{bmatrix} \frac{\partial S_1}{\partial X_1} & \frac{\partial S_1}{\partial X_2} & \frac{\partial S_1}{\partial X_3} \\ \frac{\partial X_1}{\partial S_2} & \frac{\partial X_2}{\partial S_2} & \frac{\partial X_3}{\partial S_2} \\ \frac{\partial X_1}{\partial S_3} & \frac{\partial X_2}{\partial S_3} & \frac{\partial X_3}{\partial S_3} \\ \frac{\partial X_1}{\partial X_2} & & \end{bmatrix} \\ &= [\mathbf{S}_{it}] \odot \mathbf{W}^* \mathbf{B}^* \mathbf{V}' \odot [1/\check{\mathbf{X}}_j] = [\mathbf{S}_{it}] \odot \mathbf{W}^* \mathbf{V}' \mathbf{B} \odot [1/\check{\mathbf{X}}_j] = [\mathbf{S}_{it}] \odot \mathbf{WB} \odot [1/\check{\mathbf{X}}_j] \\ &= [\mathbf{S}_{it}] \odot \begin{bmatrix} \frac{3}{\sqrt{6}}(1 - \mathbb{E}^\oplus S_1) & \frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_3 - \mathbb{E}^\oplus S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 & \frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^\oplus S_1 & -\frac{1}{\sqrt{2}}(\mathbb{E}^\oplus S_1 + 2\mathbb{E}^\oplus S_2) \end{bmatrix} \begin{bmatrix} b_{11}^* & b_{12}^* \\ b_{21}^* & b_{22}^* \end{bmatrix} \begin{bmatrix} \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \odot [1/\check{\mathbf{X}}_j] \\ &= [\mathbf{S}_{it}] \odot \begin{bmatrix} 1 - S_1 & -S_2 & -S_3 \\ -S_1 & 1 - S_2 & -S_3 \\ -S_1 & -S_2 & 1 - S_3 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \odot [1/\check{\mathbf{X}}_j] = ME(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) \\ &\Leftrightarrow E(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) = \begin{bmatrix} 1 \\ \mathbf{S}_{it} \end{bmatrix} \odot ME(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) \odot [\check{\mathbf{X}}_j] = \mathbf{WB} \end{aligned}$$

where $\mathbf{W}^* \mathbf{V}' = \mathbf{W}$ is a D, D matrix with $1 - S_i$ in the diagonal and $-S_i$ in the row i otherwise.

We then conclude that marginal effects and elasticities matrices are easy to compute using coefficients in the simplex or coefficients in the transformed space:

$$\begin{aligned} ME(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) &= [\mathbf{S}_{it}] \odot \mathbf{WB} \odot [1/\check{\mathbf{X}}_j] = [\mathbf{S}_{it}] \odot \mathbf{WVB}^* \mathbf{V}' \odot [1/\check{\mathbf{X}}_j] \\ E(\mathbb{E}^\oplus \mathbf{S}_t, \check{\mathbf{X}}_t) &= \mathbf{WB} = \mathbf{WVB}^* \mathbf{V}' \end{aligned}$$

A.3 Nullity of the sum of elasticities weighted by shares

We have to prove that $\sum_{m=1}^D e_{mlt} \mathbb{E}^\oplus S_{mt} = 0$. This is the necessary condition for new shares S'_{mt} , resulting from a change in X_{lt} , to sum up to one: $\sum_{m=1}^D S'_{mt} = 1 \Leftrightarrow \sum_{m=1}^D e_{mlt} \mathbb{E}^\oplus S_{mt} = 0$.
Proof:

$$\sum_{m=1}^D \mathbb{E}^\oplus S_{mt} = 1 \Leftrightarrow \sum_{m=1}^D \frac{\partial \mathbb{E}^\oplus S_{mt}}{\partial \log X_{lt}} = 0 \Leftrightarrow \sum_{m=1}^D \frac{\partial \mathbb{E}^\oplus S_{mt}}{\partial \log X_{lt}} \frac{1}{\mathbb{E}^\oplus S_{mt}} \mathbb{E}^\oplus S_{mt} = 0 \Leftrightarrow \sum_{m=1}^D e_{mlt} \mathbb{E}^\oplus S_{mt} = 0 \quad (18)$$

A.4 Impact measures

Table 10: Elasticity of ratios of market-shares $\frac{S_{jt}}{S_{j't}}$ relative to media investments $\check{M}_{l,t-1}$

Model A		Model B							
	\check{M}_{t-1}		$\check{M}_{C,t-1}$		$\check{M}_{P,t-1}$		$\check{M}_{R,t-1}$		$\check{M}_{Z,t-1}$
$e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j,t-1}\right)$	0.0267	$S_{C/P}$	0.0258	$S_{P/C}$	0.0127	$S_{R/C}$	0.0424	$S_{Z/C}$	0.0239
$e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j',t-1}\right)$	-0.0267	$S_{C/R}$	0.0246	$S_{P/R}$	0.0272	$S_{R/P}$	0.0208	$S_{Z/P}$	0.0325
$e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{l,t-1}\right)^*$	0	$S_{C/Z}$	0.0211	$S_{P/Z}$	0.0044	$S_{R/Z}$	0.0535	$S_{Z/R}$	0.0273

*where $l \neq j, j'$ and $S_{C/Z}$ means $S_{Citroen,t}/S_{Others,t}$ for example.

Table 11: Odds ratios of market-shares for an increase of 10% in media investments $\check{M}_{l,t-1}$

Model A		Model B							
For $\Delta = 10\%$	\check{M}_{t-1}		$\check{M}_{C,t-1}$		$\check{M}_{P,t-1}$		$\check{M}_{R,t-1}$		$\check{M}_{Z,t-1}$
OR $\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j,t-1}, \Delta\right)$	1.0025	$S_{C/P}$	1.0025	$S_{P/C}$	1.0012	$S_{R/C}$	1.0045	$S_{Z/C}$	1.0022
OR $\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j',t-1}, \Delta\right)$	0.9975	$S_{C/R}$	1.0024	$S_{P/R}$	1.0030	$S_{R/P}$	1.0026	$S_{Z/P}$	1.0031
OR $\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{l,t-1}, \Delta\right)^*$	0	$S_{C/Z}$	1.0020	$S_{P/Z}$	1.0007	$S_{R/Z}$	1.0054	$S_{Z/R}$	1.0028

*where $l \neq j, j'$ and $S_{C/Z}$ means $S_{Citroen,t}/S_{Others,t}$ for example.

Table 12: Elasticity of ratios $\frac{S_{jt}}{g(S_{-jt})}$ relative to $\check{M}_{l,t-1}$

Model A		Model B				
		$\check{M}_{C/g(-C)}$	$\check{M}_{P/g(-P)}$	$\check{M}_{R/g(-R)}$	$\check{M}_{Z/g(-Z)}$	
$e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{M}_{j,t-1}\right)$	0.0267	$S_{C/g(-C)}$	0.0239	-0.0022	-0.0176	-0.0040
$e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{M}_{l,t-1}\right)^*$	0	$S_{P/g(-P)}$	-0.0106	0.0148	0.0112	-0.0154

*where $l \neq j$.

$S_{C/g(-C)}$ means $\frac{S_{Ct}}{g(S_{-Ct})}$, where $g(S_{-Ct})$ is the geometric mean of others shares than Citroen.

Applied data analytics on multi-element geochemistry for pre-mining characterization of geological and geometallurgical attributes: Examples from the Rosemont Cu-Mo-Ag skarn deposit, Tucson, Arizona

J.C. Ordóñez-Calderón^{1,2}, S. Gelich¹, and J.F. Oliveira³

¹Hudbay, Toronto, Ontario, Canada; ordonez.jc@gmail.com

²Harquail School of Earth Sciences, Laurentian University, Sudbury, Ontario, Canada

³XPS-A Glencore Company, Falconbridge, Ontario, Canada

Abstract

Planning mining operations involves the characterization of numerous variables such as stratigraphy, alteration facies, ore grades, oxidation stage, metal recoveries, hardness, and deleterious minerals. These variables are characterized on drill core samples to assess the economic feasibility of mineral deposits. However, they are usually analyzed in <5% of the available samples. Some other variables are less reliable, for example visual identification of mineralogy and lithology in metasomatized rocks. In contrast, multi-element geochemical data is reliable and routinely collected in the majority of drill core samples. Geochemical attributes and the engineering of ore extraction are directly related to ore forming processes. As a corollary, chemical elements can be ideal predictor variables of geological and geometallurgical properties.

To circumvent problems arising from unreliable geological observations and limited geometallurgical data, this study applies data analytics to 4-acid multi-element geochemical data from the Rosemont deposit. First, a lithogeochemical and chemostratigraphic model was built using hierarchical cluster analysis on compositional variables and centered ternary diagrams. Second, predictive models were developed to map skarn alteration facies in 3D geospace using geochemical variables as inputs and quantitative mineralogy, QEMSCAN and XRD, as training outputs. Ten-fold cross-validation indicates that random forest and linear discriminant analysis are the best predictive models and outperform visual identification of skarn facies. Third, cross-validated classification and regression trees (CART) were used to identify the most relevant geochemical, mineralogical, and metallurgical variables affecting metal recoveries. The CART model recognizes 6 relevant ore types within the deposit.

The geochemical data analysis demonstrates that stratigraphy has a strong control on metal grades, whereas metal recoveries are strongly controlled by stratigraphy and oxidation processes. Applying data analytics to integrate geochemical inputs with geological and geometallurgical outputs is a powerful tool for establishing the link between ore forming processes, ore extraction, and mitigation of economic risks.

Key words: Lithogeochemistry, Chemostratigraphy, Geometallurgy, Simplex, Predictive Modeling, Compositional Data Analysis.

1 Introduction

Skarn deposits are characterized by large geological variability resulting from several factors such as (1) diverse stratigraphic and magmatic environments with a wide variety of lithologies including chemical, siliciclastic, and mixed chemical-siliciclastic sedimentary rocks, volcanic rocks, and polyphasic intrusions, (2) contact metamorphism and multi-stage metasomatic alteration resulting in recrystallization and complex mineralogical changes including garnet, pyroxene, wollastonite, vesuvianite, serpentine, amphibole, and epidote alteration, (3) layer cake stratigraphy and faulting facilitates the circulation of meteoric waters and oxidation of sulfide-rich ore bodies. All these characteristics make skarn deposits complex from a geological and a mining perspective.

The geological variability of skarn deposits complicates the visual identification of lithologies and stratigraphic domains. One of the most important steps in characterizing a skarn deposit is to establish a reliable mine-scale stratigraphy which is critical for mineral resource estimation and of economic significance for mining operations given that metallurgical parameters such as metal grades, grinding efficiency, deleterious minerals, deleterious elements, and metal recoveries are spatially related and controlled by primary stratigraphic and lithologic features.

The goal of this study was to use 4-acid multi-element geochemical data from the Rosemont deposit to develop a lithogeochemical and chemostratigraphic framework to increase the certainty of mine-scale geological domains. Simplicial projections such as tetrahedral and ternary diagrams were developed to facilitate lithogeochemical classification and to establish an informal mine-chemostratigraphy. In addition, several machine learning techniques were used to develop predictive models of skarn alteration and ore body characterization in which geochemical data plays the role of predictor of less represented output variables such as quantitative mineralogy and metallurgical properties. We present a data driven approach to characterize the geological attributes and mining properties of a mineral deposit using compositional data analysis and advanced data analytics.

2 Geology and mineralization

The Rosemont deposit, Hudbay Minerals Inc., is a copper-molybdenum-silver skarn located in the Laramide porphyry belt of Arizona, 40 km to the southeast of Tucson. The deposit contains over a billion tons of mineralized rocks hosted dominantly within a Paleozoic chemical sedimentary sequence (Fig. 1). The deposit comprises three major structural-stratigraphic domains, called (1) the Lower plate, (2) the Upper plate, and (3) the West block (Fig. 1).

The Lower plate forms an upright, east-dipping, homoclinal sequence composed dominantly of Paleozoic chemical sedimentary rocks with minor interbedded siliciclastic rocks including limestone, dolostone, marlstone, calcareous siltstone, calcareous sandstone, mudstone, siltstone, and fine-grained sandstone (Fig. 1) (Rasmussen et al., 2012; this study). Most of the economic mineralization of the deposit is contained in this domain.

The Upper plate is composed dominantly of Mesozoic and Cenozoic siliciclastic and volcanic rocks overlying and bounded to the Lower plate through a low angle fault representing a structurally overprinted unconformity (Fig. 1) (Rasmussen et al., 2012; this study). Arkose, silty sandstone, mudstone, conglomerate, and a package of andesitic volcanic rocks are typical lithologies of the Upper plate (Fig. 1).

The West block is bounded to the Lower plate by the steeply-dipping backbone fault system defining a structurally complex contact along which Precambrian granitoids are structurally interleaved with panels of Paleozoic chemical sedimentary rocks typical of the Lower Plate (Fig. 1) (Rasmussen et al., 2012; this study).

Mineralized Tertiary felsic porphyries intrude the Lower plate and are thought to be the source of mineralization and calc-silicate metasomatism of the Rosemont skarn deposit (Keith and Wilt, 1986;

J.C. Ordóñez-Calderón et al.
this study).

page 3

Calc-silicate alteration in the Lower plate and chemical sedimentary rocks of the West block is characterized by garnet, pyroxene, wollastonite, and serpentine skarn facies. In contrast, epidote skarns are the dominant alteration facies in the Upper plate.

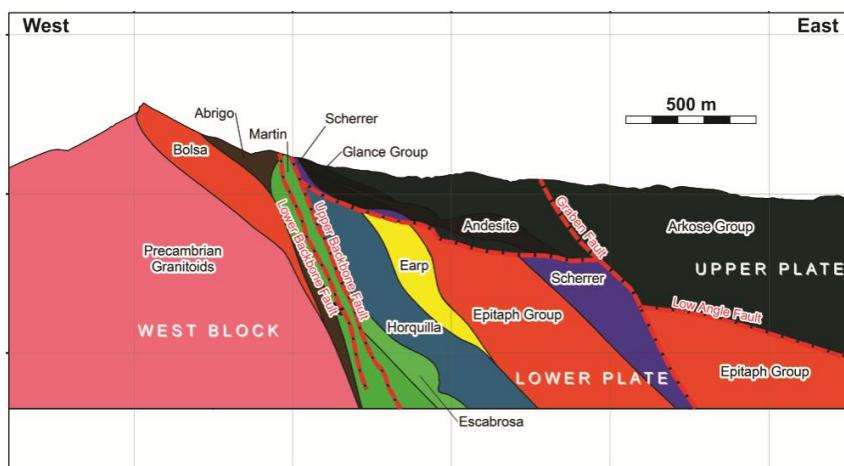


Figure 1: cross section with structural domains and stratigraphy of the Rosemont deposit. The cross section was reconstructed using core logging and the lithogeochemical model discussed in this study.

High-grade mineralization is primarily hosted in the Lower plate, whereas the Upper plate exhibits minor low-grade mineralization. The West block contains some economic mineralization, mostly hosted in structural panels of chemical sedimentary rocks.

3 Methodology

3.1 Laboratory analysis

The whole rock multi-element geochemical dataset of the Rosemont deposit comprise 33,000 drill core samples homogeneously distributed across the deposit and collected from 90 diamond drill holes sampled from top to bottom. All samples were analyzed for 41 elements by inductively coupled plasma mass spectrometry (ICP-MS) after a 4-acid digestion, and for soluble copper, a proxy for oxidation of copper sulfides, by atomic absorption spectroscopy (AAS) following a room temperature leach with sulfuric acid at 5%.

A subset of 400 samples were analyzed for quantitative mineralogy by X-ray diffraction (XRD), 100 samples for quantitative evaluation of minerals by scanning electron microscopy (QEMSCAN), and 400 samples for cation exchange capacity (CEC) to characterize swelling clays.

Ore hardness from a SAG/AG milling perspective was estimated for 140 samples through the SAG power index (SPI). The grindability of the ore was measured on 378 samples using the Bond Work index (BWi). In addition, 107 tests of total copper rougher recoveries (RCu %) were analyzed for ore type classification. The SPI, BWi, and RCu % are critical geometallurgical parameters for establishing the economic model of this mining operation.

3.2 Statistical methods

The variation matrix of Aitchison (1986) was used as a metric of similarity in hierarchical cluster analysis in order to cluster mineralogical and geochemical variables (van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). The resulting cluster dendrogram facilitates the identification of groups of variables and provides a data driven approach to select subcompositions, reduce dimensionality by compositing variables, and classification.

To circumvent the effect of closure in multivariate statistical analysis several representations of compositions were used in this study. First, the isometric log-ratio (ilr) transformation was used for K-means clustering (Egozcue et al., 2003). To facilitate interpretation of the ilr-coordinates the concept of balances between groups of parts was applied (Egozcue and Pawlowsky-Glahn, 2005). Balances are calculated using a sequential binary partition in which the D-parts are systematically divided into two non-overlapping composite variables following (D-1) steps. The procedure results in (D-1) ilr-variables, or coordinates, representing balances between groups in \mathbb{R}^{D-1} .

An additional representation of compositional data used in this study is the centered log-ratio transformation (clr) (Aitchison, 1986). The clr coefficient of a composition is obtained by dividing each component by the geometric mean of the composition and then taking the natural logarithm. The clr-transformed variables were used as input variables in different predictive models of skarn alteration facies such as support vector machines (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification and regression trees (CART), and random forests (RF).

All the compositional data analysis was conducted using the R package ‘compositions’ (van den Boogaart and Tolosana-Delgado, 2013; van den Boogaart et al., 2015). In addition, left-censored geochemical and mineralogical variables were imputed using the robust multiplicative lognormal replacement method which is part of the R package ‘zCompositions’ (Palarea-Albaladejo et al., 2014; Palarea-Albaladejo and Martín-Fernández, 2015).

4 Results and discussion

4.1 Lithogeochemical modeling

4.1.1 Clustering variables

To create a lithogeochemical model of strongly metasomatized rocks in the Rosemont deposit we selected an 18-part subcomposition comprising elements that are immobile during metasomatic alteration and contain information on rock forming processes. These components are major elements Al and P, high field strength elements (HFSE) Hf, Zr, Th, Ti, Nb, Ta, and Y, light rare earth elements (LREE) La and Ce, and transition metals (TM) Sc, Cr, Ni, Co, and V (Taylor and McLennan, 1985; MacLean and Barrett, 1993; McLennan, 2001; Kelemen et al., 2003; Ordóñez-Calderón et al., 2008, 2016; Mungal, 2014). Calcium and Mg are mobile during metasomatic alteration but were included in the subcomposition because they are the most important components of limestone and dolostone.

At least three major groups of variables can be recognized in the cluster dendrogram resulting from hierarchical cluster analysis using the variation matrix as a metric of similarity (Fig. 2). Group 1 includes Ca and Mg, Group 2 is composed of Cr, Ni, Co, V, and P, and Group 3 is composed of Hf, Zr, Th, Ti, Nb, Ta, Y, La, Ce, Sc, and Al (Fig. 2).

Group 1 represents the geochemical characteristics of limestone and dolostone. Group 2 and Group 3 represent the dominant geochemical attributes of siliciclastic sedimentary rocks and crystalline rocks, mostly volcanic and intrusive rocks (Fig. 2). Variations in Group 2 and Group 3 variables can

be attributed to provenance in siliciclastic rocks and different degrees of fractionation in crystalline rocks. These associations of variables suggest that the 18-part subcomposition contains sufficient information to investigate the lithological attributes of the Rosemont deposit.

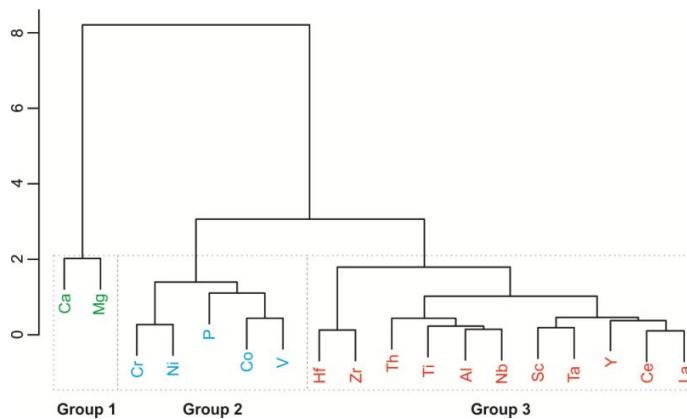


Figure 2: Cluster dendrogram of geochemical variables obtained from hierarchical cluster analysis using the variation matrix as a metric of similarity.

4.1.2 Mapping the geochemical space in the simplex

The simplex is an appropriate geometric space to map geochemical data because it represents the data in its relative scale in which ratio of parts carry most of the information (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Mateu-Figueras, 2003; Aitchison and Egozcue, 2005; Barceló-Vidal and Martín-Fernández, 2016). A graphical advantage of simplicial geometry is that the closure operation allows the visualization of geochemical data with D components in a (D-1)-dimensional space, which makes possible the visualization of a 4-part subcomposition in 3 dimensions using a tetrahedral projection or 4-part simplex.

The information extracted from the cluster dendrogram (Fig. 2) was used to create composite variables to reduce the 18-part subcomposition into 4-part and 3-part subcompositions, and to facilitate data visualization using tetrahedral and ternary diagrams.

The 4-part subcomposition comprises (1) Ca, (2) Mg, (3) Group 2, and (4) Group 3 variables (Fig. 2). The 3-part subcomposition combines Group 2 and Group 3 variables into a single composite variable. These composite variables were calculated using the geometric mean of the group. In addition, to circumvent problems with hidden structure in collapsed data clouds, the subcompositions were centered by perturbing the data matrix by the inverse of the center of the dataset.

Figure 3 shows a centered tetrahedral diagram, 4-part simplex, for the Rosemont deposit. At least two major domains can be recognized. Domain 1 represents a compositional plane that can be projected from the Ca-Mg border into the center of the Group 2-Group 3 border (Fig. 3). This plane represents chemical sedimentary rocks within the Lower plate of the Rosemont deposit. Domain 2 forms a tighter pattern and can be interpreted as a compositional line controlled dominantly by Group 2 and Group 3 variables (Fig. 3). Domain 2 represents siliciclastic sedimentary rocks and crystalline rocks occurring in the Upper plate and West block of the Rosemont deposit (Fig. 1).

The information displayed in the tetrahedral diagram can be projected into a ternary diagram with vertices Ca, Mg, and Group 2-Group 3 combined into a single variable (Fig. 4A). Empirical lithogeochemical boundaries for the Rosemont skarn deposit were established by mapping high

point-density areas using a heat map on the ternary diagram (Fig. 4A). These areas were labeled using a practical nomenclature with 7 classes reminiscent of the overall geological processes that may result in these clusters (Fig. 4A). Accordingly, high point-density areas close to the vertices of the ternary diagram were named limestone, dolostone, and siliciclastic-crystalline classes (Fig. 4A). These three classes represent the lithogeochemical end members of the Rosemont skarn deposit comprising relatively pure limestone and dolostone, with minor incorporation of siliciclastic component, as well as all siliciclastic and crystalline rocks represented by arkose, porphyry, andesite, granitoid, conglomerate, etc. The other four lithogeochemical classes on Figure 4A represent compositional mixtures of the three lithogeochemical end members described above.

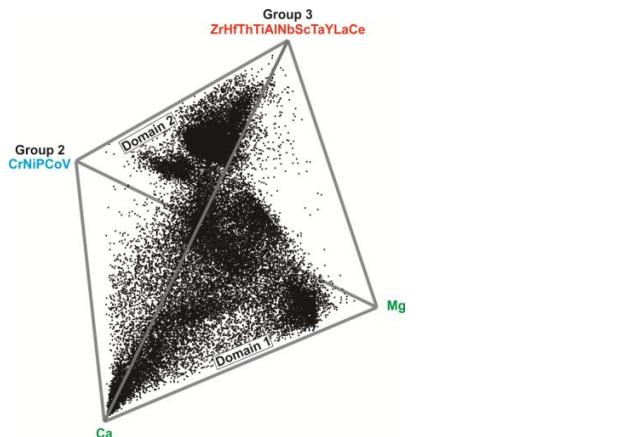


Figure 3: Centered tetrahedral plot, 4-part simplex, representing an 18-part subcomposition reduced into 4 components.

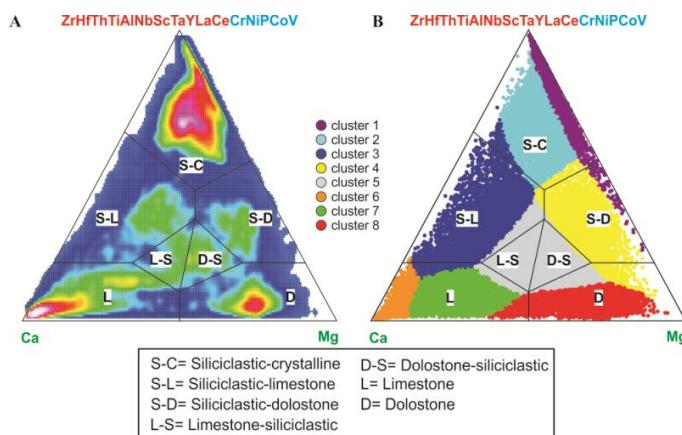


Figure 4: Centered ternary diagrams, 3-part simplex. (A) Heat map highlighting high point-density areas and empiric lithogeochemical classes. (B) Clusters resulting from K-means clustering, K=8, of 2 balances (ilr-coordinates) representing the same geochemical space given by the ternary diagram.

An alternative data-driven approach to define lithogeochemical classes is to conduct a K-means cluster analysis of observations based on the 3-part subcomposition represented in the ternary diagram (Fig. 4A). In this study, K-means clustering was applied to a data matrix of balances.

Two balances were calculated using the 3-part subcomposition represented on Figure 4A to have an

J.C. Ordóñez-Calderón et al.

page 7

equivalent geochemical space in \mathbb{R}^2 for cluster analysis. These ilr-variables were calculated using the following binary partition:

balance 1 [Mg | Ca]

balance 2 [Cr, Ni, P, Co, V, Hf, Zr, Th, Ti, Al, Nb, Sc, Ta, Y, Ce, La | Mg, Ca]

We computed a K-means clustering with K=8 to compare the cluster analysis with the empirical lithogeochemical boundaries (Fig. 4B). It is evident that clustering the 2 balances provides a remarkable similarity with the empirical boundaries chosen based on the point-density heat map (Fig. 4A).

There is a strong association of the mineralization with the lithogeochemical classes. Figure 5 shows a box plot of copper grades for samples with >1000 ppm Cu to exclude non economic mineralization. Lithogeochemical classes representing relatively clean chemical sedimentary rocks such as dolostone, limestone, and limestone-siliciclastic are associated with higher copper grades than those of mixed chemical-siliciclastic rocks such as siliciclastic-crystalline, siliciclastic-limestone, and siliciclastic-dolostone classes (Fig. 5).

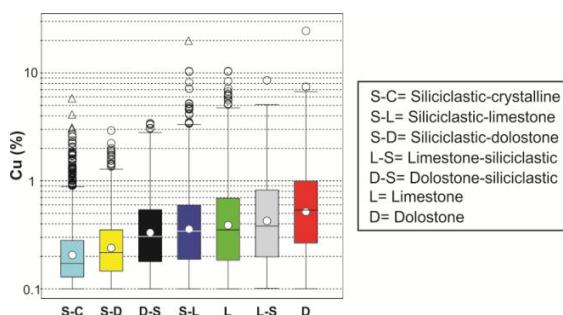


Figure 5: Box plot showing the relationships between Cu grades and the lithogeochemical classes devised in this study.

We suggest that the lithogeochemical model devised in this study (Fig. 4A) can be used as a geochemical exploration tool to assess and map the skarn fertility of chemical-siliciclastic sedimentary successions in areas permissive of skarn mineralization. Accordingly, chemical sedimentary rocks relatively clean of siliciclastic component are more fertile for economic skarn mineralization than their mixed chemical-siliciclastic counterparts.

4.1.3 Chemostratigraphy

The lithogeochemical model devised in this study was used to create a simplified chemostratigraphy to better understand the geological attributes of the Rosemont deposit. Figure 6 shows a vertical section of the deposit representing 90 diamond drill holes with 33,000 samples color coded by lithogeochemical class to visualize the spatial variability and to reveal the structural architecture of the deposit (Fig. 6).

The spatial distribution of the lithogeochemical classes shows significant continuity and clearly maps the three major structural domains of the deposit (Fig. 6). The Upper plate, dominated by the siliciclastic-crystalline class representing arkose, andesite, conglomerate, etc. The West block, evident by a sliver of rocks with siliciclastic-crystalline class attributes representing Precambrian granitoids and quartzites interleaved with dolostone, limestone, and siliciclastic-limestone classes. The Lower plate, the major host of economic mineralization, discriminated by class attributes such as limestone, dolostone, and all classes representing mixed chemical-siliciclastic sediments (Fig. 4A).

A simplified chemostratigraphic model was developed for the Lower plate based on the

lithogeochemical model. The Upper plate was not subdivided given that most of the economic mineralization is hosted in the Lower plate. Accordingly, from bottom to top the mineralized Lower plate of the Rosemont deposit comprises three chemostratigraphic units: (1) Lower limestone unit, (2) Mixed unit of chemical-siliciclastic sedimentary rocks, and (3) Upper dolostone unit (Fig. 7).

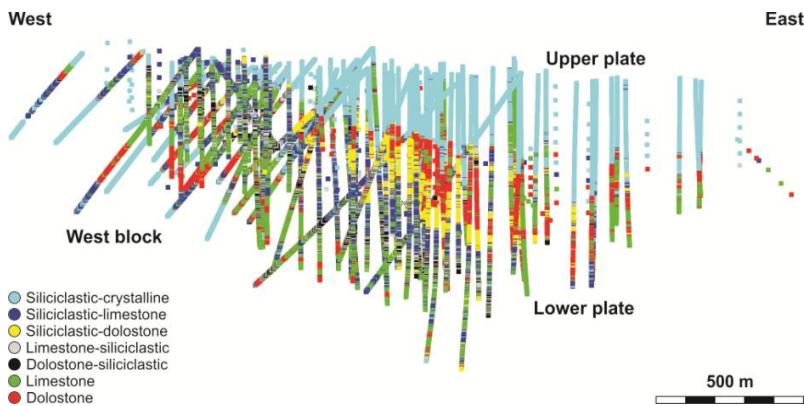


Figure 6: East-West vertical section projecting 90 diamond drill holes sampled from top to bottom and analyzed for multi-element geochemistry. The section contains 33,000 samples color coded by lithogeochemical class (Fig. 4A).

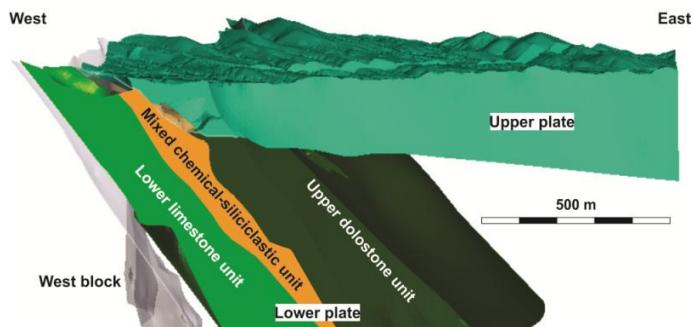


Figure 7: Geospatial 3D model with the proposed chemostratigraphy for the Rosemont deposit based on the lithogeochemical classes (Figs. 4A and 6).

From a mining perspective, the advantage of an informal chemostratigraphy over conventional stratigraphic models is that the geological variability is more efficiently characterized which has immediate applications in resource modeling and geometallurgical characterization (cf., Gregory et al., 2013; Amer et al., 2014; Maydagán et al., 2016).

4.2 Predictive modeling of skarn facies

Calc-silicate metasomatism results in large mineralogical variability of skarn deposits. These metasomatic alteration patterns can be used in exploration programs as vectors to discover new mineralization. In addition, the grindability of the ore in mining operations is strongly controlled by skarn mineralogy. Therefore, reliable mapping of skarn alteration facies is critical for exploration and mining.

Visual estimation of skarn mineralogy and the determination of the number of skarn classes are subjective and very often inconsistent within and between drill holes complicating the geospatial 3D modeling of skarn alteration facies. This issue is critical in skarns dominated by fine grain textures such as the Rosemont deposit. To circumvent these problems, we fitted 12 predictive models of skarn alteration facies to a data matrix comprising the clr coefficients of 41 geochemical variables.

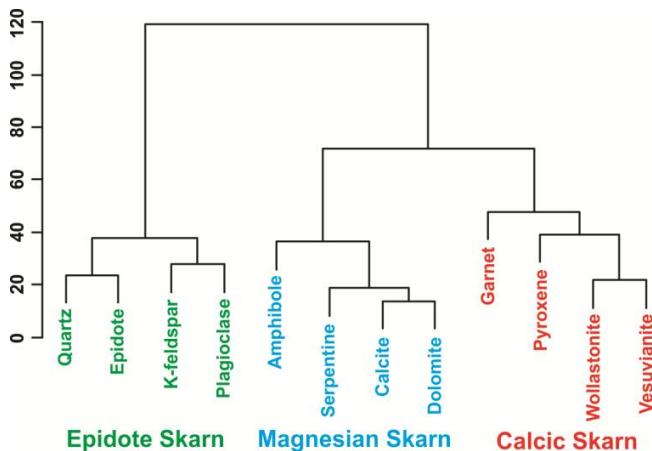


Figure 8: Cluster dendrogram of mineralogical variables obtained from hierarchical cluster analysis of XRD and QEMSCAN mineralogy using the variation matrix as a metric of similarity.

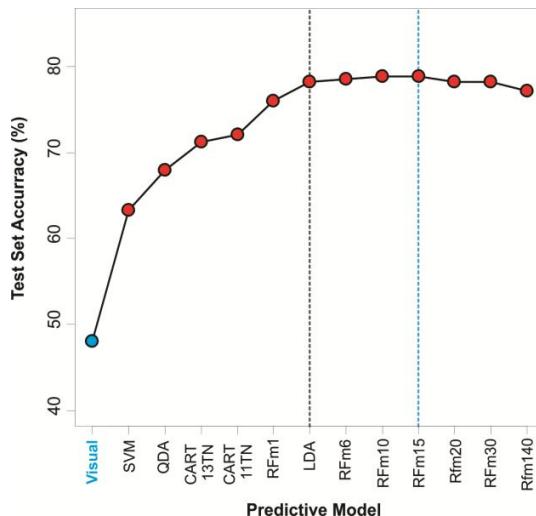


Figure 9: Percent test set accuracy of 12 machine learning algorithms developed to predict skarn classes as indicated by the dendrogram on Figure 8. The accuracy of the geologist core logging observations (Visual) is included for comparison. Abbreviations: SVM= support vector machines; QDA= quadratic discriminant analysis; CART= classification and regression trees, 13N and 11N refers to the number of terminal nodes; and RF= random forests, m6, m10, m15, etc refers to the number of random variables at each split. All random forests contain 500 individual trees.

The number of classes related to skarn alteration facies was established by performing a hierarchical cluster analysis using the variation matrix of a subcomposition of 12 minerals analyzed by XRD and QEMSCAN (Fig. 8). The cluster dendrogram suggests at least 3 broad skarn classes including (1) calcic skarn composed of garnet, pyroxene, wollastonite, and vesuvianite, (2) magnesian skarn composed primarily of serpentine, amphibole, dolomite, and calcite, and (3) epidote skarn. Samples in which at least 1 skarn-related mineral (epidote, serpentine, garnet, pyroxene, wollastonite, and vesuvianite) exceeds 5% were considered altered. A skarn class was assigned to these altered samples using a centered ternary diagram with composite variables representing the hierarchical agglomeration indicated by the cluster dendrogram on Figure 8, following the same methodology used to establish the lithogeochemical model (Section 4.1.2).

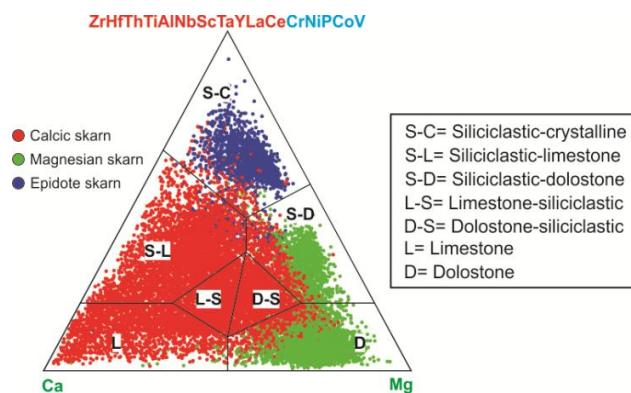


Figure 10: Predicted skarn classes plotted in the ternary diagram representing the lithogeochemical space of the Rosemont deposit.

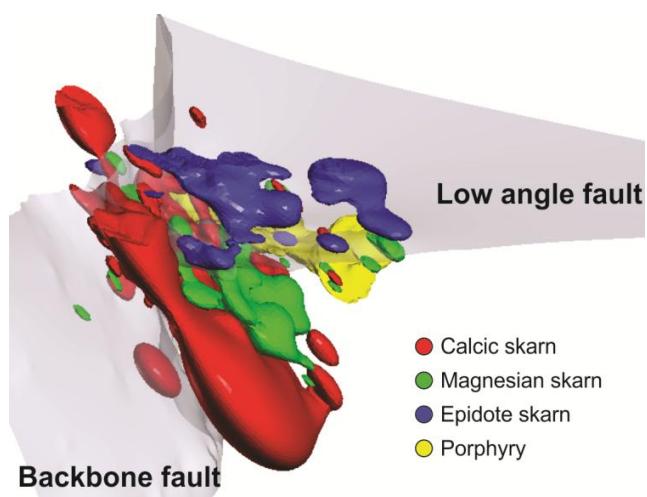


Figure 11: Predicted skarn classes displayed on the 3D geospace of the Rosemont deposit. Porphyritic intrusions, responsible for skarn alteration and mineralization are plotted for comparison.

The overall accuracy of the predictive models was assessed on test sets following 10-fold cross-validation by randomly splitting the dataset into 10 non-overlapping groups. In each of these groups 10% of the samples were extracted as a test set and the remaining samples were used as

a training set to fit the predictive model. The percent test set accuracy represents the average accuracy obtained from the 10 test sets (Fig. 9).

Figure 9 shows the cross-validation results for 12 predictive models and the accuracy of visual lithological logging. Visual estimation of skarn alteration facies is only 48% accurate, relative to XRD and QEMSCAN mineralogy, and is outperformed by all the statistical learning models used in this study (Fig. 9). The most accurate algorithm is the random forests (79% accuracy) with a random sample of 15 predictors (m) taken at each split (RFM15). All random forests models were fit on 500 decision trees bootstrapped on the training sets generated for cross-validation. It is noteworthy that linear discriminant analysis (LDA) is very close in predictive accuracy as the random forests model (Fig. 9).

The skarn alteration facies were plotted on the ternary diagram devised for lithogeochemical modeling (Fig. 10). It is evident that the epidote skarn is restricted to rocks with siliciclastic-crystalline geochemical signatures. The magnesian skarn (serpentinite-amphibole) is associated with dolostone and siliciclastic-dolostone classes. In contrast, the calcic skarn (garnet-pyroxene-wollastonite-vesuvianite) overlaps nearly all lithogeochemical classes (Fig. 10).

A geospatial 3D model of the RFM15 model was produced to investigate the spatial variability of the skarn alteration facies (Fig. 11). The spatial zoning of the skarn alteration facies is evident on Figure 11. The mineralized porphyritic intrusions are also included in the 3D model for comparison. Epidote skarns are located above the porphyries within the upper plate, which is composed of siliciclastic and volcanic rocks. In contrast, the spatial distribution of the calc-silicate and magnesian skarns reflect well the chemostratigraphy of the Lower plate, with the magnesian skarn restricted to the Upper dolostone unit (Figs. 7 and 11). The predictive 3D model presented on Figure 11 shows the epidote alteration is a vector to concealed porphyry-skarn systems and therefore an important tool for exploration.

4.3 Ore types

A 6 node classification and regression tree (CART) was used to classify ore types using 107 measurements of total copper rougher recoveries (RCu %) as the response variable. Predictor variables in the CART model include the BWi, SPI, swelling clays, magnesium clays, and the ratio of total copper/soluble copper (pctCuox) expressed as a percentage. The results of the CART model indicate that BWi and SPI are not critical parameters to delineate ore types based on rougher recoveries. In contrast, the most important variables in hierarchical order are pctCuox, swelling clays, and magnesium clays (Fig. 12).

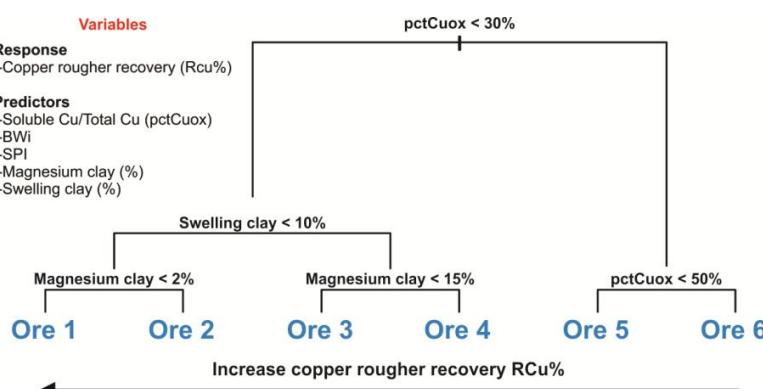


Figure 12: Classification and regression tree developed to classify ore types based on geochemical, mineralogical, and metallurgical predictor variables. Cross-validation suggests that the most optimum CART model is a tree with 6 terminal nodes.

Six ore types were defined based on the CART model. Total copper rougher recoveries systematically increase from ore type 6 to 1. Ore types 1 and 2 are clean ores (clay poor) and have the highest recoveries. Ore type 3 and ore type 4 are respectively swelling clay rich and magnesium clay rich ores. Ore types 5 and 6 are oxidized ores.

The ore type classes given by the CART model were used as categorical variables to create predictive models of ore types to populate the entire drill hole dataset with ore type classes following the same methodology used to predict skarn alteration facies (Section 4.2). The predicted ore types were used to build a geospatial 3D model of ore types (Fig. 13).

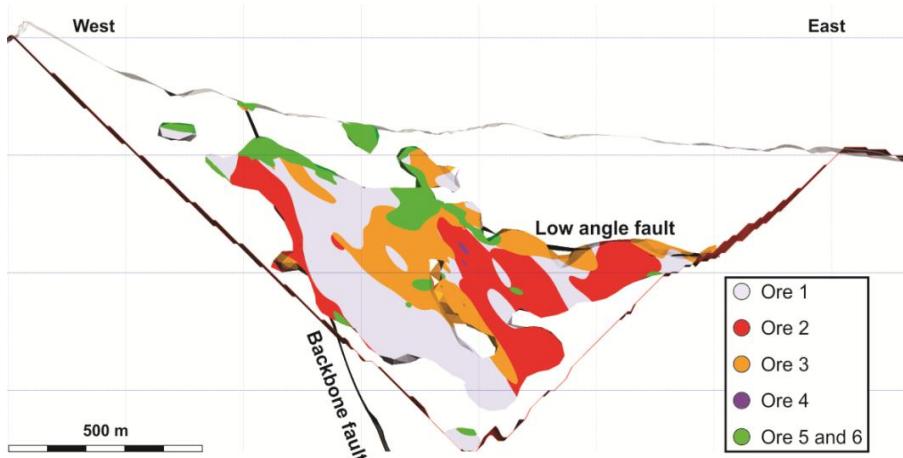


Figure 13: Cross-section representing predicted ore types across the Rosemont deposit.

The predictive 3D model of the ore types indicates that (1) oxidized ore types 5 and 6 are dominant in the uppermost parts of the deposit and along fault zones, (2) most of the deposit comprises clean ore types 1 and 2, (3) swelling clay rich ore 3 is dominant in the Mixed unit of chemical-siliciclastic sedimentary rocks defined in the chemostratigraphy (Figs. 7 and 13), and (4) magnesium clay rich ore 4 is less abundant and associated with the Upper dolostone unit (Figs. 7 and 13).

5 Conclusions

This study presented a data driven approach for systematic characterization of lithologies, stratigraphy, alteration facies, and geometallurgical properties of the Rosemont skarn deposit allowing the following conclusions:

- 1) Simplicial projections of geochemical data using tetrahedral and ternary diagrams are an effective tool to devise lithogeochemical models to circumvent subjective identification of rock types in strongly metasomatized chemical sedimentary rocks (Fig. 4A). The lithogeochemical model indicates that copper grades tend to be higher in relatively pure chemical sedimentary rocks, limestone and dolostone, in which the amount siliciclastic component is minor (Fig. 5). Therefore, the lithogeochemical model presented here provides a useful tool to assess the economic potential of chemical sedimentary rocks in environments permissive of skarn mineralization.
- 2) The spatial distribution of the lithogeochemical classes suggests that the stratigraphy of the Lower plate of the Rosemont deposit can be simplified into 3 chemostratigraphic units including a Lower limestone unit overlaid by a Mixed unit of chemical-siliciclastic sedimentary rocks, and Upper dolostone unit (Fig. 7). These chemostratigraphic units capture most of the geochemical variability of the Rosemont deposit and are strongly associated with ore grades, skarn alteration

facies, and ore types.

- 3) Several compositionally oriented predictive models of skarn alteration facies outperform visual identification of the alteration (Fig. 9). The best performing algorithm is a random forests model (RFm15), fit on 500 regression trees, with a random sample of 15 predictors (Fig. 9). The geospatial analysis of this predictive model indicates that the epidote skarn is located a few hundred meters above the mineralized porphyry (Fig. 11). This characteristic indicates that epidote alteration is an important indicator of concealed skarn-porphyry systems. In addition, the magnesian skarn is clearly related to the Upper dolostone unit defined in the chemostratigraphy.
- 4) Classification and regression trees are a simple and effective tool to classify ore types relative to metal recoveries (Fig. 12). In this study 6 ore types were defined all of which have clear structural and stratigraphic control (Fig. 13).
- 5) The data driven methodology presented here is an efficient approach to extract geological and metallurgical insight from diverse datasets routinely collected in mining operations. Compositional data analysis of geochemical and mineralogical data is at the core of the data analysis.

Acknowledgements

We thank Geologists Jeff Cornoyer and David Young for sharing their knowledge of the Rosemont deposit and for their hard work during two drilling programs conducted by Hudbay. We are thankful with Hudbay Minerals Inc., for allowing us to publish this scientific contribution.

References

- Amer, T.E., El Assay, I.E., Rezk, A.A., El Kammar, A.M., El Manawi, A.W., Abu Khoziem, H.A., 2014. Geometallurgy and processing of North Ras Mohamed poly-mineralized ore materials, South Sinai, Egypt. *International Journal of Mineral Processing* 129, pp. 12–21.
- Aitchison, J., 1986. The statistical analysis of compositional data. *Monographs on Statistics and Applied Probability*. London, Chapman & Hall, p. 416.
- Aitchison, J., Egozcue, J.J., 2005. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology* 37 (7), pp. 829–850.
- Barceló-Vidal, C., Martín-Fernández, J.A., 2016. The mathematics of compositional analysis. *Australian Journal of Statistics* 45, pp. 57–71.
- Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37 (7), pp. 795–828.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35 (3), pp. 279–300.
- Gregory, M.J., Lang, J.R., Gilbert, S., Hoal, K.O., 2013. Geometallurgy of the Pebble porphyry copper-gold-molybdenum deposit, Alaska: Implications for gold distribution and paragenesis. *Economic Geology* 108, pp. 463–482.
- Keith, S.B., and Wilt, J.C., 1986. Laramide orogeny in Arizona and adjacent regions: A stratotectonic synthesis. In Beatty, B., and Wilkinson, P.A.K. (Eds.), *Frontiers in geology and ore deposits of Arizona and the Southwest*. Arizona Geological Society Digest 16, pp. 502–554.

J.C. Ordóñez-Calderón et al.

page 14

- Kelemen, P.B., Hanghøj, K., Greene, A.R., 2003. One view of the geochemistry of subduction-related magmatic arcs, with emphasis on primitive andesite and lower crust. In Holland, H.D., Turekian, K.K. (Eds.), *Treatise on geochemistry, Volume 3*. Amsterdam, Elsevier, pp. 593–659.
- MacLean, W.H., Barrett, T.J., 1993. Lithogeochemical techniques using immobile elements. *Journal of Geochemical Exploration* 48, 109–133.
- McLennan, S.M., 2001. Relationships between the trace element composition of sedimentary rocks and upper continental crust. *Geochemistry, Geophysics, Geosystems* 2 (Paper number 2000GC000109).
- Mateu-Figueras, G., 2003. *Models de distribució sobre el simplex*. Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.
- Maydagán, L., Franchini, M., Lentz, D., 2016. Phyllosilicates geochemistry and distribution in the Altar porphyry Cu-(Au) deposit, Andes Cordillera of San Juan, Argentina: Applications in exploration, geothermometry, and geometallurgy. *Journal of Geochemical Exploration* 167, pp. 83–109.
- Mungall, J.E., 2014. Geochemistry of Magmatic Ore Deposits. In Scott, S.D. (Ed.), *Treatise on geochemistry second ed.*, Volume 13. Amsterdam, Elsevier, pp. 195–218.
- Ordóñez-Calderón, Lafrance, B., Gibson, H.L., Schwartz, T., Pehrsson, S.J., Rayner, N.M., 2016. Petrogenesis and Geodynamic Evolution of the Paleoproterozoic (~1878 Ma) Trout Lake Volcanogenic Massive Sulfide Deposit, Flin Flon, Manitoba, Canada. *Economic Geology* 111, pp. 817–847.
- Ordóñez-Calderón, J.C., Polat, A., Fryer, B.J., Gagnon, J.E., Raith, J.G., Appel, P.W.U., 2008. Evidence for HFSE and REE mobility during calc-silicate metasomatism, Mesoarchean (~3075 Ma) Ivisaartoq greenstone belt, southern West Greenland. *Precambrian Research* 161, pp. 317–340.
- Palarea-Albaladejo, J., Martín-Fernández, J.A., Buccianti, A., 2014. Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *Journal of Geochemical Exploration* 141, pp. 71–77.
- Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. zCompositions-R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 143, pp. 85–96.
- Pawlowsky-Glahn, V., Egozcue, J. J., 2001. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* 15 (5), pp. 384–398.
- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. *Modeling and analysis of compositional data*. John Wiley & Sons, Ltd., p. 247.
- Rasmussen, J.C., Hoag, C., Horstman, K.C., 2012. Geology of the Northern Santa Rita Mountains, Arizona. *Arizona Geological Society Fall Trip*, p. 24.
- Taylor, S.R., McLennan, S.M., 1985. *The continental crust: Its composition and evolution*. Oxford, Blackwell, p. 312.
- van den Boogaart, K.G., Tolosana-Delgado, R., 2013. *Analyzing Compositional Data with R*. Springer, p. 258.
- van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2015. *R Package “compositions”, Version 1.40-1. Compositional Data Analysis*, p. 264.

Balance designs revisit indices commonly used in agricultural science and eco-engineering

Running title: Compositional balances in agronomy and eco-engineering

Parent, Serge-Etienne and Parent, Léon E.

Department of Soils and Agrifood Engineering, Université Laval, Québec G1V 0A6, Canada

Table of contents

1	Abstract	2
	Keywords	2
	Abbreviations.....	2
2	Introduction.....	2
3	Compositional data analysis and log ratios.....	3
4	Soil compositions as balance systems.....	6
4.1	Soil quality	6
4.1.1	General soil compositions.....	6
4.1.2	Soil aggregation	6
4.1.3	Soil biogeochemistry	7
4.2	Soil testing.....	9
4.2.1	SLAN	10
4.2.2	BCSR	10
4.2.3	NIBC	12
5	Tissue testing	13
5.1	Early tissue nutrient interpretation methods.....	14
5.2	Weighted nutrient diagnosis.....	14
5.3	Nutrient ratios.....	15
5.4	Diagnosis and Recommendation Integrated System (DRIS)	15
5.5	Compositional data analysis of tissue compositions	17
5.6	Ionomics	18
5.7	Elaboration of nutrient standards	19

6	Animal nutrition.....	22
7	Conclusion	23
8	Acknowledgements.....	23
9	References.....	23

1 Abstract

Most agronomists and eco-engineers diagnose the interrelated soil, plant and feed compositions in agro-ecosystems in terms of total concentrations, proportions and dual ratios. However, the pre-compositional tools developed in the 20th century to diagnose compositions do not account for the special properties of strictly positive and intrinsically multivariate compositional data, i.e. data bounded to measurement scale or unit. Balance designs can provide a coherent understanding of the relationships between the components of complex soil-plant-animal systems and sub-systems. Balances are illustrated by a mobile design, where isometric log ratios (*ilr*) are computed at fulcrums from concentrations located in buckets. Power parameters are useful to normalize the distribution of *ilr* data and to regulate component accessibility in biological, physical and chemical systems. Our objective is to revisit common diagnostic indices in agricultural science and engineering using compositional balances and power parameters. Revisited soil indices are the ternary diagram, mean weight diameter, sufficiency level of available nutrients, basic cation saturation ratios, and nutrient intensity and balance concept. Revisited plant indices are the critical concentration ranges, dual ratios, stoichiometric ratios, the Kenworthy index, and the Diagnosis and Recommendation Integrated System. Revisited feed quality indices are the dietary cation-anion difference and the K/(Ca+Mg) ratio. Rather than consider compositions in terms of their isolated parts, agronomists and eco-engineers are urged to think in terms of interactive balance systems. As is the case for other disciplines, research in agronomy can benefit from compositional tools.

Keywords: Box-Cox power transformation, compositional data, fractal-like coefficient, Freundlich kinetics, hydroponics, hypocalcemia, hypomagnesemia, log ratio, nutrient interaction, nutrient solution, organic matter, phosphate sorption, soil and tissue testing

Abbreviations: *alr*, additive log ratio; BCSR, basic cation saturation ratios; *clr*, centred log ratio; *DCAD*, dietary cation-anion difference; *DPS*, degree of soil phosphate saturation; DRIS, Diagnosis and Recommendation Integrated System; *ilr*, isometric log ratio; *MWD*, mean weight diameter; NIBC, nutrient intensity and balance concept; OM, organic matter; SLAN, sufficiency level of available nutrients.

2 Introduction

More than 95% of human food is produced on soil (FAO, 2015). Soil quality, defined by chemical, physical, and biological attributes (Doran et al., 1996), affects crop productivity (Carter, 2002; Mueller et al., 2010), resistance to erosion (Eash et al., 1994),

crop nutrient requirements (Lundy et al., 2015), animal nutrition (Voisin, 1961), and human health (Brevik and Burgess, 2013). The United Nations High-level Panel on Threats identified environmental degradation as a major threat to food security (United Nations, 2004). Wise resource management requires a coherent understanding of properly balanced soil, plant, and animal feed compositions. Accordingly, various methods have been developed to diagnose soil quality, plant nutrition, and feed quality.

Most agronomists and bioengineers diagnose soil, plant, and feed compositions in terms of concentrations, proportions, and dual ratios. However, current indices ignore the special properties of the strictly positive and intrinsically multivariate compositional data – that is, data that are bounded to some measurement scale or unit, causing “resonance” between components within systems closed to, e.g., unity or 100% (Aitchison, 1986). Due to closure to the whole, one component of the system is redundant because it can be computed by the difference between the whole, such as measurement scale or unit, and the sum of the other components. Moreover, depending on the measurement scale, such as soil fresh mass, dry mass, maximum sorption and exchange capacity, or total elemental content, the interpretation of the results may differ, leading to subcompositional incoherence. Finally, confidence intervals may be improperly defined outside the compositional space (<0% or >100%).

Aitchison (1986) proposed using the additive and centered log-ratio transformations (*alr* and *clr*, respectively) to conduct statistical analyses on compositional data. Compositional log-ratio transformation techniques were pioneered by McBratney et al. (1992) in soil science and by Parent and Dafir (1992) in plant nutrition. Egozcue et al. (2003) and Egozcue and Pawlowsky-Glahn (2005) developed a concept of balances computed as the isometric log ratios (*ilr*). Filzmoser et al. (2009) recommended using the balance concept to conduct statistical analyses on environmental data. Parent (2011) proposed using nutrient balances to diagnose plant nutrients. Parent et al. (2012c) illustrated balance designs by a mobile setup with fulcrums and buckets. Parent et al. (2013b) defined ionome groups among wild and domesticated plant species using balances. Parent et al. (2014) modeled soil phosphorus (P) biogeochemistry using balances. This view was found promising by Baxter (2015) in ionomics and Némery et al. (2016) in biogeochemistry.

The objective of this review is to revisit current soil, plant and feed diagnostic indices using compositional balance and power parameters. We urge agronomists and bioengineers to change the current paradigm and think in terms of interactive balance systems rather than sets of isolated properties.

3 Compositional data analysis and log ratios

A composition comprises two or more components that are intrinsically related to each other within a closed system. For example, there are complementary hours of day and night, which total exactly 24 h and interchange all year round. Closure to the whole imposes three constraints. (1) One component, computed by the difference from the

measurement unit or scale, is redundant, leading to at least one negative correlation (i.e., if one proportion increases, at least one other proportion must decrease). Therefore, there are $D-1$ degrees of freedom in a D -part composition (Aitchison and Greenacre, 2002). (2) Components are scale-dependent and subcompositionally incoherent upon scale change (e.g., dry or fresh mass, P fractions expressed on the basis of total P), leading to spurious correlations (Chayes, 1960; Pearson, 1897; Tanner, 1949). (3) The data distribution, constrained to the compositional rather than the real ($\pm\infty$) space, is inherently non-normal because the confidence intervals may scan beyond the compositional space (<0 or $>100\%$), which is conceptually absurd (Diaz-Zorita et al., 2002).

Aitchison (1986) developed the idea of compositional data analysis and simplex geometry using ternary diagrams, a concept developed in the 18th century as Mayer's color triangle and introduced in sedimentology in the early 20th century (Howarth, 1996). He proposed using the *alr* and *clr* transformations before conducting statistical analysis on compositional data.

The *alr* transformation is computed as follows:

$$alr_i = \ln(x_i/x_D),$$

Equation 1

where alr_i is the $(D-1)^{th}$ *alr* among D components, x_i is the i^{th} component, and x_D is the D^{th} component, often called the filling value (F_v) between the full composition and the sum of quantified components (although any component could be used as denominator). F_v closes the composition to the measurement unit or scale and is computed as follows:

$$F_v = 100\% - \sum_{i=1}^{D-1} x_i,$$

Equation 2

where x_i is the i^{th} of the $D-1$ components. F_v allows the back-transformation of log ratios into familiar units. There are $D-1$ *alr* variables in a D -part composition. The *alr* returns results close to the ordinary log transformation if the filling value, used as common scale (x_j), is very large compared to other components. Because the *alr* geometry is oblique, its use for distance-based statistics is not recommended (Egozcue and Pawlowsky-Glahn, 2005).

The *clr* transformation is computed as follows (Aitchison, 1986):

$$clr_i = \ln(x_i/g[x]),$$

Equation 3

where x_i is i^{th} component, and $g[x]$ is the geometric mean across components of the composition vector x . The geometric rather than the arithmetic mean is the most appropriate central values for ratios (Fleming and Wallace, 1986). Compared to ordinary log transformation, geometric means $g[x]$ and $g[y]$ for the x and y compositional vectors must be close to each other to yield $clr_{x1} - clr_{y1} \approx \ln(x_1) - \ln(y_1)$ (Lovell et al., 2011).

Because the *clr* returns D variables, the inverse of its covariance matrix is singular, hence requiring to drop one *clr* variable, generally F_v . The *clr* has Euclidean geometry (Egozcue and Pawlowsky-Glahn, 2005), allowing computation of multivariate distances (ϵ) between two compositions as follows:

$$\epsilon = \sqrt{\sum_{i=1}^D (clr_i - clr_i^*)^2}$$

Equation 4

where clr_i and clr_i^* are the i^{th} *clr* values of the first composition and reference composition, respectively.

Even after construction of the *alr* and *clr* transformations, there was a need to develop a log-ratio expression with Euclidean geometry and returning the $D-1$ degrees of freedom of a D -part composition (Aitchison and Greenacre, 2002). Orthonormal balances can provide $D-1$ degrees of freedom due to orthogonality between log contrasts. Orthogonality is a special case of linear independence whereby vectors are at perfectly right angles to one another (Rodgers et al., 1984). Components can be arranged into orthonormal balances following a sequential binary partition that contrasts nonoverlapping subsets of components at numerator and denominator. Log ratios of subsets of components are computed as follows (Egozcue et al., 2003):

$$ilr_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \left(\frac{\sqrt[r_k]{\prod_{i=1}^{r_k} x_i}}{\sqrt[s_k]{\prod_{j=1}^{s_k} x_j}} \right),$$

Equation 5

where r_k and s_k are numbers of components in subsets of the numerator or denominator, respectively; i and j refer to components of the numerator and denominator, respectively; and $\sqrt[r_k]{\prod_{i=1}^{r_k} x_i}$ and $\sqrt[s_k]{\prod_{j=1}^{s_k} x_j}$ are geometric means of components of the numerator and denominator, respectively. Balances are reported as:

[subset at denominator | subset at numerator]

because the log ratio gets more negative as the denominator loads more in the log ratio, and inversely. In algebra, more negative values are located on the left-hand side of a vector, and more positive values are located on the right-hand side. In the mobile design (Parent et al., 2012c), balances are computed at fulcrums using the proportion or concentration values, which are located in buckets.

Because balances are orthogonal to each other, they return the same multivariate distance regardless of the balance arrangement among components. Although there are $D!(D-1)!/2^{D-1}$ possible combinations of $D-1$ orthonormal balances in a D -part composition (Pawlowsky-Glahn et al., 2011), some balances are more meaningful than

others and are thus useful to interpret the results. Subsets of components can be arranged by following some *ad hoc* theory or hypothesis, exploratory biplot analysis, for management purposes, or simply at random without any prior *ad hoc* arrangement.

Modern free software, such as R and the `compositions` package (van den Boogaart et al., 2014), Python and the `scikit-bio` package (Morton et al., 2017), CoDaPack (Comas-Cufí and Thió-Henestrosa, 2011), and Orange Data Mining (Demšar et al., 2013), can assist in performing numerical operations to provide unbiased solutions to problems related to the compositional space. Most operations can also be run using the Excel package.

4 Soil compositions as balance systems

4.1 Soil quality

4.1.1 General soil compositions

A soil textural diagram includes sand (0.053–2 mm), silt (2–53 µm), and clay ($\leq 2 \mu\text{m}$) fractions and could be expanded to include water and OM (Figure 1). Sand and silt may be further subdivided into fine, medium, and coarse fractions. Gravel fractions ($> 2 \text{ mm}$) may also be added. Pedotransfer functions relating soil properties to soil texture can be elaborated by using orthonormal balances rather than raw proportions of components. Xu et al. (2017a) added cementing agents in studies on the compaction of coarse-textured soils. Compositions may be also expressed on a volume basis.

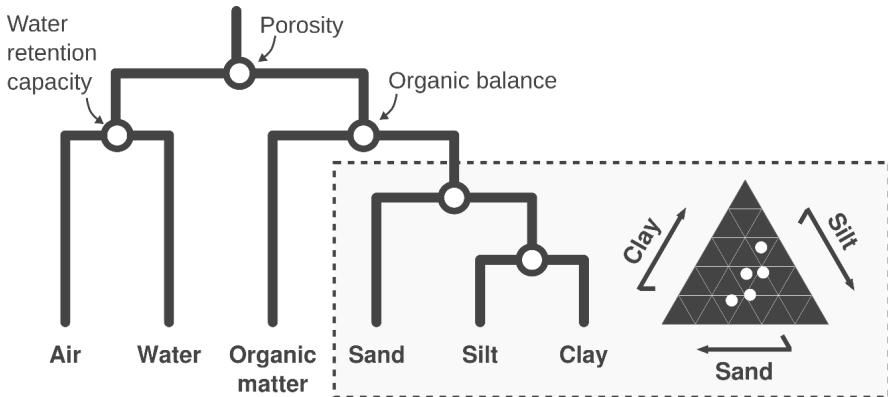


Figure 1. Basic soil constituents arranged into four balances in a mobile design.

4.1.2 Soil aggregation

A soil aggregate is made of closely packed mineral and organic particles. Aggregation provides an index of soil physical quality (Parent et al., 2012a). Soil aggregation, quantified as the distribution of aggregate-size fractions after sieving, is commonly measured as the mean weight diameter (*MWD*) computed as follows (van Bavel, 1949):

$$MWD = \sum_i^D \bar{x}_i w_i,$$

Equation 6

where \bar{x}_i is the average aggregate diameter of the i^{th} fraction, estimated as the mean value between two consecutive sieve openings; and w_i is the aggregate fraction retained on the i^{th} sieve. The sand fraction ($>53 \mu\text{m}$) is subtracted from the total soil mass on each sieve to estimate the aggregate mass retained on each sieve (Kemper and Rosenau, 1986). Gardner (1956) found that the geometric mean diameter uniquely characterized the aggregate distribution. A minimal set of five sieve-size fractions was needed to draw the cumulative particle-size distribution function (Jelinek, 1970). Soil aggregation has been assessed by using the fragmentation fractal dimension (D_f), based on optimistic assumptions on aggregate mean diameter, bulk density, and shape (Rieu and Sposito, 1991; Logsdon, 1995; Anderson et al., 1998; Parent et al., 2012a).

A balance that contrasts micro- ($<250\text{-}\mu\text{m}$) and macro- ($>250\text{-}\mu\text{m}$) aggregates avoids biases and cumbersome assumptions and allows comparisons between studies (Parent et al., 2012a, Xu et al. 2017b). The balance design of aggregates could be expanded to include several classes of macro-aggregates, clay-silt-humus associations ($<53 \mu\text{m}$) within micro-aggregates (Hassink and Whitmore, 1997), and sand-size fractions that are additive to the aggregate-size fractions on each sieve, to account for the dilution of aggregates in the entire soil mass.

4.1.3 Soil biogeochemistry

4.1.3.1 Carbon

Soil is a huge reservoir of organic carbon (C), three times as large as the vegetation of terrestrial ecosystems and twice that of the atmosphere (Zhi et al. 2014). Soils can mitigate climate change through carbon sequestration. The decomposition of native and exogenous organic matter in soils depends on biological, chemical and physical mechanisms protecting the carbon pools (C_{pool}) against microbial attacks (Stewart et al., 2008). Exogenous C sources added to soil can be fractionated into soluble substances, hemicellulose, cellulose and lignin plus cutin using the routine van Soest method, and arranged into balances also including products of decomposition (Figure 2).

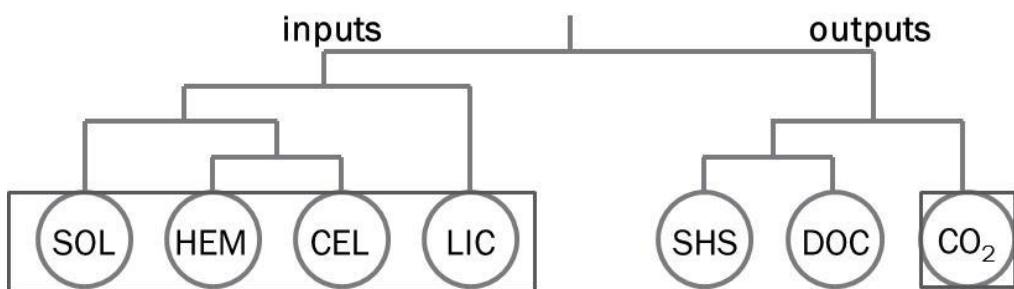


Figure 2. Balance arrangement of biochemical components of the C sources as soluble substances (SS), hemicellulose (H), cellulose (C) and lignin plus cutin (L), on the one side, and products of decomposition as CO_2 , dissolved organic carbon (DOC) and humic substances (HS), on the other. The SS, H, and C fractions are more labile than the L, HS, and DOC fractions. Components within squares are most often modeled.

After incorporation into soil, however, only the soluble and holocellulose (hemicellulose plus cellulose fractions) could be extracted and balanced (Parent and Parent, 2015). Because perturbation and powering operations are difficult to conduct, OM decomposition in soil is generally related to the initial biochemical composition of the C sources. Thuriès et al. (2001, 2002) conducted a 180-d soil incubation experiment in which they varied C sources under non-limiting supply of water and air. They related the size of the labile carbon pool to the van Soest biochemical components. Arranging the van Soest biochemical fractions into balances in the Thuriès et al. (2001, 2002) studies, Parent S et al. (2012c) found that the size of easily decomposable labile C was more closely related to balances ($R^2 = 0.924$) than to raw concentrations ($R^2 = 0.856$).

4.1.3.2 Stoichiometric ratios

The determination of nutrient dynamics requires the quantification of several pools and subpools (Stevenson, 1986). Components of the soil C, nitrogen (N), P, and sulfur (S) cycles have been modeled separately or as C:N, N:P, C:P (Delgado-Baquerizo et al., 2013), and C-N-S-P ratios (Stevenson, 1986). However, global relationships captured by balance designs have been missing from these analyses.

The C/N ratio

Parent et al. (2000), Duguet et al. (2006), and Quinche-Gonzalez et al. (2016a, 2016b) developed a soil-test N for organic soils (Histosols) that combines total C, total N, and F_v , the filling value to 100% composition. The classical C/N ratio was represented by the [N | C] balance and OM dilution in the soil mass by the [F_v | C, N] balance. Soil-test diagnosis and calibration of N dosage against crop yield were conducted successfully by using benchmark organic soils with high N-mineralization potential.

Balances in the P cycling

Overfertilization with P is a major contributor to the eutrophication of surface waters (Sims et al., 1998). Sorption of P by the heterogeneous soil surfaces has thus been studied extensively. Sorption studies are generally conducted after drying and grinding soils to <2 mm, a procedure that destroys soil aggregates and enhances the surface area of particles and reactivity per unit surface (Kopelman, 1986, 1988) by destroying soil macro-aggregates. Linquist et al. (1997) showed that the kinetics of P sorption in soils is controlled by soil aggregation because phosphate sorption occurs only within a reactive volume near the surface of soil aggregates. On the other hand, the amount of phosphate ions adsorbed on the solid phase is generally determined by the difference between added P and the P remaining in solution after reaching a given equilibrium period. In the sorption isotherm models, the relationship between the P retained by the solid phase and solution P is thus distorted by an intrinsically negative correlation between the two complementary variables (Parent and Bélanger, 1985).

In nature, most reactions occur on low-dimensional, heterogeneous surfaces (Kopelman, 1986). Powering is useful to compute effective concentrations in reactive volumes in fractal biological (Savageau, 1995) and chemical (Kopelman, 1988) systems. Chardon and Blaauw (1998) assigned fractal-like (power) coefficients to phosphate ion

concentration and time, to describe the Freundlich kinetics of phosphate sorption in soils. The distribution coefficient of phosphate ions between the solid and liquid phases provides an alternative simple *alr* expression to monitor P sorption that avoids the intrinsic negative correlation between added phosphate ions distributed between the liquid and the solid phases. Partitioning added phosphate between the solid (C_s) and the liquid (C_l) phases, we obtained a logistic variable, as follows:

$$\ln\left(\frac{C_s}{C_l}\right) = \ln(vt^n)$$

Equation 7

where v and n are fitted parameters, and $C_s + C_l = C_{added}$. The power parameter n assigned to t reflects the limited access of phosphate ions to the soil reactive volume, as determined by aggregation and soil P saturation, and mediated by agricultural practices. The theoretical maximal sorption rate occurs at $n = 1$ ($n = 1 - h$), where soil P saturation is close to zero reactive volume is maximal, i.e. there is full access to sorption sites and soil particles are completely dispersed. Additional P forms can be included in equation 7 by expanding the logistic to *ilr* variables. The *ilr* avoids subcompositional incoherence, such as the occurrence of P fractions expressed relative to the total P or soil dry mass (Parent et al., 2014; Abdi et al., 2015).

P indices have been derived from the relationship between active phosphate forms and soil sorption sites. The degree of soil phosphate saturation (*DPS*) was computed on a molar basis as the ratio of acid oxalate-extracted phosphate (P_{OX}) and oxi-hydroxides of Al (Al_{OX}) and Fe (Fe_{OX}), as follows (Breeuwsma and Silva, 1992):

$$DPS = \frac{P_{OX}}{\alpha_m(Al_{OX} + Fe_{OX})}$$

Equation 8

where α_m is the maximal saturation factor for total sorption (0.5 for mineral soils and 0.4 for organic soils). The amalgamation of Al_{OX} and Fe_{OX} reflects P sorption capacity (*PSC*). A logistic variable of two components in *DPS* within the closed *PSC* space could be computed as follows:

$$alr_{DPS} = \ln\left(\frac{P_{OX}}{PSC - P_{OX}}\right)$$

Equation 9

Equations 8 and 9 would produce comparable results if *PSC* $\gg P_{OX}$. *DPS* has been replaced by phosphate saturation ratios to run routine methodologies for fertilizer recommendations and environmental regulation (Guérin et al., 2007; Khiari et al., 2000; Leblanc et al., 2013; Pellerin et al., 2006a, 2006b).

4.2 Soil testing

Soil testing for P and potassium (K) management is a product of agricultural research urged by the growing availability of processed fertilizers (Peck, 1990). Soil-test methods

for available nutrients were developed after users recognized the fallacy of applying total elemental analysis to support decisions on fertilizer recommendations. Several jurisdictions proposed the use of mild chemical extraction methods to interpret soil tests (Peck, 1990) and determine the optimal dosage of fertilizers (Dahnke and Olson, 1990; Geraldson, 1984; McLean, 1984). Soil-test interpretation methods include the sufficiency level of available nutrients (SLAN), basic cation saturation ratio (BCSR), and nutrient intensity and balance concept (NIBC).

4.2.1 SLAN

In SLAN, soil-test P and K are scaled on a soil dry mass or scooped-volume basis, and are interpreted from well-planned field experiments. If P or K is the growth-limiting factor and other factors are at optimal or sufficient levels (De Wit, 1992; Nelson and Anderson, 1984), then crop yield must increase with soil-test P or K in control plots (zero fertilization) across trials until a yield plateau is attained. Crop response curves may be linear-plateau, polynomial, or exponential according to the Mitscherlich's law of diminishing returns (Bray, 1958; Dahnke and Olson, 1990; Prevot and Ollagnier, 1961). Crops growing in soils below critical soil-test P or K require fertilization to overcome P or K limitations.

The *alr* expression for the soil test to run SLAN can be scaled on F_v , computed by the difference between the nutrient concentration and measurement unit (kg of soil dry mass), as follows:

$$alr_{P \text{ or } K} = \ln\left(\frac{P \text{ or } K}{F_v}\right)$$

Equation 10

Because $F_v \gg$ P or K concentration, results of soil-test calibration are practically the same as using either *alr* or the raw or log-transformed P or K concentration. However, SLAN does not account for biogeochemical interactions between soil nutrients.

4.2.2 BCSR

BCSR requires quantification of the soil exchangeable cations K, calcium (Ca), and magnesium (Mg), which are usually determined by a suitable soil test, and the exchangeable acidity (aluminum [Al] and hydrogen [H]), which is most often determined by using a buffer-pH solution (Quaggio et al., 1985). Cation exchange capacity (CEC) is the sum of basic cations and exchangeable acidity expressed on a molar basis. BCSR computes the percentage of soil CEC saturated by K, Ca, Mg, and exchangeable acidity. Cation saturation ratios are computed as K/CEC, Ca/CEC, and Mg/CEC. Diagnostic dual ratios are computed as the K/Ca, K/Mg, and Ca/Mg molar ratios. BCSR has been used to guide K fertilization (McLean, 1984) and liming practices (Quaggio et al., 1985) of agricultural soils and to relate soil properties to forest decline attributed to acid rains (Ouimet et al., 2013; Ouimet and Camiré, 1995).

However, the BCSR has been associated with several problems. (1) Although spectroscopic methods show good potential to determine CEC in temperate soils

(Leblanc et al., 2016), CEC is difficult to estimate accurately in most nontropical soils (Quaggio et al. 1985). (2) There are $D \times (D - 1)/2$ computable dual nutrient ratios in D -part compositions, but $D - 1$ degrees of freedom are available for modeling purposes. For three cations, one ratio is redundant (e.g., $K/Mg = (K/Ca) \times (Ca/Mg)$). (3) BCSR fertilization recommendations are based on a conceptualized ideal soil, which is not supported by field trials (Liebhardt et al., 1981; Koppitzke and Menzies, 2007). (4) The cationic system (in $\text{mmol}_c \text{ kg}^{-1}$) is bound to CEC, which is a soil subsystem diluted into the soil mass. The use of a fixed saturation percentage target as a guide for K fertilization leads to under- or over-fertilization of low- or high-CEC soils, respectively.

Problems 2 and 4 can be handled by using balances. For problem 2, relationships between cationic species can be expressed as balances between K and divalent cations Ca and Mg to assist in K fertilization decisions, between divalent cations and exchangeable acidity to assist in liming decisions, and between Ca and Mg to assist in selection of proper liming materials (e.g., calcareous to dolomitic limes), as follows: $[Ca, Mg | K]$, $[Al, H | Ca, Mg]$, and $[Mg | Ca]$. For problem 4, cationic species can be contrasted on a mass rather than a molar basis to account for nutrient dilution in the whole soil mass (Figure 3).

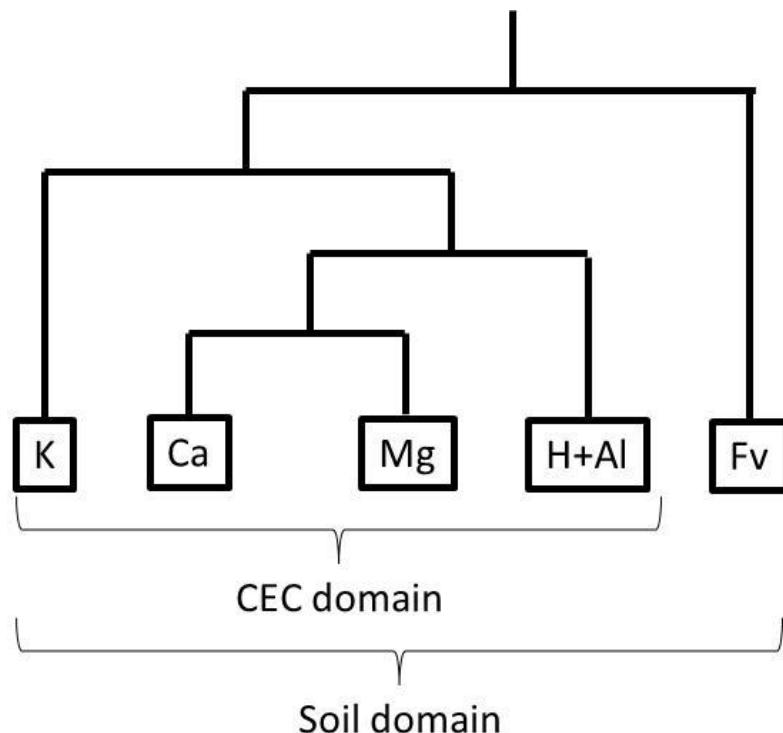


Figure 3. Domain of the cationic exchangeable complex (cations K, Ca, and Mg; soil acidity H+Al) embedded into the soil mass that includes a filling value (Fv).

Parent et al. (2012c) and Montes et al. (2016) found that guava (*Psidium guajava*) response to K fertilization was improved by using the [Ca, Mg | K] balance compared to K alone (SLAN) in Brazilian soils when CEC showed only small variations (Figure 4). The objective of liming guava soils was to reach 70% base saturation.

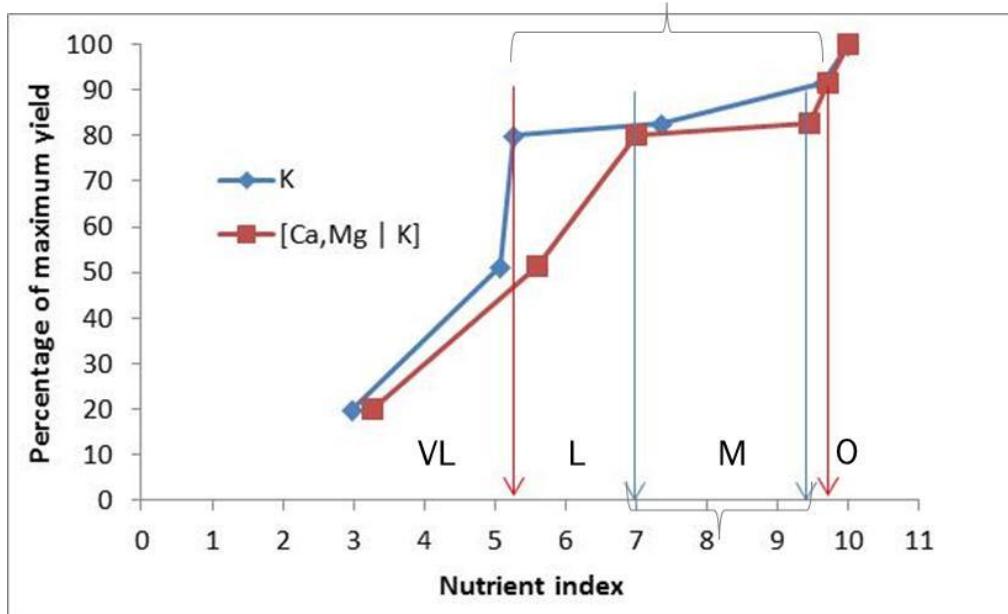


Figure 4. Soil fertility classes delineated by step of approximately two index units (VL = very low, < 50% maximum yield; L = low, 50-75%; M = medium, 75-80%; O = optimum, 80-100%) using K alone or the K balance [Ca,Mg / K]. The balance facilitates classification. Source: Parent et al. (2012c).

4.2.3 NIBC

Early studies on cationic balances in soil and hydroponic nutrient solutions focused on Ca disorders that affected the quality of horticultural crops (Carter and Webster, 1990; Geraldson, 1984). The Ca molar ($\text{mmol}_c \text{ L}^{-1}$) ratio was computed as follows:

$$\text{Ca molar ratio} = \frac{\text{Ca}^{2+}}{\text{Ca}^{2+} + \text{Mg}^{2+} + \text{K}^+ + \text{Na}^+}$$

Equation 11

Concentrations of sodium (Na), K, Ca, Mg, nitrate as nitrogen ($\text{NO}_3\text{-N}$), and chlorine (Cl) in solution, extracted under vacuum after saturating soil with distilled water to dissolve the salts, can be expressed in mg L^{-1} or mg kg^{-1} dry soil.

In soil and hydroponic nutrient solutions, the sum of cationic (+) and anionic (-) charges is constrained to be zero, hence plant nutrition research deals with bounded anionic-cationic mixtures (Schrevens and Cornell, 1993). However, plant nutrition studies where nutrient forms or concentrations are varied in nutrient solutions (e.g., Zhong et al., 2014) still disregard the inherent relationships between anionic and cationic forms of soluble

nutrients that sum to zero. The balance setup in Figure 5 provides a coherent understanding of the ever-changing nutrient mixtures in soil solutions and hydroponics as plant takes up nutrients and solution is renewed.

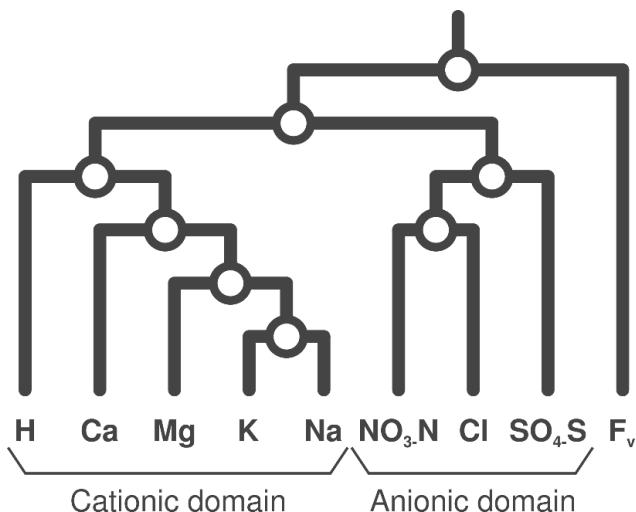


Figure 5. Cationic and anionic domains of the nutrient intensity and balance concept.

Compositional data analysis has been used to address mixture systems successfully in geochemistry (Buccianti et al., 2005; Buccianti and Pawlowsky-Glahn, 2005; Engle and Blondes, 2014; Grunsky et al., 2013) and plant nutrition studies (Lopez et al., 2002; Parent et al., 1997).

5 Tissue testing

Tissue testing involves determination of the total, soluble, and extractable fractions of elements (e.g., C, N, P, K, Ca, Mg, S) from a diagnostic tissue sample collected at a certain stage of plant development (Munson and Nelson, 1990). The scale of measurement is dry or fresh matter or sap liquid, and units of measurement are %, g kg⁻¹, mg kg⁻¹, µg kg⁻¹, g L⁻¹, mg L⁻¹, and µg L⁻¹. The plant nutrient concentration is thought to integrate all growth-influencing genetic and environmental factors (Munson and Nelson, 1990). Genetic factors establish the physiological and metabolic potentials, while environmental factors allow crops to reach a percentage of their genetic potential. Regional differences in compositions reflect variations in nutrient supply due to soil and climate conditions or poor adaptation of the species or cultivar rather than physiological requirements (Kenworthy, 1967). Moisture shortage and low temperature may limit uptake of nutrients, especially those that move slowly to the roots, producing nutrient shortage in the leaf even though soil nutrient supply appears to be adequate (Barber, 1995). Several interpretation methods have been elaborated.

5.1 Early tissue nutrient interpretation methods

Tissue nutrient diagnosis emerged in the mid-1930s with the rapid development of sampling and interpretation methods. Tissue sampling was a critical step because tissue nutrient concentration depends on sample position and sampling period (Bould et al., 1960). The hypothesis behind the interpretation of plant analysis is that a response curve can be derived from the relationship between yield and tissue nutrient composition. All other nutrients are maintained within adequate concentration ranges to avoid altering the effect of the target nutrient on yield by any harmful effect of another nutrient (Ulrich and Hills, 1967). The underlying assumption is that all factors but the ones being varied are equal or proportional under the law of the minimum, or that all factors but the ones being varied are in sufficient amounts under the law of the optimum (De Wit, 1992).

The earliest tissue test interpretation method was illustrated by a hand-drawn ternary diagram mapping relationships among three nutrients to guide fertilization (Lagatu and Maume, 1934). Commenting on 19th-century research on the minimum percentage and threshold optimum, Macy (1936) elaborated the concept of critical nutrient concentration. The plant was proposed to engage in luxury consumption of nutrients above this critical concentration and poverty adjustment below this concentration. Later, critical concentration values were set at 90–95% of the maximal yield by using crop response curves (Ulrich, 1952). A sharp calibration curve was obtained by using recently matured leaves to establish critical nutrient levels between deficiency and sufficiency. Petiole tissue and sap were also useful for nitrate ion, phosphate ion, K, and Cl analyses (Ulrich and Hills, 1967). Above the “sufficiency” plateau, the plant engaged in nutrient luxury consumption, accumulation, or contamination and nutrient accumulated without yield loss. This period was followed by generally elusively defined thresholds of nutrient excess, above which the yield decreased again due to nutrient antagonism or toxicity.

Whereas critical concentration ranges were established at a fixed plant developmental stage, tissue nutrient concentrations change with time and biomass accumulation. Justes et al. (1994) accounted for N dilution in the whole plant biomass of agricultural crops during the growing season using allometric relationships between N concentration and biomass accumulation. These relationships were called critical N curves. Timmer (1997) addressed the relationship between nutrient concentrations and seedling biomass production.

However, all of the above diagnostic methods, except those described by Lagatu and Maume (1934), ignored nutrient physiological interactions (Bates, 1971). Holland (1966) suggested running multivariate analyses across concentration values, even though concentration values are interrelated due to nutrient interactions and data redundancy, and correlations may be spurious. Alternative methods have been suggested.

5.2 Weighted nutrient diagnosis

Kenworthy (1967) claimed that nutrient concentrations should be scaled as percentages of standard values from high-performing crops. Standardized concentrations are weighted by the coefficient of variation V to derive the balance index B , as follows:

If $X < Std$: $(X/Std) \times 100 = P$; $(100 - P) \times (V/100) = I$; $P + I = B$

Equation 12

If $X > Std$, $(X/Std) \times 100 = P$; $(P - 100) \times (V/100) = I$; $P - I = B$

Equation 13

where X is nutrient concentration in tissue sample, Std is the standard nutrient concentration used to convert compositions into the percentage of the standard, P is the percentage of the standard value of normal surveyed plants, and I is the influence of variation in standard values. Equations 12 and 13 are symmetrical about 100 to allow diagnosis as relative shortage (<100) or excess (>100). Although these equations have the novel feature of using the symmetry concept to address numerically the relative nutrient shortage or excess, they do not account for nutrient interactions.

5.3 Nutrient ratios

Nutrients interact in plants (Wilkinson, 2000) and must be properly balanced (Baeyens, 1969). Interactions between two nutrients that are near critically deficient levels are important in cases where one nutrient competes or dilutes another (Marschner, 1986). Interactions are usually expressed as dual (Bergmann, 1988; Walworth and Sumner, 1987) or stoichiometric (Ingestad, 1987) ratios. These ratios can be functional, such as the N/P ratio that reflects protein synthesis and plant-available energy (Güsewell, 2004; Redfield, 1934), or operational, such as those proposed by Ingestad (1987) to scale plant nutrients on N and guide conifer seedling fertilization. Nutrient ratios are often reported assuming wrongly that pairs of elements are linearly related (Kenworthy, 1967). In addition, optimal dual ratios should not be considered alone because they could be obtained within the deficiency or toxicity range (Marschner, 1986). Meanwhile, the number of $D \times (D - 1)/2$ dual ratios computable from D -part compositions exceeds the $D - 1$ number of available degrees of freedom (Aitchison and Greenacre, 2002). Thus, most information derived from $D \times (D - 1)/2$ nutrient ratios is redundant and cannot be deciphered easily.

5.4 Diagnosis and Recommendation Integrated System (DRIS)

Ideas in the preceding sections were combined into an integrative approach, the DRIS, which uses dual ratios at high-yield level to generate nutrient standards as dual ratio means and variances (Beaufils, 1973; Walworth and Sumner, 1987). Dual ratios are weighted by their respective coefficients of variation and integrated into DRIS functions and nutrient indices, ordered from the most negative (trend towards shortage) to the most positive (trend towards excess) to facilitate interpretation (Beaufils, 1973). DRIS functions are computed as follows (Walworth and Sumner, 1987):

$$f(A/B) = [\frac{(A/B)}{(a/b)} - 1] \frac{\kappa}{cv}, \text{ if } (A/B) > (a/b),$$

Equation 14

$$f(A/B) = \left[1 - \frac{(a/b)}{(A/B)}\right] \frac{\kappa}{cv}, \text{ if } (A/B) < (a/b),$$

Equation 15

$$f(A/B) = 0, \text{ if } (A/B) = (a/b),$$

Equation 16

where A/B and a/b are dual ratios between nutrients A and B in the given and selected reference composition for crops at high-yield level, respectively; and cv is the coefficient of variation (standard deviation divided by the mean) of the dual ratio in the reference composition. Because the concentrations are reported in different units, DRIS norms depend on the unit of measurement. The factor κ is used to ensure that the same number of digits is reported. Equations 14 and 15 are intended to express the symmetry of ratio expressions (X/Y or Y/X), the choice of which depends on the highest variance ratio between dual ratios of the low- and high-yielding subpopulations. However, the selection of high yielders is questionable because high yielders showing nutrient luxury consumption, accumulation, or contamination are included in the step of computing DRIS norms.

DRIS indices I_A , I_B , and I_C are computed across nutrients A, B, and C, respectively, by averaging DRIS functions after multiplying DRIS functions by (+1) if the nutrient is in the numerator or (-1) otherwise, as follows (Walworth and Sumner, 1987):

$$I_A = \frac{f(A/B)+f(A/C)}{2}; I_B = \frac{-f(A/B)-f(C/B)}{2}; I_C = \frac{-f(A/C)+f(C/B)}{2},$$

Equation 17

$$I_A + I_B + I_C = 0,$$

Equation 18

$$|I_A| + |I_B| + |I_C| = NII,$$

Equation 19

where NII is the nutrient imbalance index computed irrespective of the sign of DRIS indices. Because DRIS indices are symmetrical, their sum is constrained to be zero (Beaufils, 1973).

Several modifications of the DRIS method have been made, including using nutrient products to account for nutrients accumulating in opposite directions with time and delineating nutrient shortage from excess by using a dry matter index as a separator (Walworth and Sumner, 1987). However, dry matter, a scale of measurement, was viewed wrongly as a component. Those modifications altered the symmetry of DRIS because indices computed from products and ratios did not sum to zero. While not reported for most crops, the accuracy of DRIS did not exceed 73% for cowpea (*Vigna* sp.) in Brazil (Wadt et al., 2016).

As DRIS is an empirical method, its scope of application is limited. Several biases in DRIS have been rectified by using log-ratio transformation techniques. When log ratios were used instead of dual ratios and variance terms were computed across multiratios, DRIS indices migrated toward *clr* results (Parent and Dafir, 1992). In contrast with DRIS, the *clr* transformation is a generic procedure amenable to multivariate analyses (Aitchison, 1986). The log ratio is useful because (1) geometric means rather than arithmetic means are the most appropriate for conducting statistical analyses on ratios (Fleming and Wallace, 1986), (2) it is common in engineering to use a logarithmic scale when ratios are greater than 10^4 (Budhu, 2010), and (3) the $\log(A/B)$ and $\log(B/A)$ ratios are reflective because $\log(A/B) = -\log(B/A)$ (Beverly, 1987a, 1987b). However, because the leaf tissue is a system bounded to 100%, a filling value must be computed to close the system to the measurement unit or scale. This value is determined by the difference between the unit of measurement scaled on dry matter and the sum of the analytical results.

5.5 Compositional data analysis of tissue compositions

The tissue compositional simplex is made of D components, as follows:

$$S^D = \left\{ x = [x_1, x_2, \dots, x_D] | x_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^D x_i = \kappa \right\}$$

Equation 20

where x_i is the i^{th} nonoverlapping component embedded within the unit of measurement κ . The filling value x_D (or F_v) is computed as the difference between κ and the sum of quantified components. Parent and Dafir (1992) rectified DRIS using *clr*, while Parent (2011) later suggested using *ilr*. Accuracies of *clr* and *ilr* exceeded 80% and even 90% across several horticultural and field crops (Badra et al., 2006; Hernandes et al., 2012; Modesto et al., 2014; Montes et al., 2016; Parent, 2011; Parent et al., 2012b, 2013a, 2013b, 2013c, 2015; Rozane et al., 2013; Souza et al., 2016). Several nutrient standards were elaborated (Hernandes et al., 2012; Montes et al., 2016; Parent et al., 1994, 2009, 2012b, 2013b, 2013c, 2015; Rozane et al., 2015; Souza et al., 2016). Fan et al. (2009) successfully related cadmium accumulation in potato tubers to balances between soil components. The design in Figure 6 illustrates one balance arrangement among the $D!(D-1)!/2^{D-1}$ possible arrangements of components in a D -part composition (Pawlowsky-Glahn et al., 2011).

Tissue nutrients are expressed as the total concentration. However, part of the total concentration is available for internal processes (Savageau, 1995) because some nutrients are more mobile or more subjected to accumulation or antagonism than others (Hill, 1980); nutrient shortage or excess affects crop yields (Ulrich and Hills, 1967). Powering is useful to normalize data (Box and Cox, 1964). Assignment of power parameters to concentrations can be viewed as a Box–Cox transformation (Box and Cox, 1964) to normalize the data. Power parameters alter the compositional simplex as follows:

$$S^D = \left\{ x = [x_1^{j_1} \cdot x_2^{j_2} \cdot \dots \cdot x_D] | x_i > 0, i = 1, 2, \dots, D; x_D + \sum_{i=1}^D x_i^j = \kappa \right\}$$

Equation 21

where x_i is the i^{th} nonoverlapping component within the unit of measurement κ , and x_D is computed by difference. The power terms j_i are optimized across concentration values to obtain normally distributed log ratio values (e.g., minimizing skewness and kurtosis).

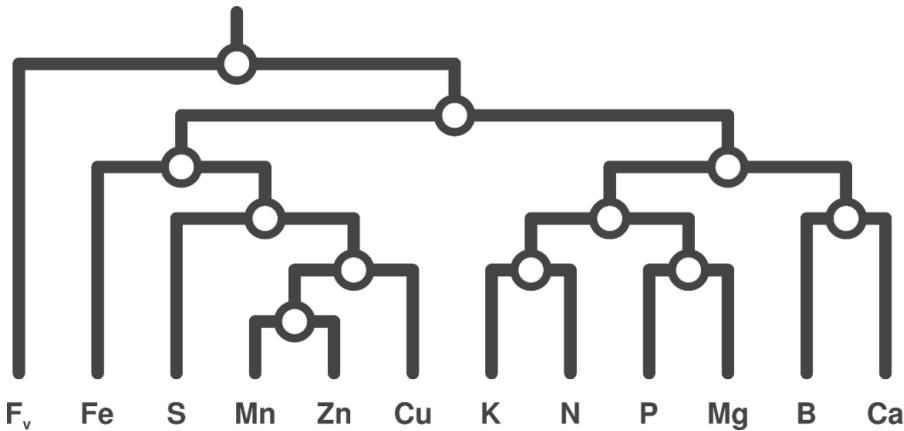


Figure 6. Nutrient balance design showing relationships between mobile and immobile macronutrients (in addition to B, which interacts with macronutrients), and between S and cationic micronutrients affected by fungicide sprays and soil properties.

5.6 Ionomics

The concept of “ionome”, the nutrient compositional vector of a cell tissue or organism, was introduced by Lahner et al. (2003) to find lines or environments that produce foods with altered nutritional profiles, or define gene by environmental effects on elemental accumulation (Baxter, 2015). The ionome has been described as the “social network of mineral nutrients” as altered by environmental factors, plant root anatomy, cell structure, production of chelating agents and sequestration within the cell (Baxter (2009)). The ionome has been treated as a collection of independent elements rather than combined traits as evidenced by interactions with each other (Wilkinson, 2000) and with biological molecules (Baxter, 2015).

To account for interactions between nutrients in living tissues and the D-1 degrees of freedom available to model compositional data, Parent et al. (2013b) used isometric log ratios to map the ionomes of the diagnostic leaves of wild and domesticated plant species from all over the world. Parent et al. (2013c) provided evidence of genotypic dominance over nutrient supply among four varieties of mango (*Mangifera indica*) grown in the state of São Paulo, Brazil. The classification provided basic information to adjust fertilization programs to the mango variety.

Attempts are being made to classify potato cultivars according to their ionome (Figure 7) to guide the fertilization of newly introduced cultivars in Quebec, Canada. Several old cultivars have been documented extensively in terms of fertilizer needs. New cultivars

are poorly documented but may show similarities with the ionomes of older ones. Recent development in potato (*Solanum tuberosum*) crop modelling to estimate the nitrogen requirements of potato cultivars only relies on the potato maturity grouping provided by the Canadian Food Inspection Agency (Parent et al., 2017). The maturity group is a potato growth trait for the time elapsed between seeding and harvest, estimated as follows: very early (65-70 days), early (70-90 days), mid-season (90-110 days), late (110-130 days) and very late (130 + days). Such classification is climate-specific and may be interpreted differently by growers. Ionome mapping may provide an additional and more objective information. The ionome data collection of the potato diagnostic leaf is at its inception.

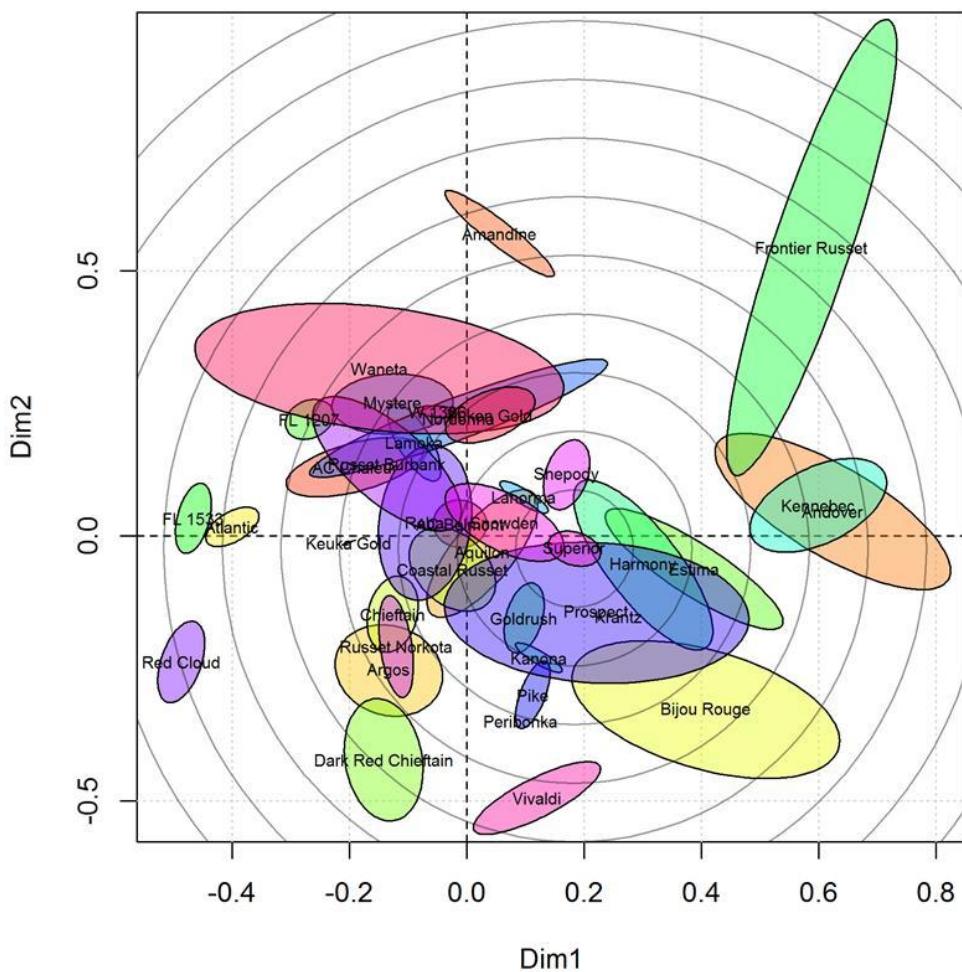


Figure 7. Ionome map of the diagnostic leaf of potato (*Solanum tuberosum L.*) cultivars.

5.7 Elaboration of nutrient standards

Several classification methods have been elaborated to separate the nutrient status of low- and high-performing crops. Webb (1972) proposed a boundary line, while Beaufils (1973) suggested using variance ratios of nutrient concentrations or ratios, assuming that

variance is higher at a lower than higher yields. Khiari et al. (2001) proposed a cumulative variance ratio function. However, the above methods may include cases of luxury consumption and contamination, which spoil nutrient standards. This problem can be avoided by using data partitioning procedures (Nelson and Anderson, 1984; Swets, 1988). The procedures can maximize the number of points in opposing quadrants by using the Mahalanobis distance computed across balances, as follows:

$$\mathcal{M} = \sqrt{(ilr_x - \bar{ilr}_y)^T COV^{-1}(ilr_y)},$$

Equation 22

where ilr_x and ilr_y are vectors of balances of the given composition x and the reference (e.g., high-yield) composition y , respectively; T is the matrix transposition operation; and COV is the covariance matrix. The Mahalanobis distance is iterated until the data partition is stabilized among quadrants or accuracy is maximized (Figure 8).

Quadrants are defined as follows: (1) true negative (TN) specimens (high yield, below-critical value); (2) true positive (TP) specimens (low yield, above-critical value); (3) false positive (FP) specimens (high yield, above-critical value; type I error); and (4) false negative (FN) specimens (low yield, below-critical value; type II error). The final partition is interpreted as follows (Parent et al., 2013c):

- Negative predictive value (NPV), computed as $TN/(TN+FN)$: probability that a balance diagnosis returns a high performance;
- Positive predictive value (PPV), computed as $TP/(TP+FP)$: probability that an imbalance diagnosis returns a low performance;
- Accuracy, computed as $(TN+TP)/(TN+FN+TP+FP)$: probability that an observation is correctly identified as balanced or imbalanced;
- Specificity, computed as $TN/(TN+FP)$: probability that a high-performance observation is balanced;
- Sensitivity, computed as $TP/(TP+FN)$: probability that a low-performance observation is imbalanced.

The joint normal distribution of TN balances is defined by a hyper-sphere rather than a hyper-cube delineated by confidence intervals (Nowaki et al., 2017). For this reason, ranges about ilr must be interpreted with care. Depending on the covariance matrix, the size of the critical hyper-ellipsoid may shrink rapidly towards ilr centroids as more balances are included. Only 0.092% of the volume of an ideal 11-dimensional hyper-cube is occupied by an embedded ideal 11-dimensional hyper-sphere. The situation is even more problematic when 12 nutrient concentration ranges are used to run separate diagnoses simultaneously, as is done usually, whereby some individual nutrient levels can fall outside critical concentration ranges while the crop is healthy.

Nevertheless, to facilitate comparison with published critical ranges, ilr values randomly combined within TN confidence intervals were back-transformed to concentration values by using the inverse ilr procedure in R (Souza et al., 2016). Minimal and maximal

concentration values are proxies for optimal nutrient ranges at high-yield level but bear no diagnostic value, as is the case for the separate *ilr* intervals. In multivariate diagnosis, the diagnostic standards are the TN *ilr* means and covariance. When the critical Mahalanobis distance exceeds the critical value, the specimen is declared misbalanced and further examined by using a mobile setup. The system is rebalanced by filling (through fertilization and liming) or emptying (by nutrient removal through harvest) the most misbalanced buckets, given prior knowledge about nutrient interactions and crop response to fertilization, liming, irrigation, and other agricultural practices.

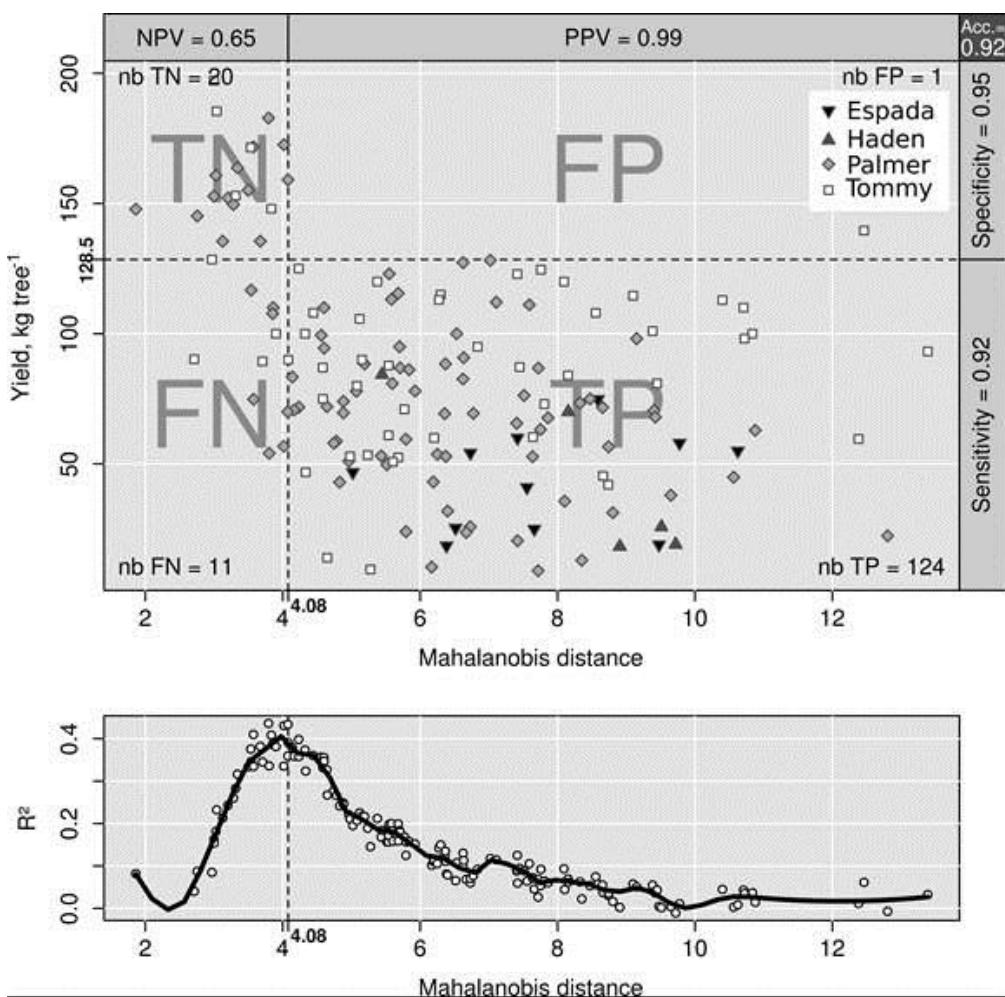


Figure 8. Partition of mango (*Mangifera indica*) data collected in 156 Brazilian orchards returning an accuracy of 92% and critical Mahalanobis distance of 4.08. TN = true negative; FN = false negative; TP = true positive; FP = false positive. NPV = negative predictive value; PPV = positive predictive value (Parent et al., 2013c).

6 Animal nutrition

Animal feed is a mixture of biochemical and mineral components. Results of feed biochemical and chemical analyses are commonly integrated into formulas to determine the suitability of the feed to sustain animal productivity and the need for supplements. Ionic imbalances in the blood of lactating ruminants may cause physiological troubles, such as hypocalcemia (Ca deficiency) and hypomagnesemia (Mg deficiency) (Grunes et al., 1970). Hypocalcemia results from the reduced ability to mobilize Ca from bones to blood, potentially leading to clinical milk fever and, eventually, to animal death (Charbonneau et al., 2006; National Research Council, 2001). Hypomagnesemia or grass tetany results from the inability to resorb Mg from bones, leading to severe convulsions and animal death (Jefferson et al. 2001). The risk of hypocalcemia is assessed from the dietary cation-anion difference (DCAD) by using cationic $[Na^+, K^+, Ca^{2+}, Mg^{2+}]$ and anionic $[Cl^-, S^{2-}, P^{3-}]$ species quantified in the feed (Charbonneau et al. 2006). The risk of hypomagnesemia is assessed from the $K/(Ca+Mg)$ molar ratio (Voisin, 1961), which should not exceed 2.2 in the feed (Jefferson et al., 2001).

Use of DCAD and the $K/(Ca+Mg)$ molar ratio has some numerical disadvantages. (1) Environments leading to hypomagnesemia and hypocalcemia do not account for complex ionic interactions. (2) Concentrations of ionic species are no more tractable after amalgamation of nutrient concentrations into ratios or regression models, hence confounding sources of imbalance. (3) The measurement scale differs between DCAD ($mmol_c \text{ kg}^{-1}$ of diet dry mass) and the $K/(Ca+Mg)$ molar ratio. Hence, it is difficult to address hypomagnesemia and hypocalcemia simultaneously. These difficulties can be overcome by the balance concept, wherein the compositional vector of hypomagnesemia is embedded in that of hypocalcemia (Figure 9).

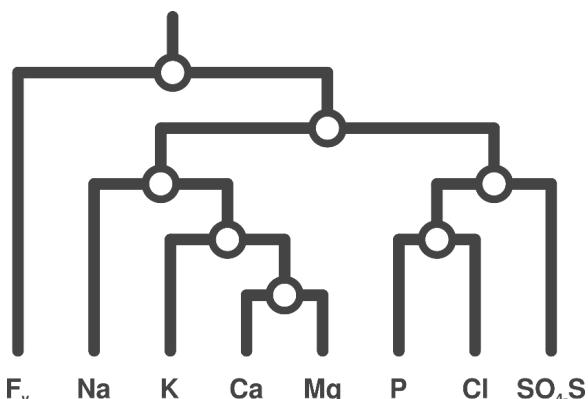


Figure 9. Mobile design embedding ionic balances for hypomagnesemia (K-Ca-Mg) into ionic balances for hypocalcaemia.

The mobile design may provide a coherent understanding of ionic relationships in the feed that influence animal health. All of the balances in Figure 8 influence hypocalcemia, whereas only the $[Ca, Mg | K]$ and $[Mg | Ca]$ balances influence hypomagnesemia. To validate this approach, the increasing pH of urine (Charbonneau et al. 2006), an index for

hypocalcemia, could be related to compositional balances for comparison with DCAD. The [Ca, Mg | K] and [Mg | Ca] balances could be calibrated against the K/(Ca+Mg) molar ratio to estimate the critical balance value for hypomagnesaemia. Feed components could be further assigned power parameters to account for differential digestibility (e.g., phytate-P vs. orthophosphate) or to rectify non-normal data distributions of the log ratio variables.

7 Conclusion

In this review paper, we revisited precompositional methods developed in agricultural research during the 20th century to diagnose soil quality, plant nutrition, and animal diets. Balance designs were presented to replace empirical indices, such as MWD, SLAN, BSCR, NIBC, DRIS and, possibly, DCAD. Balances arranged *ad hoc* between components of soil, solutions, plant, and feed mixtures are continuous orthonormal variables or coordinates amenable to multivariate and regression analysis. Although in their infancy, balance arrangements in mobile designs can provide a coherent understanding of the relationships between components of complex soil-plant-animal systems and subsystems and insightful descriptions of internal processes. Power parameters may be instrumental in weighting the course of time, the accessibility, mobility, accumulation, or nuisance of plant and feed nutrients, or to normalize the data.

Agronomists and bioengineers are urged to think in terms of interactive balance systems rather than isolated properties of compositions. The balance approach could support management decisions in precision agriculture models using compositional soil, plant, and feed data. Research in soil science, plant ionomics, epigenetics, and animal science can benefit from tools of compositional data analysis, as is the case for other disciplines that rely on compositional data.

8 Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC-DG 2254) for financial support.

9 References

- Abdi, D., Cade-Menun, B. J., Ziadi, N., and Parent, L.-É. (2015). Compositional statistical analysis of soil ³¹P-NMR forms. *Geoderma*. doi:10.1016/j.geoderma.2015.03.019.
- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.
- Aitchison, J., and Greenacre, M. (2002). Biplots of compositional data. *J. R. Stat. Soc. Ser. C Appl. Stat.* 51, 375–392. doi.org/10.1111/1467-9876.00275.
- Anderson, A.N., McBratney, A.B., and Crawford, J. W. (1998). Applications of fractals to soil studies. *Adv. Agron.* 63, 1–76.
- Badra, A., Parent, L.-É., Allard, G., Tremblay, N., Desjardins, Y., and Morin, N. (2006).

- Effect of leaf nitrogen concentration versus CND nutritional balance on shoot density and foliage colour of an established Kentucky bluegrass (*Poa pratensis* L.) turf. *Can. J. Plant Sci.* 86, 1107–1118. doi:10.4141/p05-242.
- Baeyens, J. (1969). Nutrition des plantes de culture (The Nutrition of Crop Plants). *Soil Sci.* 107, 233. doi:10.1097/00010694-196903000-00020.
- Barber, S. A. (1995). *Soil nutrient bioavailability : a mechanistic approach*. Wiley, NY.
- Bates, T. E. (1971). Factors affecting critical nutrient concentrations in plants and their environment: a review. *Soil Sci.* 112, 116–130. doi:10.1097/00010694-197108000-00005.
- Baxter, I. (2015). Should we treat the ionome as a combination of individual elements, or should we be deriving novel combined traits? *J. Exp. Bot.* 66, 2127–2131. doi:10.1093/jxb/erv040.
- Baxter, I. (2009). Ionomics: studying the social network of mineral nutrients. *Current Opinion in Plant Biology* 12, 381–386.
- Beaufils, E. R. (1973). Diagnosis and recommendation integrated system (DRIS). *Soil Sci. Bull.*, 1–132. Available at: <http://www.worldcat.org/title/diagnosis-and-recommendation-integrated-system-drис/oclc/637964264?ht=edition&referer=di> [Accessed May 2, 2013].
- Bergmann, W. (1988). *Ernährungsstörungen bei Kulturpflanzen*. Auflage 2. Stuttgart, New York: Gustav Fischer Verlag.
- Beverly, R. B. (1987a). Comparison of DRIS and alternative nutrient diagnostic methods for soybean. *J. Plant Nutr.* 10, 901–920.
- Beverly, R. B. (1987b). Modified DRIS method for simplified nutrient diagnosis of “Valencia” oranges. *J. Plant Nutr.* 10, 1401–1408.
- Bould, C., Bradfield, E. G., and Clarke, G. M. (1960). Leaf analysis as a guide to the nutrition of fruit crops. I. general principles, sampling techniques and analytical methods. *J. Sci. Food Agric.* 11, 229–242. doi:10.1002/jsfa.2740110501.
- Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Stat. Soc. Ser. B*, 26(2), 211–252.
- Bray, R. H. (1958). The Correlation of a Phosphorus Soil Test with the Response of Wheat Through a Modified Mitscherlich Equation. *Soil Sci. Soc. Am. J.* 22, 314. doi:10.2136/sssaj1958.03615995002200040013x.
- Breeuwsma, A., and Silva, S. (1992). *Phosphorus Fertilization and Environment Effects in the Netherlands and the Po Region (Italy)*. Agric. Res. Dep. Rep. 57, Winand Staring Centre for Integrated Land, Soil, and Water Res., Wageningen, The Netherlands.
- Brevik, E.C., and Burgess, L.C. (2013). *Soils and human health*. CRC Press, Boca Raton, FL.
- Buccianti, A., Nisi, B., and Vaselli, O. (2005). Thermodynamics and log-contrast

- analysis in fluid geochemistry. *CoDaWork Workshop* (Girona, Spain).
- Buccianti, A., and Pawlowsky-Glahn, V. (2005). New Perspectives on Water Chemistry and Compositional Data Analysis. *Math. Geol.* 37, 703–727. doi:10.1007/s11004-005-7376-6.
- Budhu, M. (2010). *Soil Mechanics and Foundations*. doi:10.1017/CBO9781107415324.004.
- Carter, M. R. (2002). Soil Quality for Sustainable Land Management: organic matter and aggregation interactions that maintain soil functions. *Agron. J.* 94, 38–47. doi:10.2134/agronj2002.3800.
- Carter, M. R., and Webster, G. R. (1990). Use of the Calcium to Total Cation Ratio in Soil Saturated Extracts as an Index of Plant Available Calcium. *Soil Sci.* 149, 212–217. doi:10.1097/00010694-199004000-00004.
- Charbonneau, E., Pellerin, D., and Oetzel, G. R. (2006). Impact of Lowering Dietary Cation-Anion Difference in Nonlactating Dairy Cows: A Meta-Analysis. *J. Dairy Sci.* 89, 537–548. doi:10.3168/jds.s0022-0302(06)72116-6.
- Chardon W.J., and Blaauw, D. (1998). Kinetic Freundlich equation applied to soils with a high residual phosphorus content. *Soil Sci.* 163, 30-35.
- Chayes, F. (1960). On correlation between variables of constant sum. *J. Geophys. Res.* 65, 4185–4193. doi:10.1029/JZ065i012p04185.
- Comas-Cufí, M., and Thió-Henestrosa, S. (2011). CoDaPack 2.0: a stand-alone, multi-platform compositional software. in *CoDaWork'11: 4th International Workshop on Compositional Data Analysis*, eds. J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego (Sant Feliu de Guxols).
- Dahnke, W. C., and Olson, R. A. (1990). “Soil test correlation, calibration, and recommendation,” in *Soil testing and plant analysis, Third Edition*, ed. R. L. Westerman (Madison WI: Soil Science Society of America), 45–71.
- De Wit, C. T. (1992). Resource use efficiency in agriculture. *Agric. Syst.* 40, 125–151. doi:10.1016/0308-521x(92)90018-j.
- Delgado-Baquerizo, M., Maestre, F. T., Gallardo, A., Bowker, M. a, Wallenstein, M. D., Quero, J. L., et al. (2013). Decoupling of soil nutrient cycles as a function of aridity in global drylands. *Nature* 502, 672–6. doi:10.1038/nature12670.
- Demšar, J., Curk, T., Erjavec, A., Hočevat, T., Milutinović, M., Možina, M., et al. (2013). Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* 14. Available at: <http://eprints.fri.uni-lj.si/2267/1/2013-Demsar-Orange-JMLR.pdf>.
- Diaz-Zorita, M., Perfect, E., and Grove, J. H. (2002). Disruptive methods for assessing soil structure. *Soil Till. Res.* 64, 3–22. doi:10.1016/S0167-1987(01)00254-9.
- Doran, J. W., Jones, A. J., Doran, J. W., and Parkin, T. B. (1996). “Quantitative Indicators of Soil Quality: A Minimum Data Set,” in *Methods for Assessing Soil Quality* (Soil Science Society of America). doi:10.2136/sssaspecpub49.c2.

- Duguet, F., Parent, L. E., and Ndayegamiye, A. (2006). Compositional indices of net nitrification in cultivated organic soils. *Soil Sci.* 171, 886–901. doi:10.1097/01.ss.0000235233.47804.e6.
- Eash, N. S., Karlen, D., and Parkin, T. (1994). “Fungal contributions to soil aggregation and soil quality,” in *Defining soil quality for a sustainable environment*, eds. J. W. Doran, D. C. Coleman, D. F. Bezdicek, and B. A. Stewart (Madison WI: SSSA Spec. Publ. 35,), 221–228. Available at: <http://naldc.nal.usda.gov/download/49220/PDF>.
- Egozcue, J. J., and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Math. Geol.* 37, 795–828. Available at: <http://springerlink.metapress.com/openurl.asp?genre=article&id=doi:10.1007/s11004-005-7381-9>.
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi:10.1023/A:1023818214614.
- Engle, M. A., and Blondes, M. S. (2014). Linking compositional data analysis with thermodynamic geochemical modeling: Oilfield brines from the Permian Basin, {USA}. *J. Geochemical Explor.* 141, 61–70. doi:10.1016/j.gexplo.2014.02.025.
- Fan, J.-L., Ziadi, N., Bélanger, G., Parent, L. É., Cambouris, A., and Hu, Z.-Y. (2009). Cadmium accumulation in potato tubers produced in Quebec. *Can. J. Soil Sci.* 89, 435–443. doi:10.4141/cjss08069.
- FAO, 2015. *Healthy soils are the basis for healthy food production*. Available at: <http://www.fao.org/soils-2015/news/news-detail/en/c/277682/>
- Filzmoser, P., Hron, K., and Reimann, C. (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci. Total Environ.* 407, 6100–6108. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19740525>.
- Fleming, P. J., and Wallace, J. J. (1986). How not to lie with statistics: the correct way to summarize benchmark results. *Commun. ACM* 29, 218–221. doi:10.1145/5666.5673.
- Gardner, W. . (1956). Representation of soil aggregate-size distribution by a logarithmic-normal distribution. *Soil Sci. Soc. Am. Proc.* 20, 151–153.
- Geraldson, C. M. (1984). “Nutrient intensity and balance,” in *Soil testing: correlating and interpreting the analytical results*, ed. M. Stelly (Madison, Wisconsin: American Society of Agronomy Publication), 75–84.
- Grunes, D. L., Stout, P. R., and Brownell, J. R. (1970). “Grass Tetany of Ruminants,” in *Adv. Agron.* (Elsevier {BV}), 331–374. doi:10.1016/s0065-2113(08)60272-2.
- Grunsky, E. C., Drew, L. J., Woodruff, L. G., Friske, P. W. B., and Sutphin, D. M. (2013). Statistical variability of the geochemistry and mineralogy of soils in the Maritime Provinces of Canada and part of the Northeast United States. *Geochemistry Explor. Environ. Anal.* 13, 249–266. doi:10.1144/geochem2012-138.
- Guérin, J., Parent, L.-É., and Abdelhafid, R. (2007). Agri-environmental Thresholds

- using Mehlich {III} Soil Phosphorus Saturation Index for Vegetables in Histosols. *J. Environ. Qual.* 36, 975. doi:10.2134/jeq2006.0424.
- Güsewell, S. (2004). N: P ratios in terrestrial plants: variation and functional significance. *New Phytol.* 164, 243–266. doi:10.1111/j.1469-8137.2004.01192.x.
- Hernandes, A., Parent, S.-É., Natale, W., and Parent, L. É. (2012). Balancing guava nutrition with liming and fertilization. *Rev. Bras. Frutic.* 34, 1224–1234. doi:10.1590/S0100-29452012000400032.
- Hill, J. (1980). The remobilization of nutrients from leaves. *J. Plant Nutr.* 2(4), 407–444.
- Holland, D. A. (1966). The interpretation of leaf analysis. *J. Hortic. Sci.* 41, 311–329.
- Howarth, R. J. (1996). Sources for a history of the ternary diagram. *Br. J. Hist. Sci.* 29, 337–356. doi:10.1017/S000708740003449X.
- Ingestad, T. (1987). New concepts on soil fertility and plant nutrition as illustrated by research on forest trees and stands. *Geoderma* 40, 237–252. doi:10.1016/0016-7061(87)90035-8.
- Jefferson, P. G., Mayland, H. F., Asay, K. H., and Berdahl, J. D. (2001). Variation in Mineral Concentration and Grass Tetany Potential among Russian Wildrye Accessions. *Crop Sci.* 41, 543. doi:10.2135/cropsci2001.412543x.
- Jelinek, Z. K. (1970). *Particle size analysis*. New York, NY: John Wiley & Sons.
- Justes, E. (1994). Determination of a Critical Nitrogen Dilution Curve for Winter Wheat Crops. *Ann. Bot.* 74, 397–407. doi:10.1006/anbo.1994.1133.
- Kemper, W. D., and Rosenau, R. C. (1986). “Aggregate Stability and Size Distribution,” in *Methods of soil analysis* (Madison, WI: American Society of Agronomy), 425–442.
- Kenworthy, A. L. (1967). “Plant analysis and interpretation of analysis for horticultural crops,” in *Soil testing and plant analysis, Part II*, eds. M. Stelly and H. Hamilton (Madison, Wisconsin: Soil Science Society of America), 59–75.
- Khiari, L., Parent, L.-É., and Tremblay, N. (2001). Selecting the High-Yield Subpopulation for Diagnosing Nutrient Imbalance in Crops. *Agron. J.* 93, 802. doi:10.2134/agronj2001.934802x.
- Khiari, L., Parent, L. E., Pellerin, A., Alimi, A. R. A., Tremblay, C., Simard, R. R., et al. (2000). An Agri-Environmental Phosphorus Saturation Index for Acid Coarse-Textured Soils. *J. Environ. Qual.* 29, 2052. doi:10.2134/jeq2000.00472425002900060053x.
- Kopelman, R. (1988). Fractal reaction kinetics. *Science* 241, 1620–1625.
- Kopelman, R. (1986). Rate process on fractals: theory, simulations, experiments. *J. Statist. Phys.* 42, 185–200.
- Kopitzke, P.M., Menzies, N.W. 2007. A review of the use of the basic cation saturation ratio and the “ideal” soil. *Soil Sci. Soc. Am. J.* 71(2), 259–265.

- Lagatu, H., and Maume, L. (1934). Le diagnostic foliaire de la pomme de terre. *Ann. l'École Natl. Agron. Montpellier* 22, 50–158.
- Lahner B., Gong J., Mahmoudian M., Smith E.L., Abid K.B., Rogers E.E., Guerinot M.L., Harper J.F., Ward J.M., McIntyre L. et al. (2003). Genomic scale profiling of nutrient and trace elements in *Arabidopsis thaliana*. *Nat. Biotechnol.* 21, 1215–1221.
- Leblanc, M. A., Gagné, G., and Parent, L. E. (2016). “Numerical clustering of soil series using morphological profile attributes for potato,” in *Digital Soil Morphometrics*, eds. A. E. Hartemink and B. Minasny (New York, NY: Springer).
- Leblanc, M. A., Parent, L. E., and Gagné, G. (2013). Phosphate and nitrate release from mucky mineral soils. *Open J. Soil Sci.* 3, 107–114. doi:10.4236/ojss.2013.32012.
- Liebhardt, W.C. (1981). The Basic Cation Saturation Ratio Concept and Lime and Potassium Recommendations on Delaware's Coastal Plain Soils. *Soil Sci. Soc. Am. J.* 45, 544–549.
- Linquist, B.A., Singleton, P.W., Yost, R.S., and Cassman, K.G. (1997). Aggregate size effect on the sorption and release of phosphorus in an Ultisol. *Soil Sci. Soc. Am. J.* 61, 160–166.
- Logsdon, S. D. (1995). Analysis of aggregate fractal dimensions and aggregate densities back-calculated from hydraulic conductivity. *Soil Sci. Soc. Am. J.* 59, 1216. doi:10.2136/sssaj1995.03615995005900050002x.
- Lopez, J., Parent, L. E., Tremblay, N., and Gosselin, A. (2002). Sulfate Accumulation and Calcium Balance in Hydroponic Tomato Culture. *J. Plant Nutr.* 25, 1585–1597. doi:10.1081/pln-120005409.
- Lovell, D., Müller, W., Taylor, J., Zwart, A., and Helliwell, C. (2011). “Proportions, Percentages, PPM: Do the Molecular Bioscience Treat Compositional Data Right?,” in *Compositional Data Analysis: Theory and Applications*, eds. V. Pawlowsky-Glahn and A. Buccianti (New York: John Wiley & Sons), 193–207.
- Lundy, M. E., Pittelkow, C. M., Linquist, B. A., Liang, X., van Groenigen, K. J., Lee, J., et al. (2015). Nitrogen fertilization reduces yield declines following no-till adoption. *F. Crop. Res.* 183, 204–210. doi:10.1016/j.fcr.2015.07.023.
- Macy, P. (1936). The Quantitative Mineral Nutrient Requirements of Plants. *Plant Physiol.* 11, 749–764. doi:10.1104/pp.11.4.749.
- Marschner, H. (1986). *Mineral Nutrition of Higher Plants*. London, Orlando: Academic Press doi:10.1146/annurev.es.11.110180.001313.
- McBratney, A. B., De Gruijter, J. J., and Brus, D. J. (1992). Spacial prediction and mapping of continuous soil classes. *Geoderma* 54, 39–64. doi:10.1016/0016-7061(92)90097-Q.
- McLean, E. O. (1984). “Contrasting concepts in soil test interpretation: sufficiency levels of available nutrients versus basic cation saturation ratios,” in *Soil testing: Correlating and interpreting the analytical results*, ed. M. Stelly (Madison, Wisconsin: American Society of Agronomy), 39–54.

- Modesto, V. C., Parent, S.-É., Natale, W., and Parent, L. E. (2014). Foliar Nutrient Balance Standards for Maize (*Zea mays L.*) at High-Yield Level. *Am. J. Plant Sci.* 5, 497–507. doi:10.4236/ajps.2014.54064.
- Montes, R. M., Parent, L. É., Amorim, D. A. de, Rozane, D. E., Parent, S.-É., Natale, W., et al. (2016). Nitrogen and Potassium Fertilization in a Guava Orchard Evaluated for Five Cycles: Effects on the Plant and on Production. *Rev. Bras. Ciência do Solo* 40, doi:10.1590/18069657rbcs20140532.
- Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., et al. (2017). Balance Trees Reveal Microbial Niche Differentiation. *mSystems* 2, e00162--16. doi:10.1128/msystems.00162-16.
- Mueller, L., Schindler, U., Mirschel, W., Shepherd, T. G., Ball, B. C., Helming, K., et al. (2010). Assessing the productivity function of soils. A review. *Agron. Sustain. Dev.* 30, 601–614. doi:10.1051/agro/2009057.
- Munson, R. D., and Nelson, W. L. (1990). “Principles and practices in plant analysis,” in *Soil testing and plant analysis*, ed. R. L. Westerman (Madison, Wisconsin: Soil Science Society of America), 359–387.
- National Research Council (2001). *Nutrient requirements of dairy cattle* 7th Ed. Washington DC.
- Nelson, L. A., and Anderson, R. L. (1984). “Partitioning of soil test-crop response probability,” in *Soil testing: Correlating and interpreting the analytical results*, ed. M. Stelly (Madison, Wisconsin: American Society of Agronomy), 19–28.
- Néméry, J., and Garnier, J. (2016). Biogeochemistry: The fate of phosphorus. *Nat. Geosci.* 9, 1–2. doi:10.1038/ngeo2702.
- Nowaki, R.H.D., Parent, S.-É., Cecilio Filho, A.B., Rozane, D.E., Meneses, N.B., da Silva, J.A.D.S., Natale, W., and Parent, L.E. (2017). Phosphorus Over-Fertilization and Nutrient Misbalance of Irrigated Tomato Crops in Brazil. *Frontiers Plant Sci.* (in press), Manuscript ID: 243252,
- Ouimet, R., and Camiré, C. (1995). Foliar deficiencies of sugar maple stands associated with soil cation imbalances in the Quebec Appalachians. *Can. J. Soil Sci.* 75, 169–175. doi:10.4141/cjss95-024.
- Ouimet, R., Moore, J.-D., and Duchesne, L. (2013). Soil Thresholds Update for Diagnosing Foliar Calcium, Potassium, or Phosphorus Deficiency of Sugar Maple. *Commun. Soil Sci. Plant Anal.* 44, 2408–2427. doi:10.1080/00103624.2013.803563.
- Parent, L. E. (2011). Diagnosis of the nutrient compositional space of fruit crops. *Rev. Bras. Frutic.* 33, 321–334. doi:10.1590/S0100-29452011000100041.
- Parent, L.E., and Bélanger, A. 1985. Comparison between Freundlich and fixed regression models of linuron retention by organic soil materials. *J. Env. Sci. Health A20* 3, 293-304.
- Parent, L. E., Cambouris, A. N., and Muhamenimana, A. (1994). Multivariate Diagnosis of Nutrient Imbalance in Potato Crops. *Soil Sci. Soc. Am. J.* 58, 1432-1438.

- doi:10.2136/sssaj1994.03615995005800050022x.
- Parent, L. E., Cissé, E. S., Tremblay, N., and Bélair, G. (1997). Row-centred log ratios as nutrient indexes for saturated extracts of organic soils. *Can. J. Soil Sci.* 77, 571–578. doi:10.4141/s96-073.
- Parent, L. E., and Dafir, M. (1992). A Theoretical Concept of Compositional Nutrient Diagnosis. *J. Amer. Soc. Hort. Sci.* 117, 239–242.
- Parent, L. E., de Almeida, C. X., Hernandes, A., Egoscue, J. J., Gülser, C., Bolinder, M. A., et al. (2012a). Compositional analysis for an unbiased measure of soil aggregation. *Geoderma* 179–180, 123–131. doi:10.1016/j.geoderma.2012.02.022.
- Parent, L. E., Natale, W., and Ziadi, N. (2009). Compositional nutrient diagnosis of corn using the Mahalanobis distance as nutrient imbalance index. *Can. J. Soil Sci.* 89, 383–390. doi:10.4141/cjss08050.
- Parent, L. E., Parent, S.-É., Hébert-Gentile, V., Naess, K., and Lapointe, L. (2013a). Mineral balance plasticity of cloudberry (*Rubus chamaemorus*) in Quebec-Labrador. *Am. J. Plant Sci.* 4, 1509–1520.
- Parent, L. E., Parent, S.-É., Rozane, D.-E., Amorim, D. ., Hernandes, A., and Natale, W. (2012b). Unbiased approach to diagnose the nutrient status of red guava (*psidium guajava*). in *III International Symposium on Guava and other Myrtaceae* (Petrolina: ISHS Acta Horticulturae), 145–159. Available at: http://www.actahort.org/books/959/959_18.htm.
- Parent, L. E., Parent, S. É., and Ziadi, N. (2014). Biogeochemistry of soil inorganic and organic phosphorus: A compositional analysis with balances. *J. Geochemical Explor.* 141, 52–60. doi:10.1016/j.gexplo.2014.01.030.
- Parent, L. E., Viau, A., and Ancti (2000). Organic soil quality indicators for N and P. *Suo* 51, 71–81.
- Parent, S.-É., Barlow, P., and Parent, L. E. (2015). Nutrient Balances of New Zealand Kiwifruit (*Actinidia deliciosa* cv. Hayward) at High Yield Level. *Commun. Soil Sci. Plant Anal.* 46, 256–271. doi:10.1080/00103624.2014.989031.
- Parent, S.-É., and Parent, L. E. (2015). Biochemical Fractionation of Soil Organic Matter after Incorporation of Organic Residues. *Open J. Soil Sci.* 5, 135–143. doi:10.4236/ojss.2015.56013.
- Parent, S.-É., Parent, L. E., Egoscue, J. J., Rozane, D.-E., Hernandes, A., Lapointe, L., et al. (2013b). The plant ionome revisited by the nutrient balance concept. *Front. Plant Sci.* 4, 1–10. doi:10.3389/fpls.2013.00039.
- Parent, S.-É., Parent, L. E., Rozane, D.-E., Hernandes, A., and Natale, W. (2012c). “Nutrient Balance as Paradigm of Soil and Plant Chemometrics,” in *Soil fertility*, ed. R. N. Issaka (Intech), 83–114. doi:<http://dx.doi.org/10.5772/53343>.
- Parent, S.-É., Parent, L. E., Rozane, D. E., and Natale, W. (2013c). Plant ionome diagnosis using sound balances: case study with mango (*Mangifera Indica*). *Front. Plant Sci.* 4, 449. doi:10.3389/fpls.2013.00449.

- Parent, S.-É., Leblanc, M., Parent, A.C., Coulibali, Z., and Parent, L. E. 2017. Site-specific multilevel modeling of potato response to nitrogen fertilization. *Front. Environ. Sci. - Environmental Informatics* (in review)
- Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2011). Principal balances. in *4th International Workshop on Compositional Data Analysis (Codawork 2011)*, eds. J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego (San Feliu de Guixols, Spain). Available at: <http://congress.cimne.com/codawork11/Admin/Files/FilePaper/p55.pdf>.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 60, 489–498.
- Peck, T. R. (1990). Soil testing: Past, present and future. *Commun. Soil Sci. Plant Anal.* 21, 1165–1186. doi:10.1080/00103629009368297.
- Pellerin, A., Parent, L. E., Fortin, J., Tremblay, C., Khiari, L., and Giroux, M. (2006a). Environmental Mehlich-III soil phosphorus saturation indices for Quebec acid to near neutral mineral soils varying in texture and genesis. *Can. J. Soil Sci.* 86, 711–723. doi:10.4141/S05-070.
- Pellerin, A., Parent, L. E., Tremblay, C., Fortin, J., Tremblay, G., Landry, C. P., et al. (2006b). Agri-environmental models using Mehlich-III soil phosphorus saturation index for corn in Quebec. *Can. J. Soil Sci.* 86, 897–910. doi:10.4141/S05-071.
- Prevot, P., and Ollagnier, M. (1961). “Law of the minimum and balanced mineral nutrition,” in *Plant analysis and fertilizer problems*, ed. W. Reuther (American Institute of Biological Sciences), 257–277.
- Quaggio, J. A., van Raij, B., and Malavolta, E. (1985). Alternative use of the SMP-buffer solution to determine lime requirement of soils. *Commun. Soil Sci. Plant Anal.* 16, 245–260. doi:10.1080/00103628509367600.
- Quinche-Gonzalez, M., Pellerin, A., and Parent, L. E. (2016a). Meta-analysis of lettuce (*Lactuca sativa* L.) response to added N in organic soils1. *Can. J. Plant Sci.* 96, 670–676. doi:10.1139/cjps-2015-0301.
- Quinche-Gonzalez, M., Pellerin, A., and Parent, L. E. (2016b). Onion Response to Added N in Histosols of Contrasting C and N Contents. *Am. J. Plant Sci.*, 469–478.
- Redfield, A. (1934). “On the proportions of organic derivatives in sea water and their relation to the composition of plankton,” in *James Johnstone memorial volume*, ed. R. Daniel (University Press of Liverpool), 177–192.
- Rieu, M., and Sposito, G. (1991). Fractal fragmentation, soil porosity, and soil water properties: II. Applications. *Soil Sci. Soc. Am. J.* 55, 1239. doi:10.2136/sssaj1991.03615995005500050007x.
- Rodgers, J. L., Nicewander, W. A., and Toothaker, L. (1984). Linearly Independent, Orthogonal, and Uncorrelated Variables. *Am. Stat.* 38, 133. doi:10.2307/2683250.
- Rozane, D. E., Hortense Torres, M., Antunes de Souza, H., Natale, W., and Silva, S. H.

- M.-G. da (2013). Application of a byproduct of guava processing in an Ultisol, in the presence and absence of mineral fertilization. *Idesia* 31, 89–96. Available at: <https://dialnet.unirioja.es/servlet/articulo?codigo=4522656&info=resumen&idioma=SPA> [Accessed February 24, 2016].
- Rozane, D. E., Parent, L. E., and Natale, W. (2015). Evolution of the predictive criteria for the tropical fruit tree nutritional status. *Cientifica* 44, 102. doi:10.15361/1984-5529.2016v44n1p102-112.
- Savageau, M.A. (1995). Michaelis-Menten Mechanism Reconsidered: Implications of Fractal Kinetics. *J. theor. Biol.* 176, 115-124.
- Schrevens, E., and Cornell, J. (1993). Design and analysis of mixture systems: Applications in hydroponic, plant nutrition research. *Plant Soil* 154, 45–52. doi:10.1007/BF00011070.
- Sims, J. T., Simard, R. R., and Joern, B. C. (1998). Phosphorus Loss in Agricultural Drainage: Historical Perspective and Current Research. *J. Environ. Qual.* 27, 277–293. doi:10.2134/jeq1998.00472425002700020006x.
- Souza, H. A., Parent, S.-É., Rozane, D. E., Amorim, D. A., Modesto, V. C., Natale, W., and Parent, L.E. (2016). Guava Waste to Sustain Guava (*Psidium guajava*) Agroecosystem: Nutrient “Balance” Concepts. *Front. Plant Sci.* 7, 1–13. doi:10.3389/fpls.2016.01252.
- Stevenson, F. (1986). *Cycles of soil, Carbon, nitrogen, phosphorus, sulfur, micronutrients*. New-York: Wiley Interscience.
- Stewart, C. E., Plante, A. F., Paustian, K., Conant, R. T., and Six, J. (2008). Soil Carbon Saturation: Linking Concept and Measurable Carbon Pools. *Soil Sci. Soc. Am. J.* 72, 379–392.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* (80-). 240, 1285–1293. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3287615>.
- Tanner, J. M. (1949). Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J. Appl. Physiol.* 2, 1–15. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18133122> [Accessed July 18, 2013].
- Thuriès, L., Pansu, M., Larré-Larrouy, M.C., and Feller, C. (2002). Biochemical composition and mineralization kinetics of organic inputs in a sandy soil. *Soil Biol. Biochem.* 34, 239-250.
- Thuriès, L., Pansu, M., Feller, C., Herrmann, P., and Rémy, J. C. (2001). Kinetics of added organic matter decomposition in a Mediterranean sandy soil. *Soil Biol. Biochem.* 33, 997–1010.
- Timmer, V. R. (1997). Exponential nutrient loading: a new fertilization technique to improve seedling performance on competitive sites. *New For.* 13, 275–295.
- Ulrich, A. (1952). Physiological Bases for Assessing the Nutritional Requirements of Plants. *Annu. Rev. Plant Physiol.* 3, 207–228. doi:10.1146/annurev.pp.03.060152.001231.

- Ulrich, A., and Hills, F. J. (1967). "Principles and practices of plant analysis," in *Soil testing and plant analysis. Part II*, eds. M. Stelly and H. Hamilton (Madison, Wisconsin: Soil Science Society of America), 11–24.
- van Bavel, C. H. M. (1949). Mean weight-diameter of soil aggregates as a statistical index of aggregation. *Proceedings. Soil Sci. Soc. Am.* 14, 20–23.
- van den Boogaart, K. G., Tolosana-Delgado, R., and Bren, M. (2014). "*compositions*": *Compositional Data Analysis in R package*. Available at: <http://cran.r-project.org/package=compositions>.
- Voisin, A. (1961). *Grass productivity*. London: Crosby Lockwood & Son.
- Wadt, P. G. S., Traspadini, E. I. F., Martins, R. A., Melo, F. B., Oliveira, I. J., Rodrigues, J. E. L. F., et al. (2016). Mdeidas de acuracia na qualificaçao dos diagnosticos nutricionais: teoria e pratica. in *Nutriçao e adubaçao de hortaliças, 5th Brasil. Symp. Plant Nutrition at high productivity level*, eds. R. M. Prado and A. B. C. Filho (UNESP, Jaboticaba SP, Brazil), 371–391.
- Walworth, J. L., and Sumner, M. E. (1987). The Diagnosis and Recommendation Integrated System (DRIS). *Adv. Soil Sci.* 6, 149–188. doi:10.1007/978-1-4612-4682-4.
- Webb, R. A. (1972). Use of the Boundary Line in the analysis of biological data. *J. Hortic. Sci.* 47, 309–319. doi:10.1080/00221589.1972.11514472.
- Wilkinson, S. R. (2000). "Nutrient interactions in soil and plant nutrition," in *Handbook of soil science*, ed. M. E. Sumner (Boca Raton FL: CRC Press), D89–D112.
- Xu, Yan, Jimenez, M.A., Parent, S.-É., Leblanc, M., Ziadi, N., and Parent, L.E. (2017a). Compaction of coarse-textured agricultural soils: balance models of soil compositions. *Frontiers Plant Sci.* (in review).
- Xu, Yan, Parent, S.-É., Leblanc, M., Ziadi, N., and Parent, L.E. (2017b). *Synthetic compositional index to conduct meta-analysis of soil aggregation studies*. Poster presentation, 2017 Conference of the Canadian Society of Soil Science, Trent University, Peterborough, Ontario, June 10-14.
- Zhi J, Jing C, Lin S, Zhang C, Liu Q, et al. (2014) Estimating Soil Organic Carbon Stocks and Spatial Patterns with Statistical and GIS-Based Methods. *PLoS ONE* 9(5): e97757. doi:10.1371/journal.pone.0097757
- Zhong, Y., Yan, W., Chen, J., and Shangguan, Z. (2014). Net ammonium and nitrate fluxes in wheat roots under different environmental conditions as assessed by scanning ion-selective electrode technique. *Nature Scientif. Rep.* 4, 7223. doi:10.1038/srep07223

Survey data on perceptions of contraceptive measures as compositional tables

V. Pawlowsky-Glahn^{1,3}, J.J. Egozcue², and M. Planes-Pedra¹

¹University of Girona, Girona, Spain;

²Technical University of Catalonia, Barcelona, Spain

³*vera.pawlowsky@udg.edu*

Abstract

There is a general believe that the perceived benefits of the male condom increase the likelihood of its use, while perceived harm reduces it, with a greater influence of the former on the latter. This believe takes into account that, in addition to the male and female condoms, there are different methods to avoid sexually transmitted infections/AIDS and unwanted pregnancies, such as the contraceptive pill and the morning-after pill (emergency contraception). Here we study globally and by gender the relative evaluation of young people of the possible benefits associated with the three most used methods (condoms, contraceptive pill, morning-after pill), as a possible indicator of preferential use, i.e. if their evaluation supports the general believe.

The results show that condoms are positively valued as protection against sexually transmitted infections/AIDS; as a sign of interest in protecting the health of the couple; and for their lack of side effects. The reverse occurs in the case of pills, as hormonal methods are much better valued than the condom with respect to their ability to increase the feelings of pleasure in both genders.

Examining results by gender, it can be observed that the assessments of women are less extreme than those of men in five of the six items mentioned in the previous paragraph, but are more pronounced with respect to the assumption that the use of the condom demonstrates interest to protect the health of the couple.

When confronted with the three protective methods, the preferences of the participants in the study are inclined towards condom against the contraceptive pill and the postcoital pill, but in the case of men with more intensity.

Key words: compositional data analysis, survey data, compositional table

1 Introduction

There is a general believe that the perceived benefits of the male condom increase the likelihood of its use, while perceived harm reduces it, with a greater influence of the former on the latter. This believe takes into account that, in addition to the male and female condoms, there are different methods to avoid sexually transmitted infections/AIDS and unwanted pregnancies, such as the contraceptive pill and the morning-after pill (emergency contraception). Here we study globally and depending on the gender the relative evaluation of young people of the possible benefits associated with the three most used methods (condoms (C), contraceptive pill (P), morning-after pill (M)), as a possible indicator of preferential use, i.e. if their evaluation supports the general believe.

2 Framework

2.1 Assumptions and Principles

Our approach is based on the premises that compositional data (Aitchison, 1986; Barceló-Vidal and Martín-Fernández, 2016) are vectors with strictly positive components, $\mathbf{y} = [y_1, y_2, \dots, y_D] \in \mathbb{R}_+^D$, that carry relative information. Therefore, a rescaling has no effect on the information carried by the data. Proportional vectors of positive components are equivalent, that is, compositions are equivalence classes (Barceló-Vidal et al., 2001). The usual representative of the sample space is the simplex ($\kappa = \text{constant}$),

$$\mathbb{S}^D = \left\{ \mathbf{x} = [x_1, \dots, x_D] \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^D x_i = \kappa \right\} \subset \mathbb{R}_+^D \subset \mathbb{R}^D,$$

which is obtained using the closure operator,

$$\mathbf{x} = \mathcal{C}(\mathbf{y}) = \left[\frac{\kappa \cdot y_1}{\sum_{i=1}^D y_i}, \frac{\kappa \cdot y_2}{\sum_{i=1}^D y_i}, \dots, \frac{\kappa \cdot y_D}{\sum_{i=1}^D y_i} \right],$$

just dividing each component by the sum of all components and multiplying by κ . This definition includes parts of a whole, like e.g. parts per one, percentages, ppm, molar concentrations, or relative frequencies (Pawlowsky-Glahn et al., 2015). However, the representation in the simplex is not essential for assuming that an array of positive numbers is a composition. In its place it is enough to assume the principles underlaying compositional data analysis. They are:

1. Scale invariance: scaling factors do not alter the analysis; the relevant entities are the ratios of components.
2. Subcompositional coherence: analysis of a subcomposition is compatible with that of the whole composition, i.e. subcompositional scale invariance and dominance ($d_a(x_1, x_2) \geq d_a(s_1, s_2)$) hold, and ratios of common parts are preserved.

This allows considering a table of positive scores as a composition. In the present case, we deal with a survey in which each row scores the effectiveness of the methods C, P, M with respect to 10 different situations (rows). Then, the answer of an individual to the survey consists of a table (10 rows, 3 columns). We assume that the two mentioned principles hold and each table is a composition.

2.2 Aitchison geometry of the simplex

The approach presented here benefits from the particular algebraic-geometric structure of the sample space of CoDa. Each equivalence class is considered represented in the simplex. The simplex

is endowed with the Aitchison geometry (Pawlowsky-Glahn et al., 2015; Pawlowsky-Glahn and Egozcue, 2001), which corresponds to a Euclidean type vector space structure with its operations (perturbation and powering) and a metric defined on the simplex (inner product, norm and distance).

When compositions are arranged as an (I, J) table it is necessary to redefine these geometric elements adapted to tables. They are specified in Egozcue et al. (2015). The perturbation of tables X and Y is the multiplication componentwise up to the closure, i.e. the (ij) -th entry is $[X \oplus Y]_{ij} \propto x_{ij}y_{ij}$, the latter being the entries of X and Y respectively; for a real number α and up to closure, the powering of X has the (ij) -th entry $[\alpha \odot X]_{ij} \propto x_{ij}^\alpha$.

The centered log-ratio (clr) representation of an (I, J) table, X is also an (I, J) table with entries

$$\text{clr}_{ij}(X) = \ln \frac{x_{ij}}{\text{g}_m(X)} \quad , \quad i = 1, 2, \dots, I \quad , \quad j = 1, 2, \dots, J \quad , \quad \text{g}_m(X) = \left(\prod_{i=1}^I \prod_{j=1}^J x_{ij} \right)^{1/(IJ)} .$$

The sum of all elements of matrix $\text{clr}(X)$ is null. From $\text{clr}(X)$, the closed table X is readily recovered as $X = \mathcal{C} \exp(\text{clr}(X))$. The clr representation provides a simple expression of the metric elements for tables. The inner product can be computed as

$$\langle X, Y \rangle_a = \langle \text{clr}(X), \text{clr}(Y) \rangle \quad , \quad d_a(X, Y) = d_e(\text{clr}(X), \text{clr}(Y)) \quad ,$$

where the subscript a stands for Aitchison geometry and e for ordinary Euclidean geometry, interpreting the matrices as simple vectors. The geometric elements characterizing the sample space allow to define basic statistical elements. For a sample of compositional tables, the center or mean is defined as

$$\text{Cen}(\mathbf{X}) = \frac{1}{n} \oplus \bigoplus_{k=1}^n X_k = \mathcal{C} \exp \left(\frac{1}{n} \sum_{k=1}^n \text{clr}(X_k) \right) \quad , \quad (1)$$

where \mathbf{X} denotes an n -sample of compositional tables X_k . The sample total variance of \mathbf{X} is (Egozcue et al., 2015)

$$\text{totVar}(\mathbf{X}) = \sum_{i=1}^I \sum_{j=1}^J \text{Var}(\text{clr}_{ij}(\mathbf{X})) \quad .$$

2.3 Independent and interaction compositional tables

The main contribution in Egozcue et al. (2015) is that any (I, J) compositional table has a unique decomposition into an independent and an interaction table, and that this decomposition is orthogonal in the sense of the Aitchison geometry. It can be written as

$$X = X_{\text{ind}} \oplus X_{\text{int}} \quad , \quad \langle X_{\text{ind}}, X_{\text{int}} \rangle_a = 0$$

or alternatively in terms of the respective clr's

$$\text{clr}(X) = \text{clr}(X_{\text{ind}}) + \text{clr}(X_{\text{int}}) \quad .$$

The entries of the independent part X_{ind} are obtained as the product of the corresponding elements of the geometric marginals of X . In an analysis trying to relate methods (columns) and situations (rows), the independent part is non informative, as it only reflects marginal characteristics. Removing the independent part $X \oplus ((-1) \odot X_{\text{ind}}) = X_{\text{int}}$ gives the interaction table or its clr, $\text{clr}(X) - \text{clr}(X_{\text{ind}}) = \text{clr}(X_{\text{int}})$. It is remarkable that the removal of the interaction is equivalent to a double compositional centering of X or a standard double centering of $\text{clr}(X)$. The interaction table X_{int} satisfies that its geometric marginals are uniform and that all rows and columns of $\text{clr}(X_{\text{int}})$ add to zero.

In order to analyse a sample of compositional tables, now identified with the collection of responses to the mentioned survey, a first step can be the computation of the sample center given in 1. Then attention is paid to the analysis of the interaction table of the center which is equal to $\text{Cen}(X_{\text{int}})$. The interpretation of interactions between methods and situations is easier when based on $\text{clr}(\text{Cen}(X_{\text{int}}))$. The reason for that is the properties of such a matrix: (a) their entries add to zero; (b) the squares of the entries add up to the mean simplicial deviance (Egozcue et al., 2015), a measure of cross information between methods and situations; (c) positive (negative) entries point out cells in which score is larger (less) than predicted by the independent table; as the sum of those entries is zero by rows and columns any positive entry should be compensated by negative entries in the same row and column.

3 Survey data as samples of a compositional table

3.1 Preliminaries

A group of 145 undergraduate students (76% females, 24% males) was asked to evaluate three protective measures, giving a value between 1 and 99 to each of them in 10 different situations (Grimley et al., 1995; Prat et al., 2012, 2016), summarised in Table 1. The answers of each student is here considered as a sample of a compositional table. When closed to 1, each answer can be considered as the probability the student assigns to each protective method satisfying the situation described in the question. Thus, available data constitute a sample of size $n = 145$ of $(I, J) = (10, 3)$ compositional tables. The global compositional mean or center of this data set is proportional to

Table 1: Survey data as a compositional table. P = preservative; C = contraceptive pills; M = morning after pill.

item	preventive measure	P	C	M
1	protect from sexually transmitted infections (STI)	P1	C1	M1
2	protect from pregnancy	P2	C2	M2
3	provide peace during and after intercourse	P3	C3	M3
4	economically accessible	P4	C4	M4
5	protect from the transmission of the AIDS virus	P5	C5	M5
6	evidence interest in protecting the health of the couple	P6	C6	M6
7	increase feelings of pleasure in man	P7	C7	M7
8	increase feelings of pleasure in woman	P8	C8	M8
9	are easy to use correctly	P9	C9	M9
10	do not cause side effects	P10	C10	M10

a compositional contingency table as the one shown in Equation (2), where we have the center or mean table on the left, and its decomposition in independent (middle) and interaction (right) tables as in Equation (1). A compositional contingency table can be orthogonally decomposed into an independent table, which is the closest independent table in the Aitchison sense, and an interaction table, which reflects the deviations from independence (Egozcue et al., 2015). In terms of probabilities, the interaction table can be interpreted as reflecting a lack (excess) of probability

	P	C	M
1	1.3	-0.89	-0.41
2	-0.26	0.31	-0.05
3	-0.16	0.55	-0.39
4	-0.13	-0.06	0.2
5	1.54	-1.01	-0.53
6	0.9	-0.43	-0.48
7	-1.87	1.07	0.8
8	-1.65	1.02	0.63
9	-0.47	0.15	0.31
10	0.79	-0.7	-0.09

Figure 1: Overall interaction table (clr). Colors enhance importance of interaction (I). Red: strong positive I. Ocre: medium positive I. Grey: low positive I. White: no I. Pale blue: low negative I. Aquamarine: medium negative I. Dark blue: strong negative I.

when the value is smaller (larger) than the corresponding value in the independent table.

$$\begin{array}{ccc}
 P & C & M \\
 \left(\begin{array}{ccc} 0.1047 & 0.0054 & 0.0032 \\ 0.0513 & 0.0420 & 0.0109 \\ 0.0487 & 0.0454 & 0.0066 \\ 0.0561 & 0.0277 & 0.0134 \\ 0.1105 & 0.0040 & 0.0024 \\ 0.0882 & 0.0107 & 0.0038 \\ 0.0077 & 0.0674 & 0.0192 \\ 0.0100 & 0.0663 & 0.0168 \\ 0.0405 & 0.0347 & 0.0151 \\ 0.0744 & 0.0077 & 0.0053 \end{array} \right) & = & \left(\begin{array}{ccc} 0.0366 & 0.0169 & 0.0063 \\ 0.0860 & 0.0396 & 0.0147 \\ 0.0732 & 0.0338 & 0.0126 \\ 0.0825 & 0.0380 & 0.0141 \\ 0.0304 & 0.0140 & 0.0052 \\ 0.0460 & 0.0212 & 0.0079 \\ 0.0645 & 0.0297 & 0.0111 \\ 0.0671 & 0.0309 & 0.0115 \\ 0.0830 & 0.0383 & 0.0142 \\ 0.0433 & 0.0199 & 0.0074 \end{array} \right) \oplus \left(\begin{array}{ccc} 0.0910 & 0.0101 & 0.0164 \\ 0.0190 & 0.0338 & 0.0236 \\ 0.0212 & 0.0428 & 0.0167 \\ 0.0216 & 0.0232 & 0.0301 \\ 0.1156 & 0.0090 & 0.0146 \\ 0.0611 & 0.0161 & 0.0154 \\ 0.0038 & 0.0722 & 0.0553 \\ 0.0048 & 0.0683 & 0.0466 \\ 0.0155 & 0.0289 & 0.0338 \\ 0.0548 & 0.0123 & 0.0226 \end{array} \right) \quad (2)
 \end{array}$$

3.2 Interactions between items and protective methods.

The interaction table in Equation 2 is not straightforward to interpret in its raw form, while its clr (Figure 1) representation gives a much better insight into the relations between items and protective methods. In fact, a negative value reflects a lack of probability, while a positive value reflects the contrary. The results presented in Figure 1 show that condoms are, overall, very positively valued as protection against sexually transmitted infections/AIDS (items 1 and 5), as the entries in the clr-interaction table are the largest positive values. At the same time they are considered an obstacle to pleasure during intercourse, as the entries in the clr-interaction table are the largest negative values (items 7 and 8). The opposite happens with the two types of pill, which are poorly valued as protective against infections, but positively valued regarding pleasure (items 7, 8). Other less expected results also deserve a comment. For example, the good evaluation of the condom in relation to “showing interest in protecting the health of the couple” (item 6), highlights an altruistic motivation, far from selfish motivations oriented towards one’s own health, pleasure or economy. Despite the negative evaluations of the preservative with respect to male pleasure, in previous investigations it has been shown that it is not a predictive factor of their use (Prat et al., 2016). It is surprising that the morning-after pill is negatively associated with pregnancy protection (item 2), even though there is no interaction with the other two methods.

Students were also asked for additional information, like sex and age. Age did not show any pattern related to the evaluation of the different protective measures and is therefore not presented here, while the clr-interaction tables for males and females are presented in Figure 2 for comparison. As can be observed, the clr-interaction tables for both genders are very similar, although it is evident that women have less extreme opinions than men. Item 6, “evidence interest in protecting the

1	1.45	-1.03	-0.43	1	1.26	-0.86	-0.41
2	-0.2	0.22	-0.03	2	-0.28	0.33	-0.05
3	-0.29	0.56	-0.27	3	-0.1	0.55	-0.45
4	-0.09	0.16	-0.06	4	-0.15	-0.11	0.26
5	1.76	-1.14	-0.62	5	1.48	-0.97	-0.5
6	0.8	-0.26	-0.54	6	0.93	-0.46	-0.48
7	-2.24	1.25	1.04	7	-1.75	1	0.75
8	-1.9	1.12	0.78	8	-1.58	0.97	0.61
9	-0.52	0.17	0.35	9	-0.45	0.14	0.31
10	1.28	-1.06	-0.22	10	0.63	-0.59	-0.04
	P	C	M		P	C	M

Figure 2: Interaction table (clr) by gender: males (left), females (right). Colors enhance importance of interaction (I). Red: strong positive I. Ocre: medium positive I. Grey: low positive I. White: no I. Pale blue: low negative I. Aquamarine: medium negative I. Dark blue: strong negative I.

health of the couple”, is an exception to this rule for which females show a better feeling about the condom than the males. At the same time, it is interesting that males believe more than women, that the contraceptive and the morning after pills increase the feelings of pleasure in women.

3.3 Associations between different perceptions of effectiveness of protective methods in different items.

Two parts of a composition are linearly associated if their values across the sample are proportional. To understand the associations present in a survey, a variation matrix is useful (Egozcue et al., 2013; Lovell et al., 2015). Association is measured as inversely proportional to the variance of the corresponding log-ratios, i.e. for some couple of cells X_{ij} , $X_{k\ell}$, if

$$\text{Var} \left[\ln \frac{X_{ij}}{X_{k\ell}} \right] \approx 0,$$

then the association between cells X_{ij} and $X_{k\ell}$ can be strong. The degree of association of cells in the 30×30 table is visualised in Figure 3. It shows, as expected, strong associations of P1 (protects

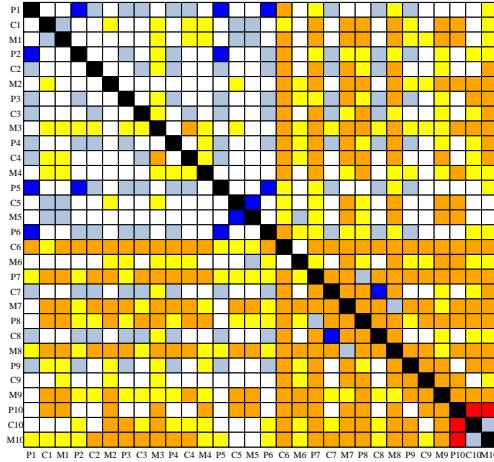


Figure 3: Heatmap of variation matrix. Colors enhance importance of association (A). Red: No A. Orange: negligible A. Yellow: very low A. White: low A. Pale blue: weak A. Dark blue: strong A.

against STI) with P2 (protects against pregnancies), P5 (protects against AIDS) and P6 (evidence

of interest in protecting the health of the partner). The same happens with C7 (increases man pleasure) and C8 (increases woman pleasure).

There are also strong associations between P2 (protects against pregnancies) and P5 (protects against AIDS), as well as between P6 (evidence of interest in protecting the health of the partner) and P5 (protects against AIDS), which also appear well associated in the heatmap of Figure 3. This is not in contradiction with the results in Figure 1, where interactions appear with opposite sign. The interaction points out that the protective value of preservatives against AIDS is better perceived than against pregnancy, and the association between P2 and P6 indicates that a good/weak perception of both protective measures is coupled. Therefore, a possible interpretation is that the score assigned to P2 and P6 measures the perception of “protection”, independently of protection against what.

3.4 Visualization: CoDa-biplot.

A biplot is a two-dimensional projection of the variables and data points of a data set. When used with compositional data the data set is first clr-transformed and then centered so that clr-variables are jointly represented with the data points (Aitchison, 1983; Aitchison and Greenacre, 2002). Here, the mean table obtained in the survey is taken as a compositional data set, where the three methods P, C, M are the compositional parts and the situations in which they are evaluated as realizations of the composition. Comparing the biplot in Figure 4 with the clr-interactions in

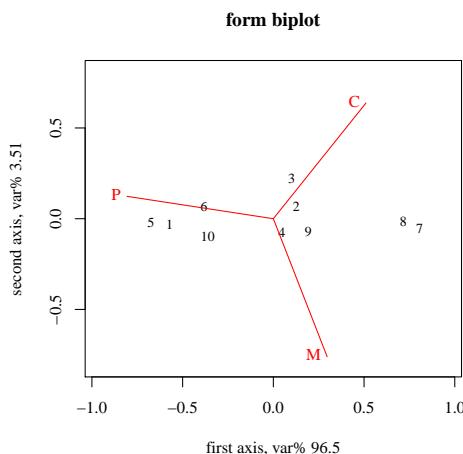


Figure 4: Form biplot of mean table. Points correspond to items in the survey, arrows to protective measures.

Figure 1 and the associations in Figure 3, it can be observed that following the arrow corresponding to P, items (points) 1 and 5 are close together, as well as items 7 and 8 at the other end, visualising the sign of the interaction and the degree of association. The biplot shows that there is essentially a one dimensional axis that classifies the items, as the first principal axis explains 96.5% of the total variance. It can be identified with the balance $\sqrt{2/3} \ln(P/\sqrt{C \cdot M})$, which is parallel to it. The preservatives (P) are positively perceived for items placed on the negative values of the principal axis, whilst items 7 and 8 (increase feelings of pleasure in man and woman) are negatively perceived. The second principal axes is essentially dominated by the balance $\sqrt{1/2} \ln(C/M)$, but its variability is small (3.51%).

The biplots by sex are represented in Figure 5. They are all very similar, although subtle differences

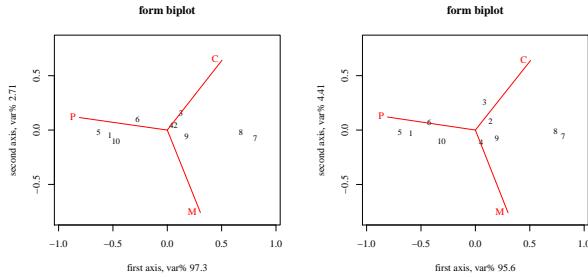


Figure 5: Form biplot of mean table. Points correspond to items in the survey, arrows to protective measures; Males (left) and females (right).

can be observed. A global difference is that the second principal axis represents a larger percentage of variance for females (4.40%) than for males (2.71%), thus indicating a greater differentiation in the perception by females. For instance, items 2 (protect from pregnancy), 3 (provide peace during and after intercourse), 4 (economically accessible), 9 (are easy to use correctly) are more dispersed along the second axis for females in accordance to the fact that women are more direct users/consumers of contraceptives and morning pills than males.

4 Conclusions

- Condoms, in general, are very positively valued (a) as protection against sexually transmitted infections/AIDS; (b) for their lack of side effects; and (c) as a sign of interest in protecting the health of the couple, inversely to what occurs in the case of pills.
- Hormonal methods clearly have a better acceptance than the condom regarding perceiving greater pleasure in intercourse, whereas the reverse occurs when it comes to feeling protected against STI. But the condom also counts in its favor with the good evaluation that it does not produce side effects and that its use shows interest in protecting the health of the couple, contrary to what is observed in the assessment of hormonal methods. The results suggest that, in general, when confronted with the three possibilities, the preferences of the participants in the study are inclined towards the condom against the contraceptive pill and the postcoital pill, but in the case of men with greater intensity.
- Examining results by gender, it can be observed that the assessments of women are less extreme than those of the men in 5 of the 6 mentioned items, but are more pronounced when they consider that the use of the condom demonstrates interest to protect the health of the couple.
- Survey data can advantageously be treated as compositional tables, as the relevant information are the ratios between the different responses and items.
- Interactions between responses and items give insight into the importance respondents give to the different possibilities.
- Standard tools of Compositional Data Analysis are useful to detect and assess associations between items/responses. Main tools are: CoDa-variation array (bivariate associations); CoDa-biplot, visualization in reduced dimension. Decomposition of the mean compositional table into independent and interaction tables.

Acknowledgements

The authors acknowledge financial support by the Spanish Ministry of Education and Science under project ‘CODA-RETOS’ (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)), by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project ‘COSDA’ (Ref. 2014SGR551), as well as by the University of Girona (MPCUDG2016/032).

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. London (UK): Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
- Aitchison, J. and M. Greenacre (2002). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 51(4), 375–392.
- Barceló-Vidal, C. and J.-A. Martín-Fernández (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics* 45, 57–71.
- Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 – The VII Annual Conference of the International Association for Mathematical Geology*, Cancun (Mex), pp. 20 p.
- Egozcue, J. J., D. Lovell, and V. Pawlowsky-Glahn (2013). Testing compositional association. In P. F. K. Hron and M. Templ (Eds.), *Proceedings of the 5th Workshop on Compositional Data Analysis – CoDaWork 2013*. ISBN: 978-3-200-03103-6, <http://coda.data-analysis.at/>.
- Egozcue, J. J., V. Pawlowsky-Glahn, M. Templ, and K. Hron (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics - Theory and Methods* 44(18), 3978–3996.
- Grimley, D., J. Prochaska, W. Velicer, and G. E. Prochaska (1995). Contraceptive and condom use adoption and maintenance: a stage paradigm approach. *Health Education Quarterly* 22, 20–35.
- Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler (2015, 03). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput Biol* 11(3), e1004075.
- Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* 15(5), 384–398.
- Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. Statistics in practice. John Wiley & Sons, Chichester UK. 272 pp.
- Prat, F., M. Planes, M. Gras, and M. Sullman (2012). Stages of change and decisional balance for condom use with a heterosexual romantic partner. *Journal of Health Psychology* 17(8), 1193–1202.
- Prat, F., M. Planes, M. Gras, and M. Sullman (2016). Perceived pros and cons of condom use as predictors of its consistent use with a heterosexual romantic partner among young adults. *Current Psychology* 35, 13–21.

Subcompositional coherence and the compositional complex

S. Sinari¹, D. Billheimer², and E.J. Bedrick³

¹BIO5 Institute, The University of Arizona, shripad@email.arizona.edu

²Epidemiology and Biostatistics, BIO5 Institute, The University of Arizona, dean.billheimer@email.arizona.edu

³Epidemiology and Biostatistics, BIO5 Institute, The University of Arizona,
edwardjbedrick@email.arizona.edu

Abstract

Aitchison (2001) defines a compositional problem as one in which only the relative sizes of the multivariate sample components are relevant. Formulated in terms of examples, the principle of subcompositional coherence is fundamental to the analysis of relative abundance data. In this paper, the notion of a compositional complex is introduced to provide a geometric interpretation of the principle of subcompositional coherence.

The complex is a combinatorial reconstruction of the closed n -dimensional simplex from all its faces, Gelfand and Manin (2003); Mac Lane (1978). Essentially, it is a collection of faces of the closed simplex and maps between these faces, also called face maps. The criterion of compatibility with these face maps is then defined to be the property of subcompositional coherence for a family of distributions on these faces.

Criteria for a family of distributions to provide a subcompositionally coherent model are developed in terms of their moments and characteristic function. The logistic normal distributions, logistic skew normal distributions and the distributions induced by elliptically symmetric distributions are shown to satisfy these criteria, substantially expanding the class of distributions available on the simplex for modeling relative abundance data.

Key words: compositional analysis, subcompositional coherence, simplicial complex, faces, face maps.

1 Introduction

Multivariate compositions are data where only the relative sizes of the components are relevant. Such data are commonplace today. Examples include RNA-Seq, metagenomics, mineral compositions of rock, ethnic compositions of towns or analysis of family budgets just to mention a few.

Scale invariance is an equivalent condition to the assumption of only relative sizes being relevant. This means any analysis done using either p or αp , where α is a positive constant, should yield same results for any composition p . A consequence of scale invariance is subcompositional coherence.

Subcompositional coherence is the requirement that if q is a composition of sub-components of p , then inference based on the components in q obtained by using either q or p should be the same.

In this paper we interpret subcompositional coherence in terms of the points on the closed unit simplex. We will begin by using the example of a tetrahedron viewed as a three dimensional realization of the four dimensional unit simplex. For the tetrahedron, we demonstrate in detail the concepts of face and face maps. The compositional complex arises as a convenient notational scheme to deal with all the faces and the face maps. We then use the complex to develop mathematical criteria for subcompositional coherence, showing in general that large classes of distributions on the unit simplex provide subcompositionally coherent models. These include the logistic normal distributions as well as those induced from skew-normal and elliptically symmetric distributions by the additive log ratio transform.

2 Example of the tetrahedron

In order to make the notation for a complex in general more accessible, we begin with the illustrative case of a tetrahedron viewed as a four dimensional unit simplex. The tetrahedron provides more "wiggle room" due to its higher dimensions than a two dimensional simplex which is closed triangle.

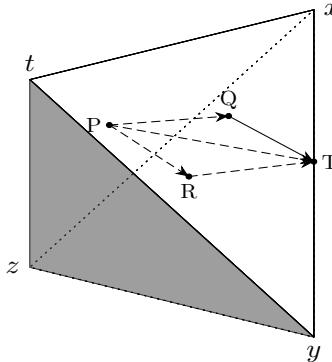


Figure 1: Tetrahedron with vertices x, y, z and t . The point P is in the interior of the tetrahedron. R and Q are the projection of P from t and z respectively. T is projection of R of from z and Q from t .

The unit simplex formed by the four standard basis vectors in \mathbb{R}^4 can be visualized as a tetrahedron Δ^3 with vertices x, y, z and t in \mathbb{R}^3 . The four dimensional co-ordinates (standard co-ordinates in \mathbb{R}^4) are in one-to-one correspondence with the three dimensional tetrahedron using the barycentric co-ordinates. Any 3 distinct vertices give a 2-dimensional face of this tetrahedron. For example, the face $\mathfrak{F}_{y,z,t}$ formed by y, z and t is shaded in gray in the Figure (1). Similarly any two distinct vertices form a 1-dimensional face, which is commonly referred to as an *edge* of the tetrahedron.

The vertices themselves can be thought of as 0-dimensional faces of the edge to which they belong. In the same spirit, we will consider the tetrahedron as the 3-dimensional face of itself.

A face map from the tetrahedron to the face formed by vertices x, y and z is given by

$$F_{x,y,z}^t = \left(\frac{x}{1-t}, \frac{y}{1-t}, \frac{z}{1-t} \right) = \left(\frac{x}{x+y+z}, \frac{y}{x+y+z}, \frac{z}{x+y+z} \right)$$

The last equality is due to the fact that $x + y + z + t = 1 ; \forall (x, y, z, t) \in \Delta^3$. Similarly if one is interested only in analysis of the two components x and y then one can take the face map

$$F_{x,y}^{z,t}(x, y, z, t) = \left(\frac{x}{1-t-z}, \frac{y}{1-t-z} \right) = \left(\frac{x}{x+y}, \frac{y}{x+y} \right)$$

Note that the face maps thus defined satisfy the following equality

$$F_{x,y}^{z,t} = F_{x,y}^z \circ F_{x,y,z}^t = F_{x,y}^t \circ F_{x,y,z}^z \quad (1)$$

We can visualize these operations in the figure above. Consider the point P in the interior of the tetrahedron. Let $R = F_{x,y,z}^t(P)$, $Q = F_{x,y,z}^z(P)$ and $T = F_{x,y,z}^t(P)$. Then Equation (1) means that any combination of face maps from the tetrahedron to the xy -edge as well as the direct descent to the xy -edge maps P to T . This leads to concept of a commutative diagram where functions or compositions thereof, with same starting and endpoints gives the same results. Another way to express Equation (1) is to say that the following diagram commutes:

$$\begin{array}{ccc} \Delta^3 & \xrightarrow{F_{x,y,z}^t} & \mathfrak{F}_{x,y,z} \\ F_{x,y,t}^z \downarrow & \searrow F_{x,y}^{z,t} & \downarrow F_{x,y}^z \\ \mathfrak{F}_{x,y,t} & \xrightarrow[F_{x,y}^t]{\quad} & \mathfrak{F}_{x,y} \end{array} \quad (2)$$

Thus if we take the collection of the distinct 2-dimensional faces denoted by

$$\Delta^2 := \mathfrak{F}_{x,y,z} \sqcup \mathfrak{F}_{x,y,t} \sqcup \mathfrak{F}_{x,z,t} \sqcup \mathfrak{F}_{y,z,t}$$

where \sqcup denotes a disjoint union, and the collection of maps

$$F_2^3 := \{F_{x,y,z}^t, F_{x,y,t}^z, F_{x,z,t}^y, F_{y,z,t}^x\}$$

then we get the complex

$$\Delta^3 \xrightarrow{F_2^3} \Delta^2$$

By taking the appropriate collections we can expand the complex to:

$$\Delta^3 \xrightarrow{F_2^3} \Delta^2 \xrightarrow{F_1^2} \Delta^1 \xrightarrow{F_0^1} \Delta^0$$

Another way to look at this complex is to consider the following lattice (diagram (3)) of subsets of the vertices of the tetrahedron.

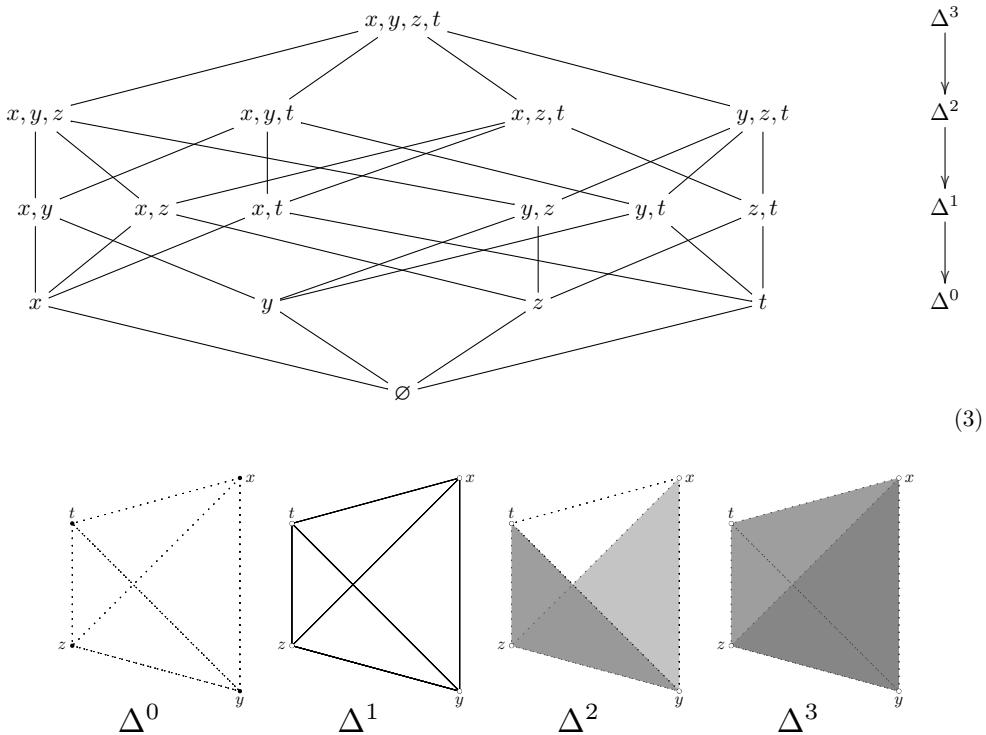


Figure 2: The figure shows the components or nodes of the complex. Going from left right we build the closed unit tetrahedron. Dotted lines and hollowed points are for visual aid only and do not form components of the node, as in Δ^2 and Δ^3 . We begin with the vertices in Δ^0 and then join them with solid lines which are components of Δ^1 , pasting the faces shown in Δ^2 and finally filling the skeleton with Δ^3 to get the solid closed tetrahedron. Only two of the four faces in Δ^2 have been shaded for visual clarity.

where subsets of same cardinality are in a single row and lines connecting them to nodes in the rows above indicate an inclusion relationship. Only partial lattice structure is shown for visual clarity. Any lines connecting beyond the immediate rows above or below are not drawn, although the full lattice should be thought of having this structure. For example, node " x, y " is not connected to the tetrahedron " x, y, z, t ". A simplex then can be thought as a map which takes each node to the face that is formed by the vertices in that node and the inclusion relationship to the corresponding face maps with the direction of the maps reversed. Thus each row represents the collection of faces of same dimension and hence the corresponding Δ^i and lines the corresponding F_j^i .

Remark 2.1. *The diagram (3) is called the Hasse diagram of the face poset of the unit simplex Wachs (2006). An important property of the diagram is the one-to-one correspondence between the nodes and the faces of the tetrahedron.*

Till now we did not assume if the faces were open or closed, i.e. the inequality defining the simplex is strict or not respectively. In other words, the three dimensional face of the tetrahedron is open if none of its components are zero. Note that all arguments done till now work with both the closed and the open faces of the tetrahedron. Since what follows in the remainder of this section is valid sometimes only for the open faces, we will consider only open faces for the remainder of this paper.

Each face of the tetrahedron can be endowed with a distribution of appropriate dimension from the same family. Let \mathfrak{D} represent this assignment operation. For example, \mathfrak{D} might be the assignment of the 3-dimensional logistic normal distribution denoted \mathfrak{D}_3 on the interior of the tetrahedron and a 2-dimensional logistic normal $\mathfrak{D}_{x,y,z}$ on the face $\mathfrak{F}_{x,y,z}$. We say that the assignment \mathfrak{D} gives subcompositionally coherent model on the tetrahedron if the distribution $(F_{x,y,z}^t)^*(\mathfrak{D}_3)$ is same as the distribution $\mathfrak{D}_{x,y,z}$, i.e., the following diagram commutes:

$$\begin{array}{ccc}
 \mathfrak{D}_3 & \xrightarrow{(F_{x,y,z}^t)^*} & \mathfrak{D}_{x,y,z} \\
 \downarrow (F_{x,y,t}^z)^* & \searrow (F_{x,y}^{z,t})^* & \downarrow (F_{x,y}^z)^* \\
 \mathfrak{D}_{x,y,t} & \xrightarrow{(F_{x,y}^t)^*} & \mathfrak{D}_{x,y}
 \end{array} \tag{4}$$

In fact, we want all possible such diagrams to commute or in terms of diagram (3) we want the image of the entire lattice to be commutative.

A few important remarks are in order.

Remark 2.2. *The process of endowing the distribution requires one to consider the open faces since the log ratio transforms are well defined only on the interior of the closed simplex, i.e. the open face.*

Remark 2.3. *We do not require the definition of composition for the collection $\{F^i\}_{i=1}^3$. Although, such as operation is well defined due to Equation (1), i.e.*

$$F_{i-2}^{i-1} \circ F_{i-1}^i : \Delta^i \rightarrow \Delta^{i-2}$$

is a well defined.

Remark 2.4. *The lattice of distributions with the maps induced by face maps (4) is commutative. However, in generality, the maps that make the lattice commute need not be induced by the face maps. We will only require that the distributions on the nodes and the corresponding maps give us a commutative diagram.*

Remark 2.5. *Notice that we have conveniently used an implicit ordering given by $x < y < z < t$ to make the notation as non-redundant as possible.*

3 Compositional complex

We begin development of notation and definitions in general. In light of remark (2.5), we will use the natural ordering of the integers which indexes the standard basis of \mathbb{R}^{n+1} . We define a face and a face map of the open simplex. The collection of these faces and corresponding maps between such collections are defined followed by the complex.

Notation 3.1. $\{e_i\}_{i=1}^{n+1}$ will denote the standard basis of \mathbb{R}^{n+1} .

Notation 3.2. For any $n \in \mathbb{N}$ define

$$[n] := \{1, 2, \dots, n\}$$

and

$$\binom{[n]}{k} := \{S \mid S \subset [n] \text{ and cardinality of } S \text{ is } k, \text{ i.e. } |S| = k \text{ and } S \text{ is ordered}\}$$

Remark 3.1. The notation (3.2) is meant to be suggestive of the cardinality of the set it is representing.

Notation 3.3. For a matrix A , A' will denote the transpose of A . A column vector $v \in \mathbb{R}^n$ will be considered both a vector and a $n \times 1$ dimensional matrix. In case of the later, v' denotes the corresponding $1 \times n$ row matrix. For any $0 < m \leq n$, m' will denote the "adjoint" of m , namely $m' = n - m$

Definition 3.1. If $S \subset [n+1]$ and $S = \{i_1 < i_2 < \dots < i_{m+1}\} \in \binom{[n+1]}{m+1}$ then

$$\Delta_S := \left\{ \mathbf{x} \in \sum_{j=1}^{m+1} x_{i_j} e_{i_j} : 0 < x_{i_j} < 1 \forall i_j \right\} \quad (5)$$

For $m = n$, we denote the simplex by Δ^n .

We call the set defined in Equation (5) to be the S face of the standard simplex. Thus a m -dimensional face is defined by $S \in \binom{[n+1]}{m+1}$.

Definition 3.2. Let $U \subset V \in \binom{[n+1]}{m+1}$. If U^c denotes the complement of U in V , then a face map from V face to the U face is defined by:

$$\begin{aligned} F_U^V : \Delta_V &\rightarrow \Delta_U \\ x = (x_v)_{v \in V} &\mapsto \alpha(x_u)_{u \in U} \end{aligned}$$

$$\text{where } \alpha = \frac{1}{1 - \sum_{v \in U^c} x_v}.$$

Definition 3.3. For $0 < m \leq n$ define

$$\Delta^m := \bigsqcup_{U \in \binom{[n+1]}{m+1}} S_U$$

the disjoint union of the m -dimensional faces of Δ^n . We define, Δ^0 as the set of vertices of Δ^n .

Then there exist maps

$$\mathcal{F}^m : \Delta^m \rightarrow \Delta^{m-1}$$

such that:

Definition 3.4. For $V \in \binom{[n+1]}{m+1}$

$$\mathcal{F}^m|_V = \{F_U^V | U \subset V \text{ and } U \in \binom{[n+1]}{m}\}$$

This gives us the complex:

$$\Delta^n \xrightarrow{\mathcal{F}^n} \Delta^{n-1} \xrightarrow{\mathcal{F}^{n-1}} \Delta^{n-2} \xrightarrow{\mathcal{F}^{n-2}} \dots \xrightarrow{\mathcal{F}^2} \Delta^1 \xrightarrow{\mathcal{F}^1} \Delta^0 \quad (6)$$

Notation 3.4. We denote this complex as the tuple (Δ, \mathcal{F}) .

4 Subcompositional coherence

A model on the complex (Δ, \mathcal{F}) , is a tuple $(\mathcal{M}, \mathfrak{p})$, consisting of a family of probability distributions \mathcal{M} and maps \mathfrak{p} (see [Wit and McCullagh \(2001\)](#); [McCullagh \(2002\)](#)).

where

$$\mathcal{M} = \{\mathcal{M}^i = \mathcal{M}(\Delta^i) = \sqcup_{k=i_1}^{i_j} \mathcal{M}_k^i\}_{i=1}^n$$

and

$$\mathfrak{p} = \{\mathfrak{p}^i = \mathcal{M}(\mathcal{F}^i)\}_{i=2}^n.$$

Here we require that \mathcal{M}_k^i , the distribution on the k -th face of Δ^i , belong to the same family of distributions $\forall i, 0 < i \leq n$ and $\forall k, i_1 \leq k \leq i_j$ giving the complex:

$$\mathcal{M}^n \xrightarrow{\mathfrak{p}^n} \mathcal{M}^{n-1} \xrightarrow{\mathfrak{p}^{n-1}} \mathcal{M}^{n-2} \xrightarrow{\mathfrak{p}^{n-2}} \dots \xrightarrow{\mathfrak{p}^2} \mathcal{M}^1$$

Let \mathcal{M} be the assignment

$$\mathcal{M} : (\Delta, \mathcal{F}) \rightarrow (\mathcal{M}, \mathfrak{p})$$

Such a model is said to be subcompositionally coherent if it satisfies:

$$\begin{array}{ccc} \Delta^m & \xrightarrow{\mathcal{F}^m} & \Delta^{m-1} \\ \downarrow \mathcal{M} & \quad \square \quad & \downarrow \mathcal{M} \\ \mathcal{M}^m & \xrightarrow{\mathfrak{p}^m} & \mathcal{M}^{m-1} \end{array} \tag{7}$$

This condition in terms of the vertices S of an arbitrary face Δ_S can be formulated as follows. Let $x, y, z, t \in S$ be distinct vertices of the Δ_S . Then the diagram (7) says that the following Equation is true:

$$\mathfrak{p}_{x,y}^{z,t} = \mathfrak{p}_{x,y}^z \circ \mathfrak{p}_{x,y,z}^t = \mathfrak{p}_{x,y}^t \circ \mathfrak{p}_{x,y,t}^z \quad \forall x, y, z, t \in S \tag{8}$$

Remark 4.1. In case of the tetrahedron, this definition is equivalent to one obtained from Figure (4) with $(F^i)^*$, the collection of maps such as $(F_{x,y,z}^t)^*$ induced by each member of F^i , taking the role of \mathfrak{p} .

Remark 4.2. These definitions can be formulated using the notions of category theory. The simplices (the simplicial category) is the manifestation of the category of all finite ordinals with weakly monotone functions as morphisms (see section 5 on "The Simplicial Category" in chapter VII on "Monoids" in [Mac Lane \(1978\)](#)). The assignment \mathcal{M} is a contravariant functor that takes the lattice of subsets of the vertices to the category whose objects are distributions from a single family and morphisms are projections onto lower dimensions. The coherence theorem of monoidal categories (see section 2 on "Coherence" in chapter VII on "Monoids" in [Mac Lane \(1978\)](#)) can then be asked of the subcategory that is the image of \mathcal{M} . This is subcompositional coherence.

5 Criteria for a distribution to be subcompositionally coherent

Having formulated subcompositional coherence as the compatibility of distributions, in the sense of Equation (8), we now determine what this condition means for distributions with finite second moments that are induced from the Euclidean space via a log ratio transform. In this paper, we will use only the additive log ratio (*alr* for short). It is well known that compositional analysis is independent of the log ratio transforms used. We first demonstrate that the induced distribution is independent of the choice of the component used to define *alr*. Second, let X_m is a random variable on the Euclidean space \mathbb{R}^m . We will determine the conditions on the family of distributions to which X_m belongs to induce a subcompositionally coherent model on the complex. For sake of clarity of notation, the following arguments are done using a single vertex j in an arbitrary face Δ_S . However, the results are true for any subset U of the vertices of Δ_S .

The distribution of a random vector X_m on \mathbb{R}^m can be characterized by its *characteristic function*:

$$\begin{aligned}\phi_{X_m} : \mathbb{R}^m &\longrightarrow \mathbb{C} \\ t &\longmapsto \mathbb{E}(e^{itX_m})\end{aligned}\tag{9}$$

The additive log ratio (*alr*) transformation with respect to j -th co-ordinate on Δ_S as:

$$\begin{aligned}alr^j : \Delta_S &\rightarrow \mathbb{R}^m \\ x &\mapsto (\log(\frac{x_{i_1}}{x_{i_j}}), \log(\frac{x_{i_2}}{x_{i_j}}), \dots, \log(\frac{x_{i_{j-1}}}{x_{i_j}}), \log(\frac{x_{i_{j+1}}}{x_{i_j}}), \dots, \log(\frac{x_{i_{m+1}}}{x_{i_j}}))\end{aligned}\tag{10}$$

The pullback

$$(alr^j)^*(\phi_{X_m}) := \phi_{X_m} \circ alr^j$$

defines a multivariate random vector X^j on Δ_S with the characteristic function given by $(alr^j)^*(\phi_{X_m})$.

Proposition 5.1. X_m^j is well defined and independent of the choice j .

Proof. Consider the map

$$alr^l \circ alr^{-k} : \mathbb{R}^m \rightarrow \mathbb{R}^m$$

The Jacobian matrix of this map is an identity $m \times m$ matrix if $k = l$. Otherwise it is given by:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & 1 & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ -1 & -1 & \cdots & -1 & \cdots & -1 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}_l$$

where the sub matrix obtained by removing the l -th row and the k' -th column is an $(m-1) \times (m-1)$ identity matrix. The k' -th column is $(0, \dots, -1, \dots, 0)^t$ with the only non-zero entry -1 is in the l -th row. The l -th row is a vector of -1 's, namely $(-1, \dots, -1)^t$. Further $k' = k$ if $k < l$ and $k' = k-1$ otherwise.

Thus one can see that X_m^l and X_m^k follow the same distribution since the absolute value of the determinant of the Jacobian is 1. \square

Let Δ_S be an arbitrary face in the simplex Δ^n and let $S_j = \{j\}^c$, i.e set theoretic complement of the j -th vertex of Δ_S . Then the following diagram commutes for $k \in S_j$:

$$\begin{array}{ccc}
 \Delta_S & \xrightarrow{p_j} & \Delta_{S_j} \\
 \downarrow alr^k & \text{Q} & \downarrow alr^k \\
 \mathbb{R}^m & \xrightarrow{\hat{\pi}_j} & \mathbb{R}^{m-1}
 \end{array} \tag{11}$$

where $\hat{\pi}_j$ is the projection in \mathbb{R}^m onto the subspace with j -th coordinate equal to 0 and the small circular arrow in the middle of the diagram denotes commutativity.

To ease the intuition let us consider what this means in terms of the tetrahedron.

If $j = t$ and $k = z$ then the diagram (11) implies

$$\begin{aligned}
 alr^z(x, y, z, t) &= (\log \frac{x}{z}, \log \frac{y}{z}, \log \frac{t}{z}) \\
 \therefore \hat{\pi}_t(alr^z(x, y, z, t)) &= (\log \frac{x}{z}, \log \frac{y}{z})
 \end{aligned}$$

and

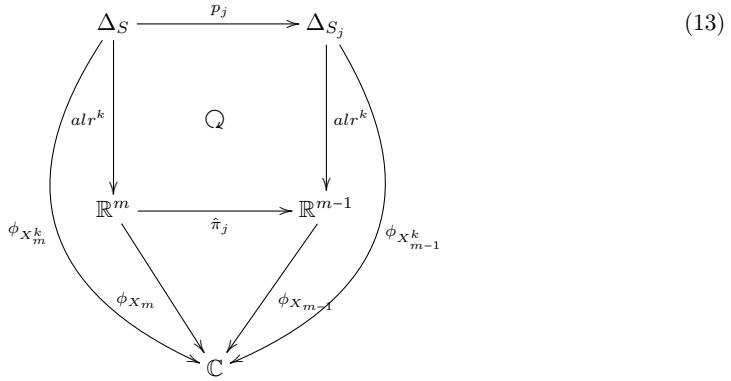
$$\begin{aligned}
 p_{x,y,z}(x, y, z, t) &= (\frac{x}{1-t}, \frac{y}{1-t}, \frac{z}{1-t}) \\
 \therefore alr^z(p_{x,y,z}(x, y, z, t)) &= \left(\log\left(\frac{\frac{x}{1-t}}{\frac{z}{1-t}}\right), \log\left(\frac{\frac{y}{1-t}}{\frac{z}{1-t}}\right) \right) \\
 &= \left(\log \frac{x}{z}, \log \frac{y}{z} \right)
 \end{aligned}$$

Remark 5.1. This just means that alr is a map of complexes from (Δ, \mathcal{F}) to (\mathfrak{R}, π) . Here (\mathfrak{R}, π) is the complex whose nodes are \mathbb{R}^n and π^n is the collection of all co-ordinate projections of \mathbb{R}^n onto \mathbb{R}^{n-1} . The choice of k is irrelevant due to proposition (5.1).

Let X_{m-1} be the random variable $\hat{\pi}_j^*(X_m)$ on \mathbb{R}^{m-1} . Thus

$$\phi_{X_m} = \phi_{X_{m-1}} \circ \hat{\pi}_j \tag{12}$$

then the following diagram commutes:



The commutativity of the diagram (13) is equivalent to the condition that the *characteristic functions* of the family of distributions induced by X_m^k and X_{m-1}^k satisfy the functional equation:

$$\begin{aligned}\phi_{X_m^k}(\mu, \Sigma, t) &= \phi_{X_{m-1}^k} \circ p_j(\mu, \Sigma, t) \\ &= \phi_{X_{m-1}^k}(p_j(\mu), P'_j \Sigma P_j, t) \\ &= \phi_{X_{m-1}^k}(\mu_j, \Sigma_j, t)\end{aligned}\quad (14)$$

where μ and Σ are the first two moments of the family of distributions. Notationally, $\mu_j = p_j(\mu)$ and $\Sigma_j = P'_j \Sigma P_j$ for a matrix representation P_j of p_j .

Thus we have shown that if $(\mathcal{M}, \mathfrak{p})$ is a coherent model with respect to the complex (\mathfrak{R}, π) , then the model induced by alr on the complex (Δ, \mathcal{F}) is subcompositionally coherent.

We know that the normal family of distributions satisfy Equation (12) and hence the induced distributions, namely, the family of logistic normal distributions provide a subcompositional coherent model.

Equation (11) in section 5 of [Azzalini and Capitanio \(1999\)](#) shows that skew normal distributions also satisfy Equation (12) and hence the induced distributions, namely, the family of logistic skew normal distributions provide a subcompositional coherent model.

Now elliptically symmetric distributions are characterized by characteristic functions which satisfy the functional equation:

$$\phi(\mu, \Sigma, t) = \phi(t' \Sigma t) \quad \forall t \in \mathbb{R}^m$$

That is the characteristic function is only the function of the quadratic form $t' \Sigma t$ on \mathbb{R}^m .

It is easy to see by taking $P_j t$ instead of t , where P_j is the projection matrix in \mathbb{R}^m with respect to the j -th co-ordinate, that elliptically symmetric distributions satisfy Equation (12) and thus induce subcompositionally coherent model in the sense of diagram (7).

Remark 5.2. *Although we do not prove it here, mixtures of subcompositionally coherent models are also subcompositionally coherent.*

6 Conclusion

Subcompositional coherence has a geometrical interpretation represented conveniently using the complex structure. This interpretation give criteria to determine if a family of distributions can yield subcompositionally coherent model on the simplex. The criteria are equality of distributions chosen on a lower dimensional face of the simplex with the one induced by any face map from higher dimensions. A procedure to construct such models using additive log ratio and distributions on the Euclidean space is presented. The procedure shows why the most commonly used family of distributions for modeling compositions, namely the family of logistic normal distributions, have the property of subcompositional coherence. The inter-dimensional nature of the subcompositional coherence is highlighted.

The complex structured explored here as well as the Hilbert space structure introduced by Billheimer et al. (2001) on the simplex are consequences of scale invariance which is naturally connected to the projective geometry. This connection will be explored in future work.

References

- Aitchison, J. (2001). Simplicial inference. In M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics*. Providence, Rhode Island: American Mathematical Society.
- Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3), 579–602.
- Billheimer, D., P. Guttorp, and W. F. Fagan (2001, December). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association* 96(456), 1205–1214.
- Gelfand, S. I. and Y. I. Manin (2003). *Methods of homological algebra* (Second ed.). Springer Monographs in Mathematics. Berlin, Heidelberg: Springer-Verlag, Berlin.
- Mac Lane, S. (1978). *Categories for the Working Mathematician*, Volume 5 of *Graduate Texts in Mathematics*. New York, NY: Springer New York.
- McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 1225–1267.
- Wachs, M. L. (2006). Poset topology: tools and applications. *arXiv preprint math/0602226*.
- Wit, E. and P. McCullagh (2001). The extensibility of statistical models. In M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics*. Providence, Rhode Island: American Mathematical Society.

Compositional data analysis of element concentrations of simultaneous size-segregated PM measurements

A. Speranza¹, R. Caggiano¹, S. Margiotta¹ and

V. Summa¹; *antonio.speranza@imaa.cnr.it*¹IMAA, Istituto di
Metodologie per l'Analisi Ambientale, CNR, 85050 Tito Scalo,
PZ, Italy

Abstract

This preliminary study presents the application of compositional data analysis on element concentrations of size-segregated PM simultaneous measurements. There is a growing interest in particulate matter (PM) due to its impact on human health, air quality and global climate change. PM is a mixture of particles suspended in the air which differ in size, chemical composition and emission sources. The assessment of the chemical composition and size distribution of PM in relation to its possible emission sources is a starting point to plan actions aimed at mitigating the levels of PM in the environment and to protect public health. Selected chemical elements have been linked to specific sources of PM including mineral matter, sea-spray, and fuel-oil combustion. However, the identification of a set of elements useful in the discrimination of specific natural and anthropogenic sources of mineral matter has proven to be problematic, since they can have the same range of chemical elements in common.

This study considers PM element concentrations of a typical suburban background site with (in-dust days) and without (non-dust days) the contribution from a Saharan dust event. The selected elements were Al, Si, Ca, Fe, Ti, Mg, Sr, commonly interpreted as related to mineral matter. The element concentrations of PM related to in-dust and non-dust days have been converted into two compositional data sets based on percentage weight. The compositional data analysis provides evidence that the two compositional data sets are statistically distinct. This outcome shows that the Saharan dust event (in-dust days) together with local sources of mineral matter (non-dust days) can determine the chemical composition of PM. Therefore, compositional data analysis allows the study of environmental sites effected by natural sources of mineral matter (e.g. Saharan dust).

Keywords: simultaneous PM measurements, PM₁₀, PM_{2.5}, PM₁, Saharan dust.

1 Introduction

Particulate matter (PM) is considered responsible for negative effects on public health (WHO, 2006), the ecosystem and the Earth's climate (Caggiano and others, 2001; IPCC 2013). PM consists of an air-suspended mixture of particles with a wide variety of sizes, shapes and chemical compositions. These particles can be inhaled and deposited in the respiratory system with consequent delivery of potentially toxic elements to the body (WHO, 2012). Indeed, a consistent number of studies have linked PM exposure to respiratory and cardiovascular diseases (Pope and others, 2004, Dockery and Pope, 1994). Moreover, the Earth's climate can be affected by PM due to its ability to absorb/scatter solar radiation and to modify the properties of weather clouds (Prospero, 2007).

The assessment of the chemical composition of PM and of its size distribution in relation to its possible emission sources is a starting point to plan actions aimed at mitigating levels of PM to protect the environment and public health (Putaud and others, 2010). In the European context, selected sets of chemical elements have been attributed to specific sources of PM. The elements Al, Si, Ca, Fe, Ti, Mg, Sr, have been mainly linked to mineral matter and African dusts, while Na, Cl have been mainly associated with marine sources. Whereas V and Ni have been mainly related to fuel and oil combustion sources (Viana and others, 2008). However, the identification of a set of elements useful in the discrimination of specific natural source of mineral matter (e.g. such as African dusts and fugitive dusts) and characteristic anthropogenic source of mineral matter (e.g. resuspended road dusts and dust from construction/demolition activities), has proven to be problematic as these sources have the same set of elements in common (Thorpe and Harrison, 2008). Nowadays, there is an increasing scientific interest in simultaneous size-segregated PM measurements (i.e. PM_{10} , $PM_{2.5}$ and PM_1 aerosol particles with aerodynamic diameters smaller than 10, 2.5 and 1 μm , respectively). Since, it has been demonstrated that the assessment of the chemical composition of these PM mass concentrations can be an effective tool in the identification of the contributions of several sources of PM (Spindler and others, 2010). Indeed, simultaneous size-segregated PM measurements have been performed on a variety of environmental sites such as urban background and traffic point sites (Rogula-Kozłowska and others, 2013), urban traffic and suburban background sites (Matassoni and others, 2011), Nordic background site and wild fire episodes (Makkonen and others, 2010), urban roadsides, urban background and rural sites (Yin and Harrison, 2008) and industrial sites (Chiari and others, 2005, 2004) among others (Speranza and others, 2016, 2014).

Since the increasing scientific interest in this topic, the aim of this study is to apply compositional data analysis to element concentrations of PM_{10} , $PM_{2.5}$ and PM_1 simultaneous measurements. Compositional data analysis started with Aitchison (1982, 1986) and has since undergone several developments and many practical applications which have led it to be considered as a consolidated technique in geosciences (e.g. Pawlowsky-Glahn and Buccianti 2002; Buccianti and Pawlowsky-Glahn 2006; Aitchison and Egozcue, 2005; Pawlowsky-Glahn and Buccianti 2011, etc.).

The main objectives of this study are: a) to use the centering and rescaling technique to improve the visualization in a triangular diagram of the element concentrations of PM_{10} , $PM_{2.5}$ and PM_1 simultaneous measurements, and b) to evaluate the compositional differences of PM relating to characteristic environmental sites with and without the effect of a Saharan dust event and to highlight the influencing processes.

2 Methodology

The mineral tracer (elements) concentrations of PM_{10} , $PM_{2.5}$ and PM_1 simultaneous measurements as reported in literature have been considered and refer to a suburban background site with (in-dust days) and without (non-dust days) the contribution of a Saharan dust episode (Matassoni and others, 2011). The selected mineral tracers are Al, Ti, Si, Ca, Mg, Fe, Sr, which have been mostly and commonly interpreted as related to mineral matter (Viana and others, 2008). The PM_1 solely for the Sr for in-dust and non-dust days was below the detection limits. The compositional data set was modified and completed using the imputation strategy described by Martín-Fernández and others (2003) (Pawlowsky-Glahn and Buccianti, 2011).

2.1 Compositional data and sample space

The sample space of a compositional observation with three components is the unit simplex

$$S_c^3 = \left\{ \mathbf{x} = (x_1, x_2, x_3) \mid x_j > 0, j = 1, 2, 3; x_1 + x_2 + x_3 = c \right\} \quad (1).$$

(Pawlowsky-Glahn and Buccianti, 2002). PM₁₀, PM_{2.5} and PM₁ simultaneous measurements are divided in terms of relative fractions as coarse [Eq. (2)], intermodal [Eq. (3)] and submicron, PM₁, mass concentrations (e.g. Colbeck and others, 2011; Kegler and others, 2001; Lundgren and others, 1996, etc.).

$$PM_{10-2.5} = PM_{10} - PM_{2.5} \quad (2)$$

$$PM_{2.5-1} = PM_{2.5} - PM_1 \quad (3)$$

These fractions are converted into compositions based on weight proportions following the strategy suggested by Aitchison (2005), Buccianti and Pawlowsky-Glahn (2005), Pawlowsky-Glahn and Egozcue (2006).

$$\mathbf{x} = \left(\frac{PM_{10-2.5}}{PM_{10}}, \frac{PM_{2.5-1}}{PM_{10}}, \frac{PM_1}{PM_{10}} \right) \% \quad (4).$$

The compositional variables of this vector are non-negative and they sum to a constant c=100, [(Eq. 1)]. Compositional data is cast into the form of a matrix where i rows represent the mineral elements and j columns represent the compositional variables.

2.2 Transformation of compositional data

The compositional data is transformed into co-ordinates using ilr (isometric log-ratio) transformations (Egozcue and others, 2003) [Eq. (5)] and [Eq. (6)].

$$ilr_1 = \frac{1}{\sqrt{2}} \ln \left(\frac{PM_{10-2.5}}{PM_{2.5-1}} \right) \quad (5)$$

$$ilr_2 = \frac{1}{\sqrt{6}} \ln \left(\frac{PM_{10-2.5} PM_{2.5-1}}{PM_1^2} \right) \quad (6)$$

The isometric co-ordinates ilr₁ and ilr₂ are inversely transformed by

$$\mathbf{x} = C \left(\exp \left(\frac{ilr_2 + ilr_1}{\sqrt{2}} \right), \exp \left(\frac{ilr_2 - ilr_1}{\sqrt{2}} \right), \exp \left(-\frac{2ilr_2}{\sqrt{6}} \right) \right) \quad (7)$$

where C is the closure operation for a vector defined below [Eq. (8)]. This operation divides each component of the vector by the sum of its components, hence scaling the vector to the constant c (Pawlowsky-Glahn and others, 2007).

$$C(\mathbf{x}) = \left(\frac{cx_1}{x_1 + x_2 + x_3}, \frac{cx_2}{x_1 + x_2 + x_3}, \frac{cx_3}{x_1 + x_2 + x_3} \right) \quad (8)$$

2.3 Triangular diagram representation, centering and rescaling technique

The compositional data sets, their centres and confidence regions can be represented using a triangular diagram. Three examples of three-part compositions and isometric log-ratios are reported in Figure 1. The triangular diagram shows the contribution of the coarse size fraction to PM₁₀, i.e. (PM₁₀-PM_{2.5})/PM₁₀ ratio, the contribution of intermodal size fraction to PM₁₀, i.e. (PM_{2.5}-PM₁)/PM₁₀ ratio and the contribution of submicron size fraction to PM₁₀ i.e. PM₁/PM₁₀ ratio. The data is displayed by Graham and Midgley (2000). Compositional data is centred using the perturbation operator of the simplex (Aitchison, 1986). The perturbation operation is defined as the perturbation \mathbf{p} applied to a composition \mathbf{x} that produces the composition $\mathbf{v} = \mathbf{p} \oplus \mathbf{x}$, with \mathbf{v} , \mathbf{p} and \mathbf{x} vectors in S^3_c (Aitchison, 2005). By perturbing a vector \mathbf{x} by its inverse $\mathbf{x}^{-1} = (x_1^{-1}, x_2^{-1}, x_3^{-1})$, it is possible to locate any composition in the baricentre of the triangular diagram. Likewise, it is possible to centre a compositional data set of size n , $\{(x_{i1}, x_{i2}, x_{i3}), i=1,2,\dots,n\}$ using the inverse of its centre \mathbf{g}^{-1} defined as Equation (10) (Buccianti and others, 1999; Martín-Fernández and others, 1999).

$$\mathbf{g} = C(g_1, g_2, g_3) \text{ where } g_j = \left(\prod_{i=1}^n x_{i,j} \right)^{\frac{1}{n}}, j = 1, 2, 3 \quad (10).$$

The centering and rescaling of the compositional data allows improved visualization of compositions close to the boundary of the triangular diagram preserving the straight lines of the grid as well as statistical properties (von Eynatten and others, 2002).

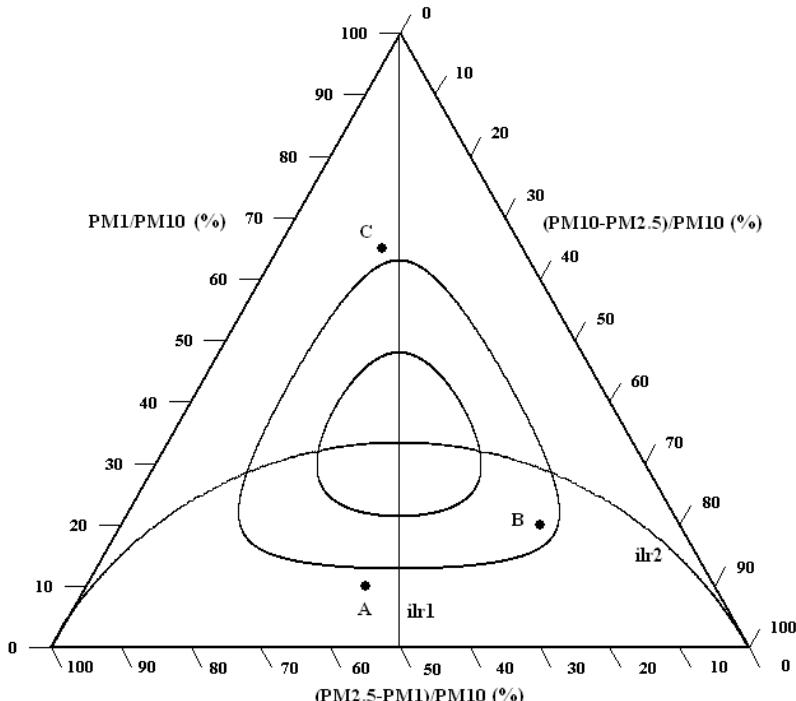


Figure 1: ilr axes and two circles in S^3 , the centre $\mathbf{g}=(1/3,1/3,1/3)*100\%$ and three compositions $A=(40, 50, 10)\%$, $B=(60, 20, 20)\%$ and $C=(15, 20, 65)\%$.

3 Results

In Figure 2, the three part compositional data of non-dust days are displayed towards the lower right corner of the triangular diagram with high values of $(PM_{10}-PM_{2.5})/PM_{10}$ ratio (coarse component) and low values of $(PM_{2.5}-PM_1)/PM_{10}$ ratio (intermodal component) and PM_1/PM_{10} ratio (submicron component). The coarse component is more dominant. The three part compositional data of in-dust days are displayed along the lower border of the triangular diagram (low values of PM_1/PM_{10} ratio). Coarse and intermodal components are dominant and comparable. The centering and rescaling technique is used to improve the visualization of the data (Fig. 3). The two data sets and their centers are clearly separated. However, in order to prove that they are statistically distinct a statistical analysis was performed.

The bivariate angle test shows that the hypothesis of normality can be accepted for the data sets referring to in-dust days and non-dust days at a significance level greater than 2.5% and 5%, respectively. The marginal test shows that the data set referring to in-dust days, ilr_1 and ilr_2 , follow a normal distribution at a significance level greater than 10%. Moreover, the marginal test shows that the group referring to non-dust days ilr_1 and ilr_2 follow a normal distribution at a significance level greater than 2.5% and 10% respectively. The numerical results are reported in Table 1 and they have been compared with critical values reported in Stephens, (1974). Therefore, for each considered data set the hypothesis of multivariate normality cannot be rejected. The two data sets were tested for hypothesis of equality in their centre and covariance structures. The results are reported in Table 2. The test values for groups related to in-dust and non-dust days are above the critical values for each considered hypothesis, thus the equality for covariance structures, (equivalent to $\Sigma_1=\Sigma_2$) centres (equivalent to $\mu_1=\mu_2$), or both has to be rejected. The data sets related to in-dust and non-dust days have to be regarded as clearly distinct for chemical composition and tracer concentrations. These results can be interpreted as following: during in-dust days, the Saharan dust episode, together with local/regional sources, can determine the chemical composition of size-segregated PM of the suburban background site. In order to evaluate the nature of the difference between the tracer concentrations in in-dust and non-dust days the perturbation difference (Aitchison, 2005, p.73) was calculated between the perturbation centres related to in-dust and non-dust compositional data sets.

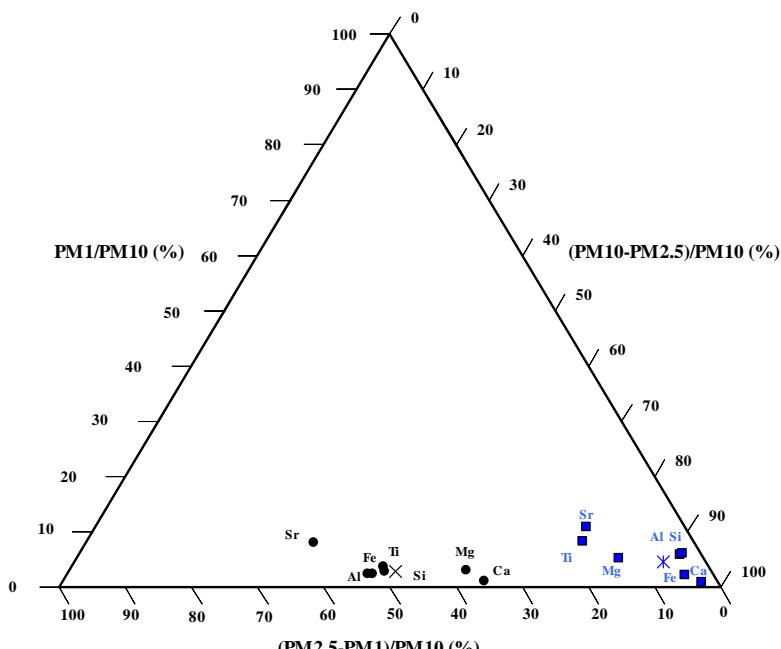


Figure 2: Distribution of data from Matassoni et al. (2011) for considered elements. The symbol \bullet in-dust days and \blacksquare non-dust days. The centre for all data set is at $g=(75.93, 19.76, 4.32)\%$, the centre \times in dust site is at $g=(49.36, 47.65, 3)\%$ and the centre $*$ non-dust site is at $g=(89.02, 6.24, 4.73)$.

Table 1. Multivariate normality tests for the two considered data sets (Matassoni et al. 2011).

non-dust days	Anderson-Darling	Cramer-von Mises	Watson
ilr ₁ marginal distribution	0.7042	0.1220	0.1219
ilr ₂ marginal distribution	0.3610	0.0542	0.0537
Bivariate angle test statistics	1.2581	0.2317	0.1718

in-dust days			
ilr ₁ marginal distribution	0.5191	0.0969	0.0966
ilr ₂ marginal distribution	0.44	0.0737	0.0736
Bivariate angle test statistics	1.071	0.1836	0.193

Table 2. Centres and covariance structure tests for the two considered data sets (Matassoni et al. 2011)

Hypothesis	Test value	χ^2 critical value ($\alpha=0.05$)	Degrees of freedom	Significance
$\mu_1 = \mu_2, \Sigma_1 = \Sigma_2$	36.870	11.07	5	0
$\mu_1 \neq \mu_2, \Sigma_1 = \Sigma_2$	9.9524	7.81	3	0.019
$\mu_1 = \mu_2, \Sigma_1 \neq \Sigma_2$	18.758	5.99	2	0.0001

The perturbation centre for in-dust days is (49.35, 47.64, 3)_(in-dust) whereas the perturbation centre for non-dust days is (89.02, 6.24, 4.73)_(non-dust). The perturbation difference is (6.29, 86.53, 7.18)_{(in-dust)-(non-dust)} suggesting that the Saharan dust event relatively increased the intermodal size fraction of the considered set of chemical tracers. Indeed, it was observed that desert dust and related mineral sources can be rich in fine particles (Dagsson-Waldhauserova and others, 2016). The non-dust and in-dust days data sets and their respective confidence regions (Pawlowsky-Glahn and others, 2015, p.224) after centering and rescaling are shown in Figure 3.

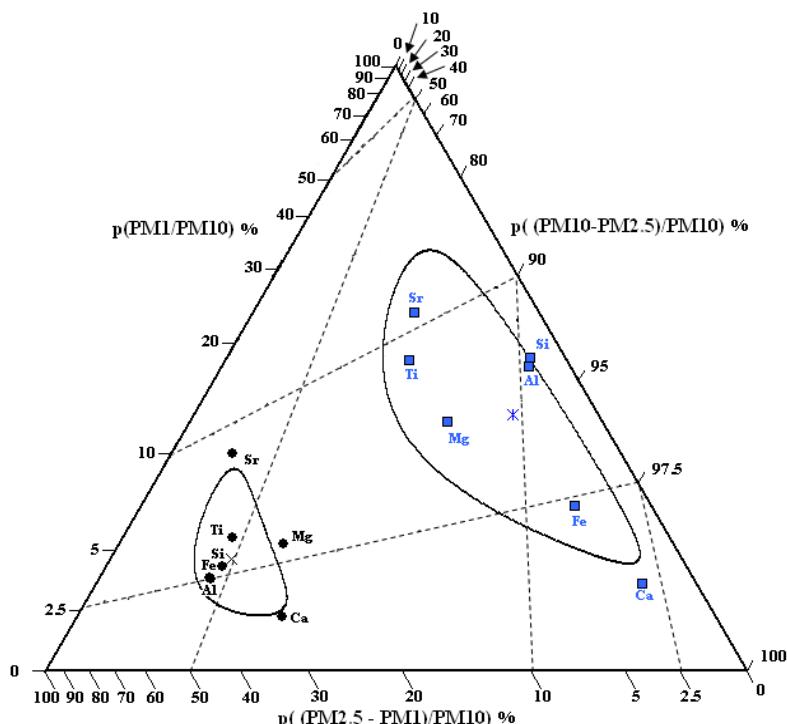


Figure 3: Distribution of data from Matassoni et al. (2011) for the considered elements after perturbation by $gm^{-3}=(4.5, 17.13, 78.41)$. The centre \times for in-dust days and the centre \times for non-dust days. The continuous lines are the confidence regions $(1-\alpha)100\% \alpha=0.05$

4 Conclusions

The statistical methods used for the analysis of compositional data allowed the validation of the differences between the investigated data sets of the related environmental site. These differences or dissimilarities can be associated with the type of mineral sources involved and possible mechanisms of addition/subtraction of materials that influences the behaviour of the environmental site.

The two data sets related to in-dust and non-dust days are clearly and statistically distinct in composition and centres. During in-dust days the contribution of the Saharan dust event alters the composition as well as the size distribution of PM, particularly the intermodal size fraction. Hence, dissimilar addition/subtraction mechanisms of mineral matters between these two sites can be observed. The compositional analysis applied to PM₁₀, PM_{2.5} and PM₁ tracer concentration simultaneous measurements is an effective technique which can be used to study environmental sites affected by several mineral sources. Moreover, the triangular diagram and centering and rescaling techniques are very important and practical tools representing compositional data of size-segregate PM mineral tracer concentration simultaneous measurements.

References

- Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 44(2), 139–177.
- Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London, 416p.
- Aitchison, J. (2005). A concise guide to compositional data analysis 2nd Compositional Data Analysis Workshop — CoDaWork'05, Universitat de Girona, Girona (2005) http://ima.udg.edu/Activitats/CoDaWork05/A_concise_guide_to_compositional_data_analysis.pdf
- Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7), 829-850.
- Buccianti, A., & Pawlowsky-Glahn, V. (2006). Statistical evaluation of compositional changes in volcanic gas chemistry: a case study. *Stochastic Environmental Research and Risk Assessment*, 21(1), 25-33.
- Buccianti, A., & Pawlowsky-Glahn, V. (2005). New perspectives on water chemistry and compositional data analysis. *Mathematical Geology*, 37(7), 703-727.
- Buccianti, A., Pawlowsky-Glahn, V., Barceló-Vidal, C., Jarauta-Bragulat, E. (1999). Visualization and modeling of natural trends in ternary diagrams: a geochemical case study. In Proceedings of IAMG (Vol. 99, pp. 139-144).
- Caggiano, R., D'EMILIO, M., Macchiatto, M., and Ragosta, M. (2001). Experimental and statistical investigations on atmospheric heavy metals concentrations in an industrial area of Southern Italy. *Il Nuovo cimento della Società italiana di fisica. C. Geophysics and space physics*, 24(3), 391-406.
- Chiari, M., Del Carmine, P., Lucarelli, F., Marcazzan, G., Nava, S., Paperetti, L., Prati, P., Valli, G., Vecchi, R., and Zucchiatti, A. (2004). Atmospheric aerosol characterisation by Ion Beam Analysis techniques: recent improvements at the Van de Graaff laboratory in Florence. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 219, 166-170.
- Chiari, M., Lucarelli, F., Mazzei, F., Nava, S., Paperetti, L., Prati, P., Valli, G., and Vecchi, R. (2005). Characterization of airborne particulate matter in an industrial district near Florence by PIXE and PESA. *X - Ray Spectrometry*, 34(4), 323-329.
- Colbeck, I., Nasir, Z. A., Ahmad, S., and Ali, Z. (2011). Exposure to PM₁₀, PM_{2.5}, PM₁ and carbon monoxide on roads in Lahore, Pakistan. *Aerosol Air Qual. Res.*, 11, 689–695.

Speranza and others

8

- Dagsson-Waldhauserova, P., Magnusdottir, A. Ó., Olafsson, H., & Arnalds, O. (2016). The Spatial Variation of Dust Particulate Matter Concentrations during Two Icelandic Dust Storms in 2015. *Atmosphere*, 7(6), 77.
- Dockery, D. W., & Pope, C. A. (1994). Acute respiratory effects of particulate air pollution. *Annual review of public health*, 15(1), 107-132.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35(3), 279–300.
- Graham, D.J., and Midgley, N.G. (2000). TECHNICAL COMMUNICATION-Graphical Representation of Particle Shape using Triangular Diagrams: An Excel Spreadsheet Method. *Earth Surface Processes and Landforms*, 25(13), 1473-1478.
- Kegler, Scott R., William E. Wilson, and Allan H. Marcus. "PM 1, Intermodal (PM 2.5-1) Mass, and the Soil Component of PM 2.5 in Phoenix, AZ, 1995-1996." *Aerosol Science & Technology* 35.5 (2001): 914-920.
- Lundgren, D. A., Hlaing, D. N., Rich, T. A., & Marple, V. A. (1996). PM10/PM2.5/PM1 data from a trichotomous sampler. *Aerosol Science and Technology*, 25(3), 353-357.
- Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3), 253-278.
- Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. In Proceedings of IAMG (Vol. 99, pp. 211-216).
- Matassoni, L., Pratesi, G., Centioli, D., Cadoni, F., Lucarelli, F., Nava, S., and Malesani, P. (2011). Saharan dust contribution to PM10, PM2.5 and PM1 in urban and suburban areas of Rome: a comparison between single-particle SEM-EDS analysis and whole-sample PIXE analysis. *Journal of Environmental Monitoring*, 13(3), 732-742.
- Makkonen, U., Hellén, H., Anttila, P., and Ferm, M. (2010). Size distribution and chemical composition of airborne particles in south-eastern Finland during different seasons and wildfire episodes in 2006. *Science of the Total Environment*, 408(3), 644-651.
- IPCC, 2013. Summary for policymakers. In: Stocker, T. F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), *Climate Change 2013: The Physical Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp
- Pawlowsky-Glahn, V., and Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.
- Pawlowsky-Glahn, V., & Buccianti, A. (2002). Visualization and modeling of sub-populations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids. *International Journal of Earth Sciences*, 91(2), 357-368.
- Pawlowsky-Glahn, V., & Egozcue, J. J. (2006). *Compositional data and their analysis: an introduction*. Geological Society, From: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, Special Publications, 264, 1-10. The Geological Society of London.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana Delgado, R. (2007). Lecture notes on compositional data analysis. <http://dugi-doc.udg.edu//handle/10256/297>
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of*

Speranza and others

compositional data. John Wiley & Sons.

9

- Prospero, J.M. (2007) *African dust: Its large-scale transport over the Atlantic ocean and its impact on the Mediterranean region.* In Regional Climate Variability and its Impacts in The Mediterranean Area (15-38). Springer Netherlands.
- Pope, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., & Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, 109(1), 71-77.
- Putaud, J.P., Van Dingenen, R., Alastuey, A., Bauer, H., Birmili, W., Cyrys, J., Flentje, H., Fuzzi, S., Gehrig, R., Hansson, H.C., Harrison, R.M., Herrmann, H., Hitzenberger, R., Hüglin, C., Jones, A.M., Kasper-Giebl, A., Kiss, G., Kousam, A., Kuhlbusch, T.A.J., Löschau, G., Maenhaut, W., Molnar, A., Moreno, T., Pekkanen, J., Perrino, C., Pitz, M., Puxbaum, H., Querol, X., Rodriguez, S., Salma, I.,... & A., Raes F. (2010). A European aerosol phenomenology–3: Physical and chemical characteristics of particulate matter from 60 rural, urban, and kerbside sites across Europe. *Atmospheric Environment*, 44(10), 1308-1320.
- Rogula-Kozłowska, W., Rogula-Kupiec, P., Mathews, B., and Klejnowski, K. (2013). Effects of road traffic on the ambient concentrations of three PM fractions and their main components in a large Upper Silesian city. *Annals of Warsaw University of Life Sciences-SGGW. Land Reclamation*, 45(2), 243-253.
- Speranza, A., Caggiano, R., Margiotta, S., Summa, V., & Trippetta, S. (2016). A clustering approach based on triangular diagram to study the seasonal variability of simultaneous measurements of PM_{10} , $PM_{2.5}$ and PM_1 mass concentration ratios. *Arabian Journal of Geosciences*, 9(2), 1-8.
- Speranza, A., Caggiano, R., Margiotta, S., & Trippetta, S. (2014). A novel approach to comparing simultaneous size-segregated particulate matter (PM) concentration ratios by means of a dedicated triangular diagram using the Agri Valley PM measurements as an example. *Natural Hazards and Earth System Sciences*, 14(10), 2727-2733.
- Spindler, G., Brüggemann, E., Gnauk, T., Grüner, A., Müller, K., and Herrmann, H. (2010). A four-year size-segregated characterization study of particles PM_{10} , $PM_{2.5}$ and PM_1 depending on air mass origin at Melpitz. *Atmospheric Environment*, 44(2), 164-173.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347), 730-737.
- Thorpe, A., & Harrison, R. M. (2008). Sources and properties of non-exhaust particulate matter from road traffic: a review. *Science of the total environment*, 400(1), 270-282.
- Viana, M., Kuhlbusch, T. A. J., Querol, X., Alastuey, A., Harrison, R. M., Hopke, P. K., Winiwarter, W., Vallius, M., Szidat S., Prévôt, A.S.H., Hueglin, C., Bloemen, H., Wählilin, P., Vecchi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitzenberger, R. (2008). Source apportionment of particulate matter in Europe: a review of methods and results. *Journal of Aerosol Science*, 39(10), 827-849.
- von Eynatten, H., Pawlowsky-Glahn, V., & Egocue, J. J. (2002). Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams. *Mathematical Geology*, 34(3), 249-257.
- WHO (World Health Organization). Regional Office for Europe, & World Health Organization. (2006). *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide.* World Health Organization.
- WHO (World Health Organization). Regional Office for Europe, & World Health Organization. (2012). *Health effects of black carbon.* World Health Organization.
- Yin, J., and Harrison, R. M. (2008). Pragmatic mass closure study for PM_{10} , $PM_{2.5}$ and PM_10 at roadside, urban background and rural sites. *Atmospheric environment*, 42(5), 980-988.

Phylofactorization - theory and challenges

Alex D. Washburne¹

¹Duke University; alex.d.washburne@gmail.com

Abstract

Data from biological communities are compositions whose parts are connected by an important sequential binary partition - the “phylogeny”, or evolutionary history of the parts. Compositional data with a natural sequential binary partition suggest the isometric log-ratio transform as a means of analyzing community ecological data. Balances in an ilr transform of the phylogeny will correspond to nodes and contrast abundances of sister clades, but traits, such as the wings of birds, arise along edges and so a natural contrast may not be between sister clades but between organisms with and without a trait. A greedy algorithm - ‘phylofactorization’ - was developed to construct an ilr transform whose balances correspond to edges along which traits arose, thereby contrasting birds to non-birds as opposed to contrasting birds to crocodiles.

In this paper, the general theory of phylofactorization is presented as a graph partitioning algorithm. A special case - regression phylofactorization - chooses ilr coordinates based on sequential maximization of objective functions from regression. The connections between regression phylofactorization and other methods is discussed, including matrix factorization, hierarchical regression, factor analysis and latent variable models. Open challenges in the statistical analysis of phylofactorization are presented, including criteria for choosing the number of factors and approximating null-distributions of commonly used test-statistics and objective functions. As a graph-partitioning algorithm, cross-validation of phylofactorization across datasets requires graph-topological considerations, such as how to deal with novel nodes and edges and whether or not to control for partition order. These challenges carry major implications for the biological sciences and are a promising area of future work.

Key words: compositional data, community ecology, isometric log-ratio, phylogeny, phylofactorization, graph-partitioning, greedy algorithm, regression, cross-validation

1 Introduction

It's easy to take for granted the elegance of the particular spherical coordinates used to define locations on the surface of the Earth. First, the approximately constant radius of the Earth provides a natural reduction in the dimension of GPS data from 3 to 2 dimensions when changes in elevation are negligible. When changes in elevation are non-negligible, changes in radius correspond to changes in air temperature and atmospheric pressure independent of changes in the other two dimensions. Second, the choice of latitude to correspond to angular deviation from the equator in the direction of the axis about which the Earth spins may seem obvious, but on an abstract sphere one could choose any axis to define latitude. Latitude relating to the spin of the Earth about its axis - and closely corresponding to the revolution of the Earth about the sun - is a "natural" choice of a variable that yields latitudinal associations with other important measurements such as climate (tropical, subtropical, temperate and arctic), positions of stars in the sky, and more. Longitude does not have an obvious reference, so it is set by convention to be zero at the Prime Meridian, which conveniently passes through the Royal Observatory in Greenwich, England (the French detested this convention, and used their own Paris Meridian until the early 20th century). Some coordinates are natural choices, whereas others are left to convention. If we didn't know the Earth was a sphere and spinning about an axis, one would be pleased to discover two coordinates which so closely correspond to changes in important environmental meta-data.

While distances often motivate coordinates for algebraic ease, the choice of coordinates is still separate from the choice of distances. One could define the same distance - as the crow flies or "as the gopher burrows" - on the same sphere independently of the choice of coordinates. Thankfully, the distance between Sienna and Los Angeles remains the same, regardless what bizarre coordinates mathematicians are working with. The point being: choice of coordinates is separate from the choice of distance, and changing coordinates is motivated by dimensionality reduction, natural directions one might travel or along which meta-data change, and remaining coordinates can be filled in ad hoc, by convention, or for convenience.

Compositional data bound to the surface of the simplex yield many choices of distance, the most commonly used of which is the Aitchison distance [Aitchison, 1986]. Fixing the definition of distance between points, there are many choices of coordinates motivating more elegant, relevant, and easy-to-analyze quantities for data on the simplex. The isometric log-ratio (ilr) transform [Egozcue et al., 2003, Egozcue and Pawlowsky-Glahn, 2005] is a change of basis to variables whose Euclidean distances are equal to Aitchison distances of the original, compositional variables, and whose values reflect a contrast or the differences in relative abundance of two groups. The ilr transform can be defined through a sequential binary partition defining which groups of parts to contrast for each coordinate but, in choosing which isometric log-ratio transform to use for analysis and cross-validation across datasets, one encounters similar considerations as those encountered when choosing spherical coordinates for the Earth [Pawlowsky-Glahn and Buccianti, 2011].

This paper is about a method - phylofactorization [Washburne et al., 2017] - whose aim is to construct an isometric log-ratio transform for biological data which yields coordinates that are consistent with the compositional nature of the data, meaningful (biologically interpretable), and convenient (coordinates along which there are predictable changes). Much like spherical coordinates for GPS data require geography and astronomy to motivate, phylofactorization requires some biological background to motivate the coordinates and precisely why we are choosing them.

First, one needs justification for the analysis of communities as compositions or, generally, as objects prone to geometric changes and thus justifiably analyzed by isometric log-ratios. Then, one needs to know about the tree of life (the phylogeny) as a graph with no cycles, a sequential binary partition connecting species in a community. Phylofactorization often uses the ilr transform as a contrast of two groups separated by edges in the phylogeny. Variables corresponding to edges in the phylogeny yield biologically meaningful coordinates corresponding to putative functional ecological traits, axes about which biologists might expect there to be predictable changes with

environmental meta-data, which may allow further development of community ecological theory [McGill et al., 2006].

This paper discusses the general theory of phylofactorization as a graph-partitioning algorithm iteratively cutting the phylogeny at edges which maximize an objective function. Edges in the phylogeny separate the community into two disjoint groups of species, and thus, when analyzing community composition, one can use ilr coordinates as quantities of interest corresponding to edges. Phylofactorization done to maximize objective functions from regression on ilr coordinates is called “regression-phylofactorization”; the relationship between regression-phylofactorization and matrix factorization, hierarchical regression, factor analysis, and latent variable models is discussed. Future research directions are discussed, including criteria for choosing the number of factors, the null distribution of test-statistics from phylofactorization to allow hypothesis-testing, and graph-topological considerations when cross-validating phylofactorization across datasets with non-identical, or even disjoint, sets of species. Future research along these directions carries major implications for methods to diagnose disease [Kostic et al., 2014], modulate microbial communities [Rajpal and Brown, 2013], and make inferences about the habitat associations of unclassified species in the tree of life [Letunic and Bork, 2007].

2 Communities as Compositions

Communities are assemblages of organisms in a study area or a sampling design. The definition of a community is often arbitrary and based on the organisms we happen to observe or we think are important. Due to the challenge of sampling communities, communities are rarely the entire set of species in a region of space - rather, communities will be defined as the assemblage of trees in a forest, grasses in a grassland, mammals in Yellowstone National Park, or birds in a patch of rainforest. For example, one could analyze the community of non-human animals around Sienna. If a researcher counted the non-human animals around Sienna, they would observe counts, integer-valued random variables of the numbers of birds, cows, sheep, cats, dogs, and other non-human animals in the community.

Community ecology studies how communities assemble, function, and respond in light of biotic and abiotic conditions. A common question is: how do communities differ along environmental gradients or between sample sites? One of the most famous and pioneering examples of this line of questioning was Alexander von Humbolt’s observation that how plant communities change with increasing latitude is very similar to how plant communities change with increasing altitude, implicating climate (rainfall, length of growing season, etc.) as an important abiotic factor in community structure and function. For an example closer to CoDa 2017, how would the animal community in Sienna differ from the animal community in the Mediterranean Sea, and how much of that is due to latitude (likely very little) versus the fact that one habitat is terrestrial and one is marine? We obtain count data of organisms, and wish to make inferences on how these counts change and which environmental meta-data are most important.

Compositional data analysis has a long marriage with geosciences, but it is somewhat unfamiliar to community ecology, and so it’s worthwhile to take a few paragraphs to motivate compositional data analysis in community ecology. The first step in this paper is to motivate communities as compositions or, more generally, as comprised of quantities which exhibit geometric changes most appropriately analyzed in terms of the log-ratios in compositional data analysis. Once we believe communities are best analyzed in terms of log-ratios, we can move on confidently to discuss the ilr transform of a community as a means of analyzing problems like the von Humboldt’s, and then we can discuss how the ilr transform allows us to choose very meaningful coordinates in light of the evolutionary tree.

First, consider the nature of the data one obtains when sampling communities. Suppose we have a small sample size where we observe 10 doves, 30 starlings, and 10 cats in one day, and, on a

rainy day, we observe 2 doves, 6 starlings, and 2 cats. Did the population of doves decrease by 8, or did our sampling effort/ability change? Many animals are not as active on rainy days, so one would clearly attribute this to a change in effort or detectability. To control for changing effort or detectability, a more robust analysis would limit itself to inferences on the relative abundances of doves, starlings, and cats, possibly incorporating sampling effort/ability as weights for averages, regression, etc.

Are communities only compositions when we obtain a small sample size? What if we sampled every non-human animal in Sienna, and the total community size changes (e.g it grows as people acquire fewer cats and fewer birds are killed, or shrinks in a drought)? In small islands, fragments of habitat or enclosures, or extensively sampled habitats with easily detectable organisms, one knows absolute abundances - how should we analyze changes in absolute abundances?

Biologists have long struggled to define first principles, but I will propose two such principles central to the definition of “life” - reproduction and death. Every organism can die, and just about every organism can reproduce. If there was no reproduction and every organism had a constant probability of dying per unit time, the expected population dynamics would be an exponential decrease in population size, or linear decreases in log-abundances over time. If there is reproduction, and the propensity for birth exceeds that of death, the expected population dynamics would be an exponential increase in population size, or linear increases in log-abundances. More complicated state-dependent propensities can yield complicated dynamical systems, yet, near an equilibrium where birth rates equal to death rates, all such dynamical systems exhibit exponential approach to or explosion away from equilibria, most conveniently analyzed as changes in log-abundances. If changes in environmental conditions lead to changes in fitness, i.e. per-capita propensities for reproduction and death, we would see short-term, geometric changes in population size. Motivated by these first principles of per-capita birth and death yielding a tendency for exponential growth and decay, one is not without justification to default to analyzing absolute abundances on a log-scale, with changes modeled as log-ratios. There is empirical support for analyzing population dynamics on a log-scale: Kalyuzhny et al. [2014] looked at bird populations across North America from 1966 to 2014 and found that the counts of hundreds of sub-populations of birds exhibit temporal fluctuations more consistent with a geometric Brownian motion than by an arithmetic Brownian motion, more naturally and coherently analyzed in terms of changing log-ratios of abundances. Populations move more like a stock price than a particle under a microscope.

So, if we count only a few birds, we are best confined to using log-ratios, with appropriate treatment of zeros, to infer changes in relative abundances. If we count all the birds, we are at least somewhat justified in using log-ratios to infer changes in absolute abundance. Changes in equilibrium population sizes and community compositions may be neither arithmetic nor geometric, as changes in rainfall can lead to complicated, neither linear nor exponential, increases in the number of trees in a region, so there is no law that log-ratios are always appropriate - they should be justified in each case. However, most datasets do not sample entire communities and so, often, changes in community composition are justifiably analyzed as changes in log-ratios.

Most phylofactorizations of community ecological data have focused on a recent and highly relevant class of data - “microbiome” data, collected by sequencing microbes in a region such as soils, tongues, or the human gut. These data are obtained by amplicon sequencing - sequencing “barcode” genes and counting the number of different types of “barcodes” obtained - where absolute counts depend on the amount of reagents we put into a machine (the sequencing depth) and not absolute abundances of microbes in the community. These sequence-count data are as compositional as the bird counts in Sienna - the absolute number of counts relates to effort, and only inferences on relative abundances can be made. While these data are compositional, best analyzed with log-ratios and Aitchison distances, the choice of coordinates remains: changes are relative, but which groups are changing, and relative to which other groups are they changing? Microbiome datasets often come with the evolutionary tree articulating the common ancestry and origins of species in the dataset, a natural scaffolding for choosing groups and, therefore, choosing coordinates.

3 The Evolutionary Tree

All cellular life arose from a common ancestor, and so all community ecological datasets contain parts connected by the evolutionary tree also known as the phylogeny. The phylogeny simplifies biological data into a nested hierarchy of lineages - doves are birds, birds are vertebrates, vertebrates are animals. Going the other direction, animals can be split into vertebrates and invertebrates, vertebrates can be split into amniotes and anamniotes, and so on to the doves, starlings and cats found in Sienna.

The evolutionary tree contains a natural means of dimensionality reduction in biological data. Consider describing the 66,000 vertebrate species based on whether they live on land or in water. One could capture all of the land/water associations with 66,000 variables, one for each species. Alternatively, one could note that all vertebrate lineages before tetrapods (e.g. fish, sharks, rays) live in the water, whereas tetrapods tend to live on land. One variable indicating whether a species is a tetrapod or not, a variable corresponding to one edge in the phylogeny, captures most of the 66,000 species' land/water associations by accurately guessing the habitat of over 30,000 fish, sharks and rays and correctly guessing the habitats of most reptiles, birds and mammals. A handful other, similarly constructed variables will finish the job: whales and dolphins live in the water relative to other tetrapods (a second variable splitting whales & dolphins from all other tetrapods), seals/sealions/walruses live in the water (a third variable splitting seals/sea-lions/walruses from all tetrapods that are not whales/dolphins), and so on (partitioning amphibians that live on land or in the water will likely require some more variables). A handful of well-chosen coordinates corresponding to edges on the tree of life can explain most of the variance in land/water associations in a 66,000-dimensional dataset.

The utility of the phylogeny for dimensionality reduction stems from its correspondence to traits, the products of evolution by natural selection, features which determine where an organism lives, what it eats and how it responds to changing environmental conditions. The edge along which tetrapods arose is the edge along which limbs and lungs arose, allowing organisms to walk on land and breath air. The edge separating whales & dolphins from their most recent common ancestor on land is the edge along which flippers and blowholes arose, allowing them to swim and conveniently gasp for air at the surface of the water. While we're more familiar with the traits of animal lineages and their ecological functions, we know very little about the traits on the microbial tree of life. For microbes, we have a phylogeny, often constructed from the "barcodes" mentioned above, but we know almost nothing about which traits arose on which lineages and what functions those unknown traits may have. The study of microbial communities can be improved by considering traits, either explicitly as measured or implicitly as latent variables on the tree of life, as determinants of disease, habitat association, or response to perturbation [Martiny et al., 2015].

We'll be dealing with the phylogeny in this paper as a mathematical structure used to make inferences about locations of putative traits driving changes in community composition. As a mathematical structure, the phylogeny is a connected graph containing no cycles. For a given gene, the true structure is a sequential binary partition - a strictly bifurcating graph where all internal nodes have three neighbors and all tips have one. However, phylogenies are never known - they are estimated - and often we don't know how to resolve the relationships between ancient nodes as no one was there to see what happened. We use contemporary data to guess what might have happened. Unresolved nodes are "polytomies", and a tree with polytomies is not a sequential binary partition; an unresolved polytomy will have one parent and more than two descendants. All edges, however, connect only two nodes. The edges in the phylogeny have lengths, referred to as "branch lengths", which approximately correspond to the time between speciation events. Edges are the locations along which traits arise.

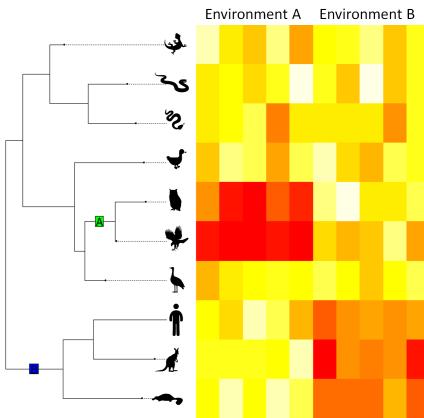


Figure 1: Phylogenetically-structured compositional data. A simulated dataset of relative abundances of 10 species across 5 samples in each of two environments. Blocks of data correspond to the phylogenetic structure one would observe if monophyletic clades have common patterns of abundance across environments. In this example, raptors (owls & hawks) are hyper-abundant in Environment A, whereas mammals (humans, kangaroos and platypus) are hyper-abundant in Environment B. Common habitat associations of organisms with common ancestry may be caused by functional ecological traits.

4 Phylofactorization

If we observe changes in the composition of a community of species connected by an evolutionary tree, which coordinates best capture the changes and relate to putative traits?

Consider a simulated dataset of community compositions, $\mathbf{x}_j \in \Delta^D$ for $D = 10$, across 10 samples $j = 1, \dots, 10$, illustrated in Figure 1. The dataset consists of a set of vertebrates one might find in Australia sampled in 5 different sites in each of two different environments. In these data, we consider traits driving habitat associations. Traits arise along edges and, in the example in Figure 1, two traits drive differential abundances. One trait, “A”, arose along the common ancestor shared by hawks and owls and leads to an increased abundance in environment B. Another trait, ‘B’, is shared by mammals and leads to an increased abundance in environment B. A default analysis of community compositions would motivate using a clr or an ilr transform.

Analysis of an arbitrary ilr transform may indicate differences between the two environments, but arbitrary ilr coordinates do not map to traits. We can do better. The phylogeny is a sequential binary partition, but analysis of an ilr transform constructed from the rooted phylogeny will obtain variables corresponding to changes in sister clades - such as changes in hawks and owls relative to changes in cranes. The inelegance of the ilr transform of a rooted phylogeny comes from the variables corresponding to nodes, not edges. The changes we aim to detect are not changes of sister clades relative to one-another, but changes in one clade with a trait relative to the rest of the organisms without a trait, possibly controlling for other traits we have already identified as important. Such changes correspond to edges, giving us coordinates that can be interpreted as differential abundances of species with and without various traits. Constructing ilr coordinates corresponding to edges yields coordinates corresponding to latent variables - traits. However, the edges don’t define a sequential binary partition - we need to choose one.

4.1 General algorithm for phylofactorization: a graph-partitioning algorithm without a balance constraint

Phylofactorization [Washburne et al., 2017] is a greedy algorithm for constructing a sequential binary partition corresponding to edges in the phylogeny. Each edge, e , in the phylogeny separates the community into two disjoint groups of species, R_e and S_e , containing r_e and s_e species, respectively. Phylofactorization requires an objective function, $\omega(\mathbf{X}, R, S)$, of the dataset, $\mathbf{X} = (x_{i,j})$ and the disjoint index sets of the two groups, R and S . The most general form of phylofactorization defines a graph partitioning algorithm [Buluç et al., 2016].

The general algorithm for phylofactorization follows:

1. Compute objective function corresponding to each edge, e , with partition p_e separating two groups, R_{p_e} and S_{p_e} : $\omega_{p_e} = \omega(\mathbf{X}, R_{p_e}, S_{p_e})$
2. Identify $e^* = \max_e \omega_{p_e}$
3. Cut the tree at e^* , creating two disconnected graphs
4. Repeat 1-4 until stopping criterion is reached.

The general algorithm is illustrated in Figure 2.

For example, one could define the objective function as a function of an ilr balance for each edge. Recall the ilr transform of a given sample corresponding to a generic partition, p , of two disjoint index sets, R and S , containing r and s species, respectively, can be written as

$$y_p = \sqrt{\frac{rs}{r+s}} \log \left(\frac{g(\mathbf{x}_R)}{g(\mathbf{x}_S)} \right) \quad (1)$$

where $g(\cdot)$ is the geometric mean, $\mathbf{x}_R = (x_{i_1}, \dots, x_{i_r})$ for all elements $i_i \in R$ and $\mathbf{x}_S = (x_{l_1}, \dots, x_{l_s})$ for all elements $l_i \in S$. The balance, y_p , can also be obtained by projecting log-relative abundances, $\log(\mathbf{x})$, onto the balancing element, \mathbf{v}_p , whose i th element is

$$v_{p,i} = \begin{cases} \sqrt{\frac{s}{r(r+s)}} & i \in R \\ -\sqrt{\frac{r}{s(r+s)}} & i \in S \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

For each sample, j , one obtains $y_{p,j}$ and can define the objective function as the variance of an ilr transform:

$$\omega_V = \sum_{j=1}^m \frac{(y_{p,j} - \bar{y}_p)^2}{m-1}. \quad (3)$$

Maximizing ω_V is similar to a principal components analysis - the edge, e^* , with corresponding partition, p^* , which maximizes ω_V defines an ilr balancing element, \mathbf{v}_{p^*} , onto which projection of log-relative abundances maximizes variance. Phylofactorization by ω_V may yield edges separating the most different species, corresponding to the most important traits in the dataset which may or may not be predictable with existing meta-data. One could devise other objective functions with ilr balances, such as $\omega_{L_1} = |\bar{y}_p|$. The balancing elements in the ilr transform are convenient tools for phylofactorization, allowing edges to be boiled down into variables interpretable as contrasts of groups split by the edge. The variance of the contrasts from projection onto balancing elements does not change with r or s , allowing a more fair competition of the edges in the phylogeny to be the site of a partition.

As a graph partitioning algorithm, phylofactorization has dual interpretations in terms of the edges cut and the sub-graphs that remain un-cut. After k iterations of phylofactorization, the

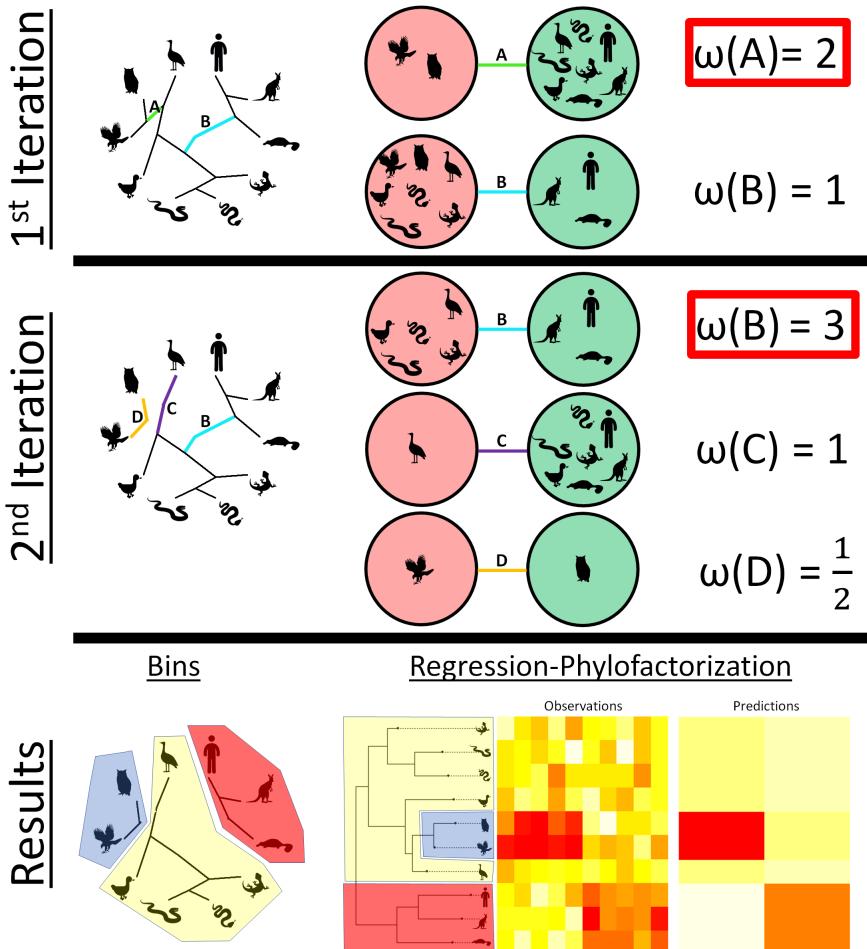


Figure 2: Phylofactorization. Phylofactorization is a graph partitioning algorithm which iteratively cuts edges in the phylogenetic tree based on objective functions, ω . In the first iteration (**top row**), all edges are considered, but we depict the two edges with traits affecting abundance patterns in figure 1. Each edge separates the group of D species into two groups. Edge A separates the raptors (hawk and owl) from all other species, and edge B separates the mammals (platypus, kangaroo, and human) from all other species. An objective function, ω , often a measure of contrast of the two groups is calculated for each edge, and the edge which maximizes the objective function is selected. The phylogeny is cut/partitioned at that edge - edge A, in the illustrated example - and the algorithm continues. In the second iteration (**middle row**), all edges are considered but each edge separates only those species in its sub-graph. Since edge A was cut in the first iteration, in the second iteration edge B contrasts mammals from all non-raptor species, not mammals from all other species as it did in the first iteration. The edge which maximizes the objective function, in this case edge B, is chosen and the tree is cut. (**bottom row**) After k iterations of phylofactorization, one has identified k edges or chain of edges (see chain labeled edge D) of interest which partition the phylogeny into $k + 1$ bins of species. In the special case of regression-phylofactorization, edges yield contrasts between groups via projection of log-relative abundances or absolute abundances onto a balancing element, v_p , corresponding to the partition, p , defined by the edge, and the objective functions are statistics from regression on the balance. Regression phylofactorization outputs predictions of k balances yielding low-rank predictions of abundance patterns, recreating the major blocks of variation illustrated here.

species will be split into $k + 1$ groups or “bins”, referred to as “binned phylogenetic units”. Where phylofactorization correctly identifies edges corresponding to functional ecological traits, the bins correspond to species with similar functional ecological traits. For instance, splitting vertebrates along the three edges: “tetrapod”, “Cetacean” (whale/dolphin), and “Pinniped” (seals, sea lions & walruses), will result in four groups: (1) non-tetrapod vertebrates, (2) Cetaceans, (3) Pinnipeds, (4) tetrapods which are neither Cetaceans nor Pinnipeds. Group (1) has gills and fins and lives in the water, group (2) has flippers and blowholes and lives in the water, group (3) has flippers and eats fish and lives in the water, and group (4) has legs, lungs and lives on land. Phylofactorization can be seen as having two complimentary goals: accurately identifying edges corresponding to functional ecological traits, and accurately binning species into groups with similar traits. Novel objective functions or phylofactorization algorithms can be evaluated by their ability to perform each of these two goals.

In the general theory of phylofactorization, there are many unanswered questions for future research. The full connections to graph partitioning and implications of such connections on the algorithm have yet to be expounded, the utility of different objective functions for inferring different traits and ecological processes has been largely unexplored, and the performance of Hastie-Tibshirani [Friedman et al., 2001] stochastic algorithms may yield more accurate inferences but has yet to be investigated.

4.2 Regression phylofactorization

Often, researchers are interested how communities change across sample sites or environmental gradients such as latitude, pH in soils, or rainfall. For such problems, each community composition, \mathbf{z}_j , comes with m associated environmental meta-data measurements, $\mathbf{z}_j = (z_{1,j}, \dots, z_{m,j})$. Phylofactorization can be utilized by defining objective functions from regression on an ilr balance, such as the explained variance, F -statistics, or t -statistics from a focal coefficient in multiple regression. For compositional data, regression-phylofactorization performs regression on the ilr balances, y_{p_e} corresponding to the partition, p_e , for each edge, e . The regression takes the form

$$y_{p_e} \sim \beta_0 + \beta_1 z_1 + \dots + \beta_m z_m. \quad (4)$$

and one selects the edge, e^* , with partition p_{e^*} , which maximizes the objective function of regression. Regression-phylofactorization is a form of hierarchical regression in which each iteration considers candidate variables, y_{p_e} , to maximize an objective function from regression and successive iterations consider variables with non-overlapping partitions, thereby controlling for previous inferences. However, instead of a strict order for hierarchical regression as used in ordinary least-squares, phylofactorization hosts a competition at each iteration, k , among a set of candidate variables.

The objective function originally considered, and commonly used in exploratory regression-phylofactorization, is the explained sum of squares from regression,

$$\omega_{\text{ExVar}}(p) = \frac{1}{n-1} \sum_{j=1}^n (y_{p,j} - \bar{y}_p)^2 - (\hat{y}_{p,j} - \bar{y}_p)^2 \quad (5)$$

where \bar{y}_p is the sample mean of ilr balance corresponding to partition, p , and $\hat{y}_{p,j}$ is the fitted value of regression on the ilr balance, y_p . The use of explained sum of squares as an objective function was motivated by the fixed total variance in a compositional dataset, irrespective of the choice of ilr transform, which implies that each iteration of phylofactorization has candidate edges, e , competing for a fixed amount of remaining variance to explain. Explicitly,

$$\text{totvar} = \sum_i \text{Var} [\text{clr}_i(\mathbf{X})] = \sum_i \text{Var} [\text{ilr}_i(\mathbf{X})] \quad (6)$$

for any choice of ilr transform. After k iterations, the previously chosen ilr balances with partitions $\{p_i\}_{i=1}^k$ will explain a sub-total variance,

$$\hat{\sigma}_k^2 = \sum_{i=1}^k \omega_{\text{ExVar}}(p_i) \quad (7)$$

leaving candidate ilr balances competing for $\text{totvar} - \hat{\sigma}_k^2$ variance in the dataset. Maximizing the explained variance from regression at each iteration of phylofactorization produces a set of ilr balancing elements with fitted values that provide a low-rank approximation of a dataset, $\text{clr}(\mathbf{X})$, discussed further in the next section: Phylofactorization and factor analysis.

Regression-phylofactorization can also be used in sample-site classification, a crucial problem for disease detection, by reversing the regression:

$$z \sim \beta_0 + \beta_1 y_{p_e} \quad (8)$$

and maximizing the F -statistic to identify phylogenetic bioindicators of disease. Ideally, one would like to use multiple regression to produce multiple bioindicators of disease, but the power of such approach is limited by the sequential nature of phylofactorization. The computational time needed to perform multiple regression of the form

$$z \sim \beta_0 + \beta_1 y_{p_1} + \beta_2 y_{p_2} \dots + \beta_k y_{p_k} \quad (9)$$

grows exponentially with k , and many problems of partitioning a graph into k bins are NP hard. However, solutions or algorithms which perform reasonably well at finding the set of partitions which maximize the F -statistic in the regression of Equation (9) may improve the classification of disease, such as inflammatory bowel disease [Kostic et al., 2014], and identification of microbial drivers of environmental conditions.

4.3 Phylofactorization and factor analysis

Recall that the ilr transform in Equation (1) can be re-written as a projection of log-relative abundances, $\log(\mathbf{x})$, onto a basis vector called the balancing element, \mathbf{v} , defined in Equation (2). In K iterations of regression phylofactorization, one has a set of balancing elements, $\{\mathbf{v}_k\}_{k=1}^K$, giving a set of K balances, $\{\mathbf{y}_k\}_{k=1}^K$, and their fitted values from regression, $\{\hat{\mathbf{y}}_k\}_{k=1}^K$. Let V be a $D \times K$ matrix whose columns are balancing elements, \mathbf{v}_k , Y the $K \times n$ matrix whose rows are ilr balances, \mathbf{y}_k corresponding to the respective balancing elements in the columns of V , and whose columns are samples. Let \hat{Y} the same as Y but containing fitted balances, $\hat{\mathbf{y}}_k$. Regression-phylofactorization can serve as a rank K approximation of the $\text{clr}(X)$ transformed data matrix

$$\text{clr}(\mathbf{X}) = VY + \epsilon \quad (10)$$

$$\text{clr}(\mathbf{X}) = V\hat{Y} + \hat{\epsilon} \quad (11)$$

where $\epsilon \in \mathbb{R}^{D \times n}$ is the error of the approximation by ilr balances and $\hat{\epsilon} \in \mathbb{R}^{D \times n}$ the error of the approximation by fitted ilr balances. The matrix illustrated in the end of Figure 2 is an example of Equation (11), a rank 2 approximation of $\text{clr}(X)$ using the fitted ilr balances from regression. In this sense, phylofactorization is a change of basis and a matrix factorization algorithm.

In the original paper, phylofactorization was also hypothesized to be a form of constrained factor analysis of the clr -transformed data, as the use of ilr balancing elements allows the construction of a set of orthonormal axes that allow low-rank predictions of the data. Factor analysis obtains low-rank approximations of a data matrix, $Z \in \mathbb{R}^{D \times n}$ through the product of an orthogonal matrix of “loadings”, $L \in \mathbb{R}^{D \times K}$ and a matrix of “factors” $G \in \mathbb{R}^{K \times n}$, plus an error matrix $\epsilon_z \in \mathbb{R}^{D \times n}$,

$$Z = LG + \epsilon_z. \quad (12)$$

such that (1) G and ψ are independent, (2) $\mathbb{E}[G] = 0$, and (3) $\text{Cov}[G] = I$. Phylofactorization was hypothesized to be a form of factor analysis due to its apparent similarity in constructing an orthogonal matrix, V , a matrix Y corresponding to a latent variable (differential abundance of traits) and constrained to axes which are balancing elements of edges in the phylogeny.

However, the claim that phylofactorization is a form of factor analysis is not true for the default regression-phylofactorization maximizing ω_{ExVar} . While phylofactorization clearly obtains a low-rank approximation nominally similar to factor analysis, it is clear that $\mathbb{E}[Y]$ is not necessarily zero - there could be imbalances in the data such that one group, R is, on average, more abundant than its complement, S , causing $g(\mathbf{x}_R) > g(\mathbf{x}_S)$ and $\mathbb{E}[y_p] > 0$ for a given row of Y . Furthermore, while it has been conjectured that the factors are sequentially independent, under the assumption that an ilr balance at one partition contains no information about the ilr balance at subsequent partitions (the converse is not necessarily true: ilr coordinates along the root path of a sequential binary partition may be correlated under increases in single clades, as highlighted in the original phylofactorization paper), the conjectured independence of sequential factors (1) is a conjecture and (2) does not imply that $\text{Cov}[Y] = I$ for a given phylofactorization. Whether or not sequential factors are independent depends on the underlying statistical model of how abundances change and whether or not a factor was correctly identified. There may be other algorithms for regression-phylofactorization which are a form of factor analysis, such as phylofactorization of standardized datasets (thereby assisting $\mathbb{E}[y_p] = 0$) with objective functions which minimize off-diagonal elements of $\text{Cov}[Y]$ or yield $\text{Cov}[Y] = I$ asymptotically..

Despite phylofactorization not being factor analysis *sensu stricto*, there are conceptual similarities between phylofactorization, factor analysis, principal components analysis, and other methods. Such connections may carry implications or suggested solutions for many of the challenges highlighted in the next section. The orthonormal basis $\{\mathbf{v}_k\}_{k=1}^K$ chosen by phylofactorization is a low-rank set of orthogonal axes along which meaningful change occurs in a compositional dataset. Where the objective function is to maximize the variance of the ilr balance, phylofactorization becomes like PCA in finding orthonormal axes sequentially maximizing the variance of data projected onto the axis. Since the ilr transforms are all rotations of one-another, there may be connections with rotation methods in factor analysis to allow for convenient interpretations of the method. For instance, the bins described in the results of Figure 2 can be constructed either by factoring as illustrated (edge A and then edge B), or by a re-ordering of the factorization of those same edges (edge B and then edge A) - the ilr bases corresponding to those two examples are rotations of one-another. Finally, the “factors” in factor analysis are latent variables, and regression phylofactorization is inferring latent variables - traits and their differential abundance - that underlie covariances among species. In this sense phylofactorization is also a latent variable model, and could be implemented with more complicated functional responses of traits than those captured by regression on ilr balances. Phylofactorization also constructs a decision tree for the classification of species [Rokach and Maimon, 2014] - given a new species, we ask a set of questions to make estimates about its habitat associations (Is it a tetrapod? If yes, is it a whale/dolphin? If no, is it a seal/sea-lion/walrus?).

The place of phylofactorization in the mathematical, statistical and computational literature, and its connections with other methods is still being resolved and may likely vary depending on the choice of objective function and particularities of the analysis. Future work integrating phylofactorization into more general mathematical and statistical literature may provide theoretical connections which suggest more efficient algorithms, more powerful tests, more apposite objective functions, and more interpretable results.

5 Challenges of Phylofactorization

The novelty and biological utility of phylofactorization leaves open many avenues of future research. As articulated above, more needs to be done to understand the generality, utility, and limitations of

phylofactorization. In addition to articulating the place of phylofactorization in graph partitioning algorithms, understanding the pros, cons and interpretations of different objective functions for regression phylofactorization, and understanding the relationship between phylofactorization and other statistical methods, there are several more focused challenges to the maturation of phylofactorization which, if addressed, can dramatically improve its utility in the biological sciences. Two of those challenges are illustrated here: criteria for choosing the number of factors and cross-validation.

Much like factor analysis, there is the need for criteria to choose the number of factors, a problem intimately tied with the null distributions of test statistics under phylofactorization. At which iteration, K , should we stop factorization to control our family-wise error rate of inferences on edges or control error rates in some other sense? Calculation of the Marchenko-Pastur distribution [Marchenko and Pastur, 1967] for the asymptotic distributions of singular values aided the development of PCA into an inferential tool - what is the null distribution of the variance of the dominant edge's ilr balance from phylofactorization with ω_V ? What is the null distribution of the F-statistics from regression-phylofactorization? Answering these questions can move phylofactorization from an exploratory to an inferential tool.

Phylofactorization makes predictions assignable to a universal tree of life, which allows comparison of datasets with unrelated species for cross-validation of phylofactorization. If we discover the ratio of birds to mammals shifts dramatically across two Australian habitats - tree tops and the desert - we may be interested in cross-validating such findings for birds and mammals on other continents, even if the sets of species across continents are completely disjoint. Cross-validation introduces graph-topological and statistical challenges including those common to cross-validation and those common to ilr analysis including the nested dependence of variables. Answering these questions can allow widespread use of phylofactorization for comparison across datasets, annotations on the tree of life, and development of disease detection or, more broadly, community classification despite novel or non-overlapping species.

Phylofactorization can be implemented in the R package `phylofactor`. To encourage future progress on these challenges, some key functions have been constructed to assist future research along these lines. The R package is available on GitHub at <https://github.com/reptalex/phylofactor>, and novel methods for choosing numbers of factors, calculating quantiles for null distributions of test statistics, and cross-validation can be added to the R package to improve the robustness of the tool.

5.1 Criterion for choosing the number of factors

How many factors should we choose? How many traits can we confidently say are contributing to observed patterns of community composition? We can perform phylofactorization for $D - 1$ iterations to construct a full basis, but, for big datasets such as the Earth Microbiome Project [Gilbert et al., 2010] containing hundreds of thousands of species and hundreds of thousands of samples, phylofactorization can be prohibitively expensive, even with appropriate parallelization. Even where computation time is not a limiting reagent, downstream analysis and interpretation of phylofactorization can be time-intensive. Choosing the number of factors requires constructing an algorithm that stops phylofactorization once there appears to be no more signal in the data, thereby reducing computational time and saving researchers from having to hypothesize putative traits along insignificant edges in the phylogeny likely to be false-positives.

The original paper proposed a stopping function for regression-phylofactorization based on a Kolmogorov-Smirnov test of the uniformity of the distribution of P -values resulting from F -tests, referred to as the KS stopping criterion. The justification for the KS stopping criterion was that phylofactorization, at each iteration, is often obtaining test statistics for each edge and, if the null hypothesis is true, the P -values from multiple hypothesis tests should be uniform. The KS stopping criterion was demonstrated to allow a conservative stopping criterion for simulations of up to

10 clades with randomly assigned geometric changes in one of two environments. The original test did not specify an alternative hypothesis for the KS test, but this author has found the alternative hypothesis that empirical distribution of P-values is greater than the uniform null distribution is more robust; two-sided KS tests can lead to missed stops as the P-value distribution shifts from right-skewed to left-skewed in a single factor, and downstream factors cause the P-value distribution to become even more left-skewed; missed-stops can lead to a full phylofactorization, which can be prohibitively costly for large datasets.

The KS stopping criterion performs well, but other algorithms could outperform it and simultaneously obtain more accurate statements of certainty about the number of factors. There are many criteria for choosing the number of factors in factor analysis which may motivate robust stopping criteria for phylofactorization. Here, we'll build on the previous literature by comparing the KS stopping criterion to a popular alternative in factor analysis. Horn's parallel analysis [Horn, 1965] chooses the number of factors based on simulation of the mean, null distribution of eigenvalues of the correlation matrix for standard Gaussian data. The eigenvalues are ranked from largest to smallest and the researcher retains factors whose variances are greater than the similar-ranked eigenvalues from the null correlation matrix.

As an analog of Horn parallel analysis, we perform regression-phylofactorization on null datasets comprised of i.i.d. standard log-normally distributed abundances. Specifically, 300 datasets X were simulated with $D = 32$, $n = 10$ and $\log(x_{i,j}) \stackrel{i.i.d.}{\sim} N(0, 1)$. For each dataset, one corresponding independent variable, z , was simulated where $z_j \stackrel{i.i.d.}{\sim} N(0, 1)$. Log-normally distributed abundances were used, as opposed to compositional abundances drawn from a logistic-normal, for ease (the resulting balances are Gaussian either way) and because for each sample j , with total abundance $C_j = \sum_i x_{i,j}$, the isometric log-ratios are invariant to C_j ,

$$\log \left(\frac{g\left(\frac{\mathbf{x}_R}{C_j}\right)}{g\left(\frac{\mathbf{x}_S}{C_j}\right)} \right) = \log \left(\frac{g(\mathbf{x}_R)}{g(\mathbf{x}_S)} \right). \quad (13)$$

A single, random phylogeny with $D = 32$ species was simulated using the function `rtree` from the R package `ape` [Paradis et al., 2004] and used across all 300 replicate simulations. To simulate non-null datasets, we added an association between z and a set of b clades for $b \in \{2, 4, 8, 16\}$. Clades were drawn at random by considering nodes in the tree including the tips, not including the root and its first daughter to ensure b correspond to unique factors (the entire community descends from the root, and the sets of descendants of each of the root's two daughters are each other's complement set - compositional increases in one are equivalently modeled as compositional decreases in the other). Effects, β , were drawn at random from $\{3.1, 4.1, \dots, 18.1\}$ without replacement - the 0.1 included to ensure effects were not an integer multiple of one-another as such effects can cancel each other out exactly and are improbable in nature. The sign of association was drawn at random from $\alpha = \{-1, 1\}$, and the effect was simulated by first simulating null datasets, X , as described above, and then, for each clade c , with corresponding effects β_c , the abundances of all descendants were reassigned as

$$x_{i,j} \leftarrow x_{i,j} e^{\alpha \beta_c z_j} \quad (14)$$

for all $j \in c$.

The results of our simulations and comparison are visualized in the first row of Figure 3. The explained variance (EV) for each factor, averaged across all 300 replicates, exhibits a sharp decrease near the true number of factors, b . The EV from log-normal null datasets exhibits a steady decay until the last factors, where the rate of decay increases. The average null EV curve appears to cross the empirical EV curve at or slightly after the true number of factors. We use this result to construct a “LN” stopping criterion: a conservative number of factors to include is one less than the first factor at which the average null EV curve is greater than the empirical EV curve.

To compare the KS stopping criterion with the LN stopping criterion, 300 datasets were simulated

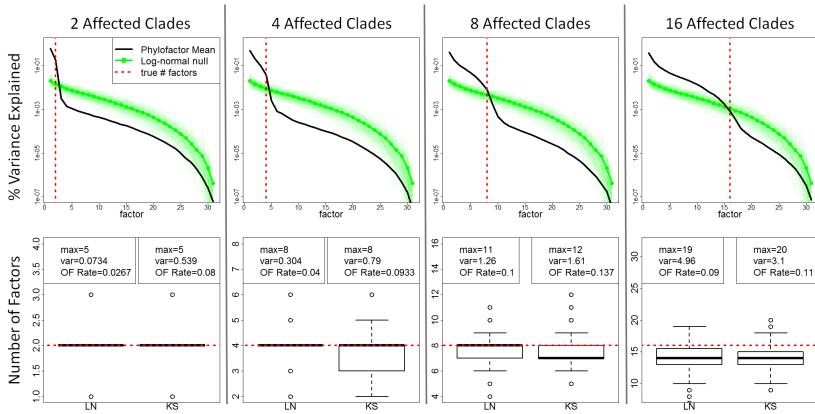


Figure 3: Performance of LN and KS stopping criteria. A challenge of phylofactorization is determining the number of factors, K , to include in an analysis, and stopping criteria aim to stop the computationally intensive iteration of phylofactorization. Abundances of $D = 32$ species across $n = 10$ samples were simulated as i.i.d. standard log-normal random variables. To simulate effects, a set of b clades were associated with environmental meta-data, \mathbf{z} , where $z_j \stackrel{i.i.d.}{\sim} N(0, 1)$. Regression-phylofactorization was performed on 300 dataset for each $b \in \{2, 4, 8, 16\}$ and for data with and without effects, with objective function ω_{ExVar} . (**top row**) The percent explained variance (EV) decreases with factor, k , and the mean EV curve for data with b affected clades intersects the mean EV curve for null data near where $k = b$, motivating a stopping criterion (LN) based on phylofactorization of null datasets to be evaluated and compared to the standard KS stopping criterion. (**bottom row**) The LN stopping criterion has a lower over-factorization (OF) rate than the standard KS stopping criterion (where OF rate is the fraction of the 300 factorizations of data with simulated effects in which $K > b$). Both criteria can be modified to be made more conservative (e.g. the P-value threshold for the KS stopping criterion can be lowered, or the LN criterion can be modified to consider when a quantile of the null EV curves, higher than the mean null EV curve, crosses the empirical EV curve). The KS stopping criterion is far less computationally intensive for large datasets, but the LN criterion may facilitate further statistical statements regarding the null distribution of test statistics in phylofactorization.

with b affected clades having an association with meta-data, and effects drawn and implemented as described above. Two phylofactorizations were implemented, one in which the phylofactorization continued until the P -value from the KS test exceeded $P = 0.01$, and another in which phylofactorization continued until the empirical EV was less than the average null EV. In both cases, the factor which satisfied the stopping criterion ($P > 0.01$ for KS, $EV_{\text{null}} > EV_{\text{obs}}$ for LN) was not included. The empirically observed number of factors, b^* , for each stopping criterion was compared to the true number of affected clades, b .

The performances of the different stopping criteria are visualized in the second row of Figure 3. The KS stopping criterion has a higher over-fitting rate than the LN, especially for low numbers of affected clades. Such high over-fitting rate could be remedied by a lower P -value threshold for the KS test. However, the KS stopping criterion is much less computationally intensive - it doesn't require null phylofactorizations which can be prohibitively costly for large datasets. The performance gains from our analog of Horn's parallel analysis may not outweigh the computational costs, especially for large datasets (a microbiome dataset of intermediate size may have 1,000 species and dozens of samples). An analytical solution for the mean EV curves under phylofactorization of standard log-normal data may prove less computationally intensive and yield a more robust stopping criterion than the KS stopping criterion. Other stopping criteria may also consider using rank estimation of X to bound criteria for choosing the number of factors. To assist future research in this area, the R package `phylofactor` contains a function `pf.nullsim` to simulate null phylofactorization of log-normally distributed data, as done here, for comparison with future stopping

criteria.

A related problem to the criteria for choosing the number of factors is the more general problem of null distributions of test-statistics under phylofactorization. For a given factor, what is the probability of an objective function as large or larger than the one observed? For a Bayesian analysis, how does an observed objective function, such as a large explained variance in regression-phylofactorization, change a prior distribution of effects on edges in the phylogeny? Phylofactorization aims to be an inferential tool making inferences on edges in the phylogeny, but, due to the complexity of statistical inference in phylofactorization, currently it is mostly an exploratory tool. Analytical derivations or estimations of the null distribution of test-statistics from phylofactorization, in particular for those from regression-phylofactorization, can move phylofactorization from an exploratory tool to a more grounded, inferential tool.

For example, consider the F-statistic from regression on $y_{p_{e^*}}$ for a factored edge, e^* , from regression-phylofactorization. Under a generalized linear model of null data like Equation (4), the F-statistic will follow an F-distribution with degrees of freedom $d_1 = m$ and $d_2 = n - 1 - m$. However, regression-phylofactorization on a tree with no polytomies and an objective function to maximize the F-statistic will choose, F_{max} , the largest of $2D - 3$ F-statistics. If the F-statistics were independent and identically distributed, we could use the fact that

$$P(F_{max} < f) = P(F_1 < f \cap \dots \cap F_{2D-3} < f) \quad (15)$$

to calculate $P(F_{max} < f) = P(F_i < f)^{2D-3}$. If the F-statistics were independent, the problem of the null distribution of regression-phylofactorization F-statistics would be solved.

Such approximation works well for the early factors of large trees but the approximation worsens for small trees and later factors of large trees due to an increasing percent-overlap of the groups, $\{R_e, S_e\}$, for each edge causing increased dependence among the F-statistics, F_i . There is a need for more accurate and universal approximations that are robust to tree size and factor number. Such approximations are crucial for disease detection and community classification, for which accurate false-positive rates are necessary for clinical use.

As with stopping criteria, phylofactorization of null datasets can be used to produce approximate statements of significance where such statements are desired. Alternatively, if one is interested in the significance of a particular factor, k , one can obtain the sub-graphs considered at that level of factorization, the corresponding groups R_e and S_e , and simulation of the null distribution - by constructing null IIR balances using log-normal data as above - may yield accurate null distributions. In anticipation of such uses, the R package `phylofactor` contains a function `pf.getGroups` to obtain the set of groups, $\{R_e, S_e\}$, considered at a given iteration of phylofactorization. Analytical tools, or computational tools that don't require repeated phylofactorization (e.g. direct, null-simulation of test-statistics under the dependence observed in the phylogeny), approximating the null distribution of test-statistics and the likelihood functions of parameters (e.g. β from regression) given observed test-statistics can greatly improve phylofactorization as an inferential tool and allow Bayesian phylofactorization with priors over edges and their associated effects.

Future work on stopping criteria and null distributions of test statistics may allow researchers to make more accurate statements about their uncertainty in the quantity of factors and their particular locations in the phylogeny. Accurate stopping criteria and null distributions can allow biologists to make more nuanced predictions from phylofactorization. For instance, quantifiable certainty about the number of factors and their locations can allow researchers to compare the complexity of trait-habitat associations tested; if a treatment affects only two or three clades out of 2,000, further research understanding the effect of treatments on microbes need only focus downstream physiological studies on two or three clades. Management of a community through such a treatment, as one may hope to modulate communities for improved human health, may be allow precise interventions to affect a few clades. However, if a treatment affects hundreds of clades, and that number S^* can be claimed to be significantly higher than another, management of a community through treatment becomes more complex. Knowing whether a treatment targets

1, 10 or 100 clades, and knowing how uncertain we are about the effects on each clade, can allow rapid progress on modulation of microbial communities.

5.2 Cross-validation

If we find mammals are more abundant than reptiles in high, northern latitudes (the mammal/reptile ratio increases as we move North), we can easily probe the generality of this result by cross-validating our finding in southern latitudes. We can easily cross-validate studies in macroscopic ecology because we can readily identify members of the same clades across environments (there is a clear definition of what is a ‘mammal’ and what is a ‘reptile’). Cross-validation of findings in microbial communities, for which most of the phylogeny is not annotated, requires consideration of the logic underlying phylogenetic cross-validation. Phylofactorization makes inferences that correspond to edges, and consequently can permit cross-validation across studies, including those with disjoint sets of species.

The idea behind cross-validation of phylofactorization is that a set of K edges or chains of edges $\{e_{i,1}\}_{i=1}^{i=K}$ are identified in a dataset \mathbf{X}_1 with phylogeny T_1 , and a subset $A \subseteq \{1, \dots, K\}$ of those edges, $\{e_{i,1}\}_{i \in A}$, are assessed for their agreement in a second dataset, \mathbf{X}_2 , with a phylogeny T_2 . The species composition of T_1 and T_2 can range from identical to disjoint, and both T_1 and T_2 are sub-graphs of a universal phylogeny, T_U . Cross-validation could be between two datasets, or it could be implemented repeatedly between two subsets of one dataset to prevent over fitting. The challenges of cross-validation of phylofactorization are both topological and statistical.

The topological challenges of cross-validation require translating the logic behind cross-validation of phylogenetic inferences, such as the mammal/reptile comparison above, onto graph-topology. How can we compare mammals/reptiles in Australia with the same in North America, when the sets of species are completely disjoint? By reliable identification of the clades, ‘mammals’ and ‘reptiles’, i.e. the groups, R_e and S_e in phylofactorization. Two major issues arise: interruptions of factored edges and whether to include previously-identified factors, which implies downstream edges contrast sub-graphs, or not, which may make cross-validation of downstream edges more robust to erroneous or irrelevant edges in previous factors. These challenges are illustrated in Figure 4.

Interruptions are nodes present in T_U and T_2 which were not present in T_1 . For instance, comparing birds to non-bird reptiles in Europe (with phylogeny T_1) is contrasting taxa along a single edge in T_1 ; such a comparison in Australia (with phylogeny T_2) would be complicated by the presence of crocodiles, whose lineage “interrupts” the edge in-between the most recent common ancestor of birds and the most recent common ancestor of birds and lizards in Europe. There are two options for how to deal with interruptions: one could either ignore the interruption (remove all descendants of the interrupting node from downstream analysis), or use the interruption to refine the location of the phylogenetic factor (does the defining feature of birds come before or after crocodiles?). Ignoring interruptions allows direct comparison of groups which led to the inference of a factor in the original dataset, but refining the location of a phylogenetic factor may increase the power across datasets by increasing the number of taxa and allowing more focused downstream studies of microbial physiology. The R package `phylofactor` contains a function `crossVmap`, which takes as input the groups, $\{R_{e_1}, S_{e_1}\}$ and the universal tree T_U , and inputs either the groups $\{R_{e_2}, S_{e_2}\}$ corresponding to the direct comparison (ignoring interruptions), or the set of all possible groups $\{R_{e_{2,i}}, S_{e_{2,i}}\}$ corresponding to each edge in the link of edges to permit refinement of the location of the phylogenetic factor (e.g. one can obtain the set of groups, {birds + crocodiles, reptiles} and {birds, lizards+crocodiles}, to determine where the meaningful difference between birds and reptiles arose).

The second topological challenge is how to deal with the hierarchical structure of factors. Given an ordered set of factored edges $\{e_i\}_{i=1}^{i=K}$, should one cross-validate the edges in the order in which they were factored for the original dataset, or should one perform some analog of phylofactorization

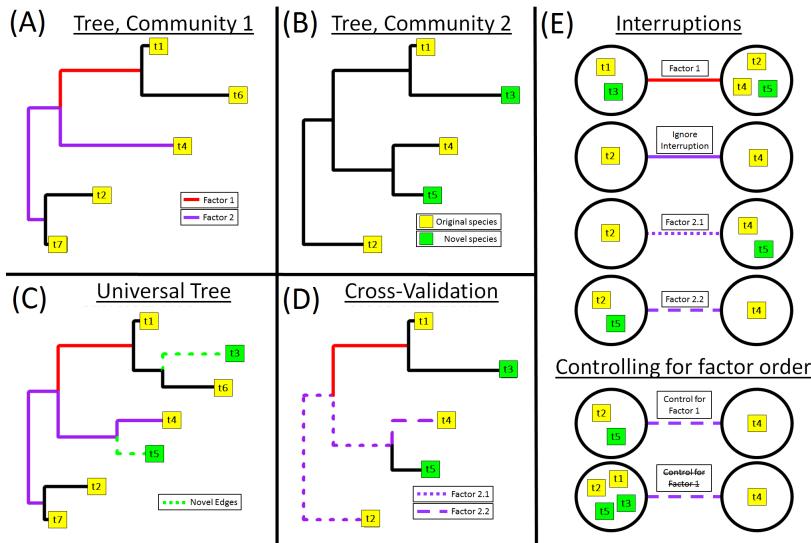


Figure 4: Topological considerations for cross-validation. Phylofactorization constructs coordinates that can be used to compare datasets and cross-validate findings, but comparisons require consideration of the graph-topology underlying phylofactorization. **(A)** One community with two factors. The second factor forms a partition separating t4 from {t2,t7}. The second factor does not correspond to a single edge, but instead a chain of two edges. **(B)** A second community, missing species t6 and t7, but containing novel species t3 and t5. **(C)** All factors can be mapped to chains of edges on a universal phylogeny. Novel species “interrupt” edges in the original tree, and cross-validation requires deciding what to do with novel species and interrupted edges. Species t3 does not interrupt a factored edge, and so t3 can be reliably grouped with t1 in factor 1. However, species t5 interrupts one of the edges in the edge-path of factor 2. **(D-E)** Interruptions can be ignored, or they can be used to refine the location of important edges (illustrated in Factor 2.1 and Factor 2.2). Another topological and statistical question is when/whether it is most appropriate to control for factor order. For instance, controlling for factor order with Factor 2.2 would partition t4 from {t2,t5}. Not controlling for factor order would partition t4 from {t1,t2,t3,t5}. Controlling for factor order allows direct correspondence between phylofactorization across datasets, but it may cause later factors to be sensitive to errors with earlier factors, which may be false-positives in the original dataset or irrelevant in novel datasets. Two applications for robust cross-validation of phylofactorization are disease diagnosis (e.g. Crohn’s disease) and annotation of the universal phylogeny to allow prediction environmental associations of novel microbes.

on this set of edges? The former allows a strict comparison of phylogenetic factorization across datasets, whereas the latter may be less sensitive to errors in the exact sequence of factors and the presence of erroneous edges or edges corresponding to traits with functions in one community but not in others.

The statistical challenges of cross-validation include standard challenges of compositional data analysis (e.g. the nested dependence of pre-determined ilr balances) and particular challenges to statistical analysis under different topological considerations listed above. Many of these challenges may have easy solutions already in existence but unknown to this author. Robust cross-validation within datasets, by repeated training and testing of subsets of the same dataset, can reduce over fitting (and thereby be related to the criteria for determining the number of factors), and robust, one-time cross-validation between two datasets can allow phylogenetically informed diagnostics of microbial community state. For example, if one finds phylogenetic factors driving Crohn’s disease [Kostic et al., 2014], it may be possible to develop diagnostic tools to assign the likelihood a patient has Crohn’s disease given a sample of their intestinal microbial communities, even if their intestinal microbial communities have novel species and interrupting nodes. For another example, cross-validation can allow refinement of the location of edges and allow for annotation of a universal tree of life to allow for prediction of environmental associations of novel, uncultivated microbes.

6 Discussion

Communities are compositions of species, and species have traits, evolved by natural selection, which determine their relative abundances, habitat associations and responses to perturbation. Traits can be mapped to edges on the phylogeny, and consequently one can describe communities as the species or as mixtures of traits, latent variables defined through phylogenetic coordinates. Phylofactorization is a method for choosing phylogenetic coordinates along which communities change, coordinates whose positions are interpretable as latent variables or functional ecological traits that evolved along the phylogeny. The phylogeny is a graph, and phylofactorization is a graph-partitioning algorithm with no balance constraint which iteratively chooses edges of importance and then partitions the phylogeny along that edge. Often, to choose which edge to cut, one is interested in a measure of difference between the two groups, as partitioning the graph along edges that “best differentiate” two groups may result in a set of sub-graphs with low within-group differences and “differences” can be the presence/absence of a trait. The balancing elements from the isometric log-ratio transform serve as a standardized measure of difference and a means to construct an orthonormal basis from phylofactorization.

An important special-case of phylofactorization is regression-phylofactorization. Because communities can often be justifiably analyzed as compositions, especially microbial communities sampled through amplicon sequencing, regression-phylofactorization entertains balances corresponding to a partition along each edge and partitions the graph along whichever edge maximizes an objective function for regression. Regression-phylofactorization can be interpreted as hierarchical regression, construction of a decision tree about species’ associations with meta-data, and a latent-variable model whose latent variables are traits, with similar challenges to factor analysis (although, strictly speaking, it is not necessarily factor analysis). By constructing a sequential binary partition corresponding to a sequence of edges in the phylogeny which explain the most total variance in a compositional dataset, regression-phylofactorization constructs a low-rank approximation of a dataset with orthonormal “loadings” (balancing elements) whose balances correspond to the relative abundances of organisms with and without a putative trait, controlling for other, previously identified putative traits.

Much like PCA, factor analysis and other clustering algorithms in their initial conception, phylofactorization is a predominantly exploratory tool, but future work on criteria for choosing the number of factors and the null distributions of test-statistics can assist the use of phylofactorization for rigorous statistical inference on edges in the phylogeny and associations between clade

abundances and meta-data. Significant progress towards understanding microbes and their traits can be made with robust cross-validation of phylofactorization to allow comparison of inferences across datasets, including datasets containing novel species. Cross-validation of phylofactorization requires careful consideration of the graph-topological inferences being made, but such tools can permit novel methods for disease detection and refined inferences on a universal phylogeny to allow predictions of habitat associations of novel, uncultivated microbes.

The challenges listed above are a small subset of future research directions which may collectively bring about more accurate, inferential phylofactorization and greatly accelerate research in the microbiome world for which most of the phylogeny is unannotated and most species have never been observed under a microscope. There are many other challenges not listed here but which are still important. For instance, like all methods in compositional data analysis, there is a need to understand the appropriate treatment of zeros for phylofactorization, which may depend on the underlying model of count data. For some ecological effects there may be more appropriate contrasts of two groups split by a partition, such as arcsines of differences of total relative abundance in two groups, R and S , used for testing of neutral drift in ecological time-series [Washburne et al., 2016]. To assist future research on phylofactorization, the R package `phylofactor` has several functions aimed at providing user-interface with internal objects for method development. The R package is available at <https://github.com/reptalex/phylofactor>.

Acknowledgments

Phylofactorization was conceived under the support of Diana Nemergut, who passed away as the method was being developed by this author. Under support from D. Nemergut's start-up funds, generously made available by Duke University, the method was further nourished by the networking, support and fruitful discussions with Juanjo Egozcue, Vera Pawlowsky-Glahn, Rob Knight, Noah Fierer, Justin Silverman and Lawrence David.

References

- John Aitchison. The statistical analysis of compositional data. 1986.
- Aydin Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent advances in graph partitioning. In *Algorithm Engineering*, pages 117–158. Springer, 2016.
- Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005.
- Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- Jack A Gilbert, Folker Meyer, Dion Antonopoulos, Pavan Balaji, C Titus Brown, Christopher T Brown, Narayan Desai, Jonathan A Eisen, Dirk Evers, Dawn Field, et al. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Standards in genomic sciences*, 3(3):243, 2010.
- John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

- Michael Kalyuzhny, Efrat Seri, Rachel Chocron, Curtis H Flather, Ronen Kadmon, and Nadav M Shnerb. Niche versus neutrality: a dynamical analysis. *The American Naturalist*, 184(4):439–446, 2014.
- Aleksandar D Kostic, Ramnik J Xavier, and Dirk Gevers. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–1499, 2014.
- Ivica Letunic and Peer Bork. Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128, 2007.
- Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.
- Jennifer BH Martiny, Stuart E Jones, Jay T Lennon, and Adam C Martiny. Microbiomes in light of traits: a phylogenetic perspective. *Science*, 350(6261):aac9323, 2015.
- Brian J McGill, Brian J Enquist, Evan Weiher, and Mark Westoby. Rebuilding community ecology from functional traits. *Trends in ecology & evolution*, 21(4):178–185, 2006.
- Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.
- Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.
- Deepak K Rajpal and James R Brown. Modulating the human gut microbiome as an emerging therapeutic paradigm. *Science progress*, 96(3):224–236, 2013.
- Lior Rokach and Oded Maimon. *Data mining with decision trees: theory and applications*. World scientific, 2014.
- Alex D Washburne, Joshua W Burby, and Daniel Lacker. Novel covariance-based neutrality test of time-series data reveals asymmetries in ecological and economic systems. *PLoS Comput Biol*, 12(9):e1005124, 2016.
- Alex D Washburne, Justin D Silverman, Jonathan W Leff, Dominic J Bennett, John L Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969, 2017.

Appendix: Mathematical Notation

- D : number of species in a community.
- n : number of communities sampled in a dataset.
- m : number of meta-data variables
- p : partition index. Each partition defines two disjoint index sets over the species in the community, $\{R_p, S_p\}$. p_e denotes partition defined by a given edge, e
- $p|k$: generic partition of one remaining group after k previous partitions.
- e : edge index. In this paper, edges will often have corresponding partitions dependent on the sub-graphs in which the edge is found.
- Δ^{D-1} : Simplex $\{\mathbf{x} \in \mathbb{R}^D | x_i \geq 0 \forall i \text{ and } \sum_i x_i = 1\}$.
- \mathbf{x}_j : compositional vector in sample j , $\mathbf{x} \in \Delta^{D-1}$.

- C_j : total abundance of a non-normalized abundance vector, $\sum_i x_{i,j}$.
- v_p : balancing element corresponding to partition, p .
- $v_{p|k}$: balancing element orthogonal to k previously defined balancing elements $\{v_l\}_{l=1}^k$
- $y_{p,j}$: ilr transform, a.k.a. balance, of sample j determined by partition, p . Used to denote candidate balances when the full sequential binary partition is not yet chosen.
- z_j : m -vector of environmental meta-data in sample j .
- X : Data matrix, $(x_{i,j}) \in \Delta^{D-1 \times n}$, whose rows are species and columns are compositional vectors, x_j .
- R_p : index set of species in one of two, disjoint groups defined in a partition. Absence of subscript implies a generic partition, sometimes subscript e is used as shorthand to denote partition defined by a given edge, p_e .
- S_p : index set of species in the complementary group to R_p , i.e. the other group defined in a particular partition.
- $g(\mathbf{x})$: geometric mean of \mathbf{x} .
- $\text{clr}_i(\mathbf{x})$: i th element of centered log-ratio transform, $\text{clr}_i(\mathbf{x}) = x_i/g(\mathbf{x})$.
- $\text{ilr}_i(\mathbf{x})$: i th element of a given isometric log-ratio transform (related to y_i , but corresponding to a complete transform instead of a candidate ilr balance).
- $\text{clr}(X)$, $\text{ilr}(X)$: clr or ilr-transformed dataset. In this paper, these operations are applied to columns of X .
- ω : arbitrary objective function.
- ω_V : variance of ilr balance
- ω_{ExVar} : explained variance from regression on ilr balance

Finding the centre: corrections for asymmetry in high-throughput sequencing datasets

Jia R. Wu¹, Jean M. Macklaim¹, Briana L. Genge¹, and Gregory B. Gloor^{1,2}

¹Dep't of Biochemistry, U. Western Ontario, London, Canada, N6A 5C1

²Dep't of Applied Mathematics, U. Western Ontario, London, Canada, N6A 5C1

gbgloor@gmail.com

Abstract

High throughput sequencing is a technology that allows for the generation of millions of reads of genomic data regarding a study of interest, and data from high throughput sequencing platforms are usually count compositions. Subsequent analysis of such data can yield information on transcription profiles, microbial diversity, or even relative cellular abundance in culture. Because of the high cost of acquisition, the data are usually sparse, and always contain far fewer observations than variables. However, an under-appreciated pathology of these data are their often unbalanced nature: i.e, there is often systematic variation between groups simply due to presence or absence of features, and this variation is important to the biological interpretation of the data. A simple example would be comparing transcriptomes of yeast cells with and without a gene knockout. This causes samples in the comparison groups to exhibit widely varying centres. This work extends a previously described log-ratio transformation method that allows for variable comparisons between samples in a Bayesian compositional context. We demonstrate the pathology in modelled and real unbalanced experimental designs to show how this dramatically causes both false negative and false positive inference. We then introduce several approaches to demonstrate how the pathologies can be addressed. An extreme example is presented where only the use of a predefined basis is appropriate. The transformations are implemented as an extension to a general compositional data analysis tool known as ALDEx2 which is available on Bioconductor.

Key words: transcriptome, Bayesian estimation, count composition, sparse data, high throughput sequencing, robust estimation, qPCR

1 Introduction

High throughput sequencing (HTS) technology is used to generate information regarding the relative abundance of features. In these designs, DNA or RNA is isolated, a library is made from a sample of the nucleic acid, and a random sample of the library is sequenced on an instrument. The output is a set of short sequence tags, called reads, which are mapped to example sequences for each feature to generate a table of read counts per feature for every sample. Traditionally, samples comprise a set of features whose identity depends on the experimental design. For example, features are genes in the case of RNA-seq or metagenomic sequencing, or are operational taxonomic units (OTUs) when the objective is identifying microbial diversity.

These data are often analyzed by count based methods, such as negative binomial or zero-inflated Gaussian models, that assume the features are independent and identically distributed for statistical tests (Auer and Doerge 2010; Anders et al. 2013). However, the capacity of the instrument used for HTS imposes an arbitrary upper limit on the total number of reads observed. Thus, data collected from high throughput sequencing are count compositions, and so counts per feature are not independent when collected in this way. In addition, several other pre- and post-sequencing steps contribute to make the data compositional (Gloor et al. 2016). Traditional tools do not address the compositional nature of HTS data (Fernandes et al. 2014; Gloor et al. 2016) and assume that the features are sufficiently independent when there are enough of them, or when they fulfill certain statistical properties (Weiss et al. 2016), although much effort is placed on ‘normalizing’ the data to have a consistent read depth (Sun et al. 2013; McMurdie and Holmes 2014).

Formally, Aitchison (1986) defined a composition as a vector \mathbf{x} of positive values $x_1 \dots x_D$ whose features sum to an arbitrary constrained constant α . Absolute values of features in a composition are uninformative, and the only information provided in compositional data are the relative magnitudes of the ratios between the pairs of components. For example, the only knowledge available is that the gene 1:gene 2 ratio is 5, but the absolute abundance of either is unavailable. Aitchison demonstrated that compositional data can be properly analyzed by log-ratios between the features, since these data carry only relative information (Aitchison 1986).

One way of satisfying the need to examine the ratios between parts is to use the centred-log-ratio (CLR) transformation proposed by Aitchison, defined as:

$$\mathbf{x}_{clr} = \log\left(\frac{x_i}{g(\mathbf{x})}\right)_{i=1 \dots D}$$

where \mathbf{x}_{clr} = A composition transformed by CLR

x_i = A feature of the non-transformed composition (\mathbf{x})

D = The number of features of \mathbf{x}

$g(\mathbf{x})$ = Geometric mean of D features of \mathbf{x}

Since all arbitrary sums are the same this led to the concept of a composition as an equivalence class where composition \mathbf{x} can be scaled into an identical composition \mathbf{y} by multiplication of a constant α (Barceló-Vidal et al. 2001). Thus, in the ideal case, we can discuss any composition as being a proportion scaled by α without loss of precision. Indeed, the CLR, and indeed any ratio-based method is, at least in theory, scale-invariant because if the parts of \mathbf{x} are counts with $\alpha = N$ reads, then:

$$\mathbf{x}_{clr} = \log\left(\frac{Nx_i}{g(N\mathbf{x})}\right) = \log\left(\frac{x_i}{g(\mathbf{x})}\right). \quad (2)$$

The important caveat that limits this ideal situation when dealing with high throughput sequencing data is that the total read count, α , for each observation must be roughly similar.

Aitchison (1986) also defined the ALR, the additive log-ratio as:

$$\mathbf{x}_{alr} = \log\left(\frac{x_i}{x_D}\right)_{i=1\dots D-1} \quad (3)$$

where, following from above, \mathbf{x}_{alr} is the composition transformed by ALR, and the denominator is the D^{th} feature of \mathbf{x} , which by convention is the feature chosen to be constant.

In the ALR, the log-ratio is thus determined by selecting one presumed invariant feature as the denominator. The ALR is surprisingly similar to the relative qPCR approach in common use in molecular biology that measures relative abundance of molecules in a mixture (Thellin et al. 1999; Vandesompele et al. 2002). Here, the feature of unknown abundance is determined relative to the abundance of a feature of (presumed) known abundance, which can be a housekeeping gene or can be a DNA molecule of known amount added to the mixture. It is well known that the relative abundance measure will change when a different DNA species is used as the denominator, leading to the use of multiple (presumed invariant) features in some cases. Thus, the ALR and CLR can be viewed as the two limits of a continuum of incomplete knowledge about the proper internal standard, or basis, by which relative abundance should be judged. The ALR uses one presumed constant feature as the basis; while the CLR presumes that the majority of features are not changed, leading to the use of the geometric mean of all features as the basis. We can however, choose to use combinations of other features as the basis.

For convenience, the analyses and discussion here are drawn from RNA-seq, or transcriptome, experiments where the data are exploring the relative abundance of features that are gene transcripts found in cells in an environment. However, the examples, results and conclusions apply without restriction to metagenomic sequencing, microbial diversity sampling (by 16S rRNA gene sequencing) or to in-vitro selection experiments (Fernandes et al. 2014).

It is common for HTS data to be sparse, that is, for a given sample to contain features with counts of 0. Furthermore, the sparsity of the samples is affected by the total number of reads obtained for each sample. Each sample in a transcriptome contains between thousands and tens of thousands of features each of which may have a potential dynamic range of over 4 orders of magnitude. In many cases a transcriptome dataset will be composed of several groups, where the expression of a feature (gene) is so low that it is below the detection limit in one group, and very high in another group. The expression of genes in biological systems is linked, and some genes control the expression of other genes, either by increasing or decreasing their relative abundance. Furthermore, the cell has a built-in control system whereby gene expression itself appears to be a composition, that is, the expression levels of all genes in a cell are constrained by an absolute upper bound (Scott et al. 2010). Note however, that this does not mean that a population of cells will have total gene expression with an upper bound, since the cells themselves can change in both absolute and relative abundance in a mixture.

2 Statement of the Problem

The assumption being made when using the CLR transformation to identify features that differ between groups is that most features are either invariant or varying at random when comparing the two groups. It is worth noting that this assumption is also made by essentially all differential abundance tools. This assumption is broken if there is any sort of systematic variation between groups. For example, when comparing microbial diversity between sampling sites or conditions, organisms present in one sub-site or condition may be absent from another (Macklaim et al. 2015; Hummelen et al. 2010; Gajer et al. 2012). In the case of multi-organism RNA-seq (meta-RNA-seq), organisms resident in one condition may have a different expression profile and abundance than those resident in a second condition (Macklaim et al. 2013). In the case of a single-organism RNA-seq, samples from one condition may contain more genes than samples from another condition (Lang and Johnson 2015; Peng et al. 2014; Zhao et al. 2013; Gierliński et al. 2015). These differences are represented by either zeroes or low count features that occur systematically in only

one group.

The potential for a change in cell number and the potential for expression linkage of genes in biological systems, coupled with the inability to collect a large enough number of sequence reads, can lead to experiments with an apparent or a real asymmetry in relative abundance of many genes or features. Such an asymmetry will result in mis-centering of the data when conducting differential abundance analyses, largely, but not exclusively because of the effect on the geometric mean upon which the CLR depends. The asymmetry will also affect the scale-invariance of the data, since a value of 0 is not scaled when multiplied by a constant. Note, that it is also entirely possible for the dataset *as a whole* to be centred, but for the particular comparison of interest to not be centred. This could arise because of a systematic experimental bias that is unknown to the investigator.

Throughout, we use two plots to summarize the location of the features in multivariate datasets. The Bland-Altman (BA) plot (Altman and Bland 1983) plots the mean log-ratio abundance on the x-axis and the difference between groups on the y-axis. The BA plot is efficient at showing the relationship between (relative, mean log-ratio) abundance and difference, but contains little information on the per-feature dispersion in the data. The Effect size plot (Gloor et al. 2016) complements the BA plot by showing the relationship between a measure of dispersion (on the x-axis) and the difference between groups (on the y-axis). All plots are in log units calculated a base of 2. The ratio between these two values is a proxy for the effect size statistic calculated by ALDEEx2. Difference and dispersion are calculated using methods that are indifferent to distributional assumptions and are defined in the methods section. We also provide supplementary examples of the same data using compositional biplots (Aitchison and Greenacre 2002) to demonstrate that similar pathologies can occur when the data are displayed in this way. Figure 1 shows that incorrect estimates of the location of the data can be achieved with seemingly minor variation within simulated data. The goal is to identify a basis that best represents each sample so features can be accurately compared even when the data contains an asymmetry.

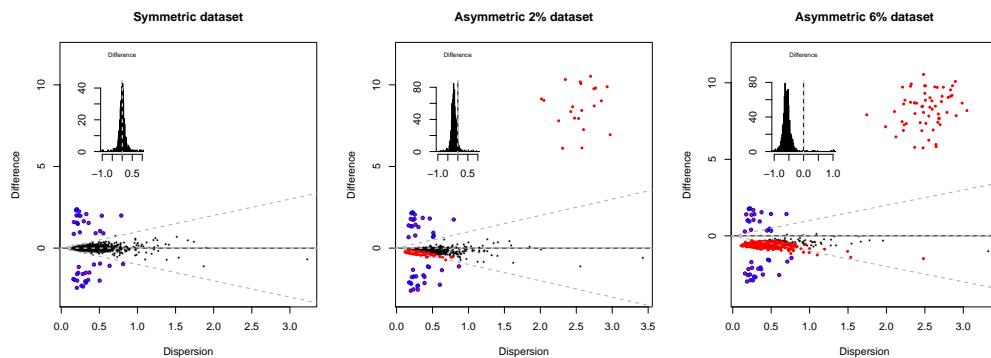


Figure 1: Effect plots of simulated asymmetric data that illustrates the problem. The effect plots show the difference between two conditions in simulated RNA-seq data with 1000 genes where 40 genes are modelled to have true difference between groups. Each point is a feature (gene), they are coloured in black if not different between groups, red if identified as being statistically significantly different between groups, and red with a blue circle if they are one of the 40 genes modelled to be true positives. The red points in the top right quadrant are the genes modelled to be asymmetrically variable between groups. These are also true positive features, but are not part of the initial modelled true positives. The inset histograms show the distribution of the differences between groups as calculated by ALDEEx2, and the vertical line shows a difference of 0. These x-axis of these plots are truncated to show only differences near the midpoint.

3 Methods

It is worth recalling that essentially all HTS data come from underpowered experimental designs, in the sense that there are more parts than there are samples: indeed it is common, because of cost to conduct and analyze only pilot-scale experiments. Thus, the strength of evidence for statistical inference must be weak, but paradoxically, the parts that are identified as differentially abundant must appear to be much more different between groups than the actual data support (Halsey et al. 2015). These can only be validated by replication or meta-analysis (Cumming 2008), both of which are rare in both the transcriptome and microbiome fields.

When estimating differential abundance it is important to properly estimate the dispersion, τ , of the j^{th} feature for all samples; dispersion can be represented by the following simple model:

$$\tau_j = \nu_j + \epsilon_j \quad (4)$$

where ν represents the underlying biological variation and ϵ represents the stochastic error from all the steps involved in the collection, preparation, and sequencing of the dataset. The majority of extant analysis tools utilize point estimates of both parameters. First, it is generally assumed that ϵ is small relative to ν . Second, it is assumed that there is some underlying similarity in the distribution of ν and ϵ for all features in all samples at a given relative abundance level. That is, if the j features were ordered by abundance, that the expected value of ν_j would be $\sim \sum(\nu_{j-m} \dots \nu_{j+m})/2m$ where m is some small offset in the abundance index. Similar logic applies to estimating the expected value of ϵ , but many tools offer more complex additional models to estimate these parameters for troublesome data.

However, we observed that ϵ can be exponentially larger than ν at the low count margin (Fernandes et al. 2013; Gloor et al. 2016), and that properly accounting for this realization alone can result in an excellent fit to even problematic data. Thus, a reliable analysis can be obtained by incorporating an ‘in silico’ technical replication which explicitly models the variation in ϵ as a probability density function on a per feature, per sample basis; in other words that $\tau_j = \nu_j + f(\epsilon_j)$. This approach is implemented in the ALDEx2 Bioconductor package and substantially reduces the false positive identification rate in microbiome and transcriptome data while maintaining an acceptable true positive identification rate (Thorsen et al. 2016).

The differences between groups, dispersion within groups and relative abundance were calculated using the ALDEx2 R package that uses Bayesian modelling that generates a probability function for ϵ_j that can be used to place bounds on the uncertainty of the the observed data (Fernandes et al. 2013; Gloor et al. 2016). If there are two groups, A and B, this requires that the data comparison is properly centred on the difference between these groups. ALDEx2 has been shown to give meaningful and reproducible results, even on sparse, asymmetric datasets using many different experimental designs (Fernandes et al. 2013; Macklaim et al. 2013; Fernandes et al. 2014; McMurrough et al. 2014), although as shown here the asymmetry can still affect the outcome.

I will adhere to the following notation when describing this process:

- indices will be denoted as lower case, italic; i.e., i, j, k, n , except for the case of the number of features in a composition, when a D will be used.
- a vector will be denoted in bold, lower case, italic; i.e., the i^{th} sample vector will be \mathbf{s}_i . Vectors derived from this vector will follow the same notation and contain D features.
- a matrix or array will be denoted in upper case, roman text; i.e., S

The starting point for analysis is an n samples $\times D$ features array. The sample vector contains the number of reads mapped to any of the j features in the i^{th} sample, $\mathbf{s}_i = [j_1, j_2 \dots j_D]$, where $i = 1 \dots n, j = 1 \dots D$. The total number of counts is irrelevant and determined by the machine

(Gloor and Reid 2016; Gloor et al. 2016). These data are compositional and are an example of an equivalence class with $\alpha_i = \sum \mathbf{s}_i$. In theory, the vector \mathbf{s}_i can be adjusted to a unit vector of proportions, $\mathbf{p}_i = [p_1, p_2 \dots p_D]$, i.e. $\alpha = 1$, without loss of information by the maximum likelihood (ML) estimate $\mathbf{p}_i = \mathbf{s}_i/\alpha_i$. In this representation, the value of the j^{th} feature is a ML estimate of the probability of observing the counts conditioned on the fractional f that the feature represents in the underlying data and on the total read depth for the sample; i.e., $\mathbb{P}_{i,j}(f_{i,j}|\alpha_i)$. However, the maximum likelihood estimate will be exponentially inaccurate when the dataset contains many values near or at the low count margin (Newey and McFadden 1994) as is common in sparse HTS data. Instead we use a standard Bayesian approach (Jaynes and Bretthorst 2003) to infer a posterior distribution of the unit vector directly from \mathbf{s}_i , by drawing k random Monte-Carlo instances from the Dirichlet distribution with a uniform, uninformative prior of 0.5, i.e.:

$$\mathbf{P}_{i(1\dots k)} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_k \end{pmatrix} = \begin{pmatrix} p_{i,11} & p_{i,21} & p_{i,31} & \dots & p_{i,D1} \\ p_{i,12} & p_{i,22} & p_{i,32} & \dots & p_{i,D2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{i,1k} & p_{i,2k} & p_{i,3k} & \dots & p_{i,Dk} \end{pmatrix} \sim Dirichlet_{(1\dots k)}(\mathbf{s}_i + 0.5) \quad (5)$$

This approach has consistent sampling properties and removes the problem of taking a logarithm of 0 when calculating the CLR because the count 0 values are replaced by positive non-zero values that are consistent with the observed count data (Fernandes et al. 2013; Gloor et al. 2016). Each of the Monte-Carlo instances, by definition, conserves proportionality and accounts for the fact that there is more information when α_i is large than when it is small. This partially restores scale invariance to the data by providing a distribution of values where the uncertainty of features scales inversely with the read depth (Fernandes et al. 2013; Gloor et al. 2016).

Each of the Monte-Carlo Dirichlet instances are CLR-transformed row-wise using Equation 1 and the entire dataset is stored in the array C with dimension n, D, k . Note that the CLR itself is scale invariant since the same output vector is obtained for all members of the equivalence class. For convenience a logarithm of 2 is used for the CLR transformation so that differences can be expressed in a intuitive scale.

Summary statistics from the distribution of CLR values for each feature can be calculated and reported as either expected values or as medians of the distributions (Fernandes et al. 2013). If we have two groups, A and B, where the indices of the samples in the first group are $1 \dots i_a$ and the indices of the samples in the second group are $i_a+1 \dots n$, then the distributions of CLR values for the j^{th} feature of the two groups can be contained in the vectors: $\mathbf{a}_j = \mathbf{C}_{(1\dots i_a)j(1\dots k)}$ and, $\mathbf{b}_j = \mathbf{C}_{(i_a+1\dots n)j(1\dots k)}$. Summary statistics use for plotting and analysis are:

- Log-ratio abundance of a feature is the median of the joint distribution of CLR values from groups A and B; i.e., it is the median of $\mathbf{a}_j \cup \mathbf{b}_j$.
- Dispersion is the median of the vector $\Delta_{\mathbf{a}_j \vee \mathbf{b}_j} = \text{maximum}(|\mathbf{a}_j - \mathbf{a}_{\langle j \rangle}|, |\mathbf{b}_j - \mathbf{b}_{\langle j \rangle}|)$, where $\langle j \rangle$ indicates a random permutation of the vector. The reported dispersion for each feature is denoted as $\tilde{\Delta}_{\mathbf{a}_j \vee \mathbf{b}_j}$ and is a conservative surrogate for the median absolute deviation when \mathbf{a}_j and \mathbf{b}_j contain many entries (Fernandes et al. 2013).
- Difference between groups is the median of the vector $\Delta_{\mathbf{a}_j - \mathbf{b}_j} = (\mathbf{a}_j - \mathbf{b}_{\langle j \rangle})$, i.e., $\tilde{\Delta}_{\mathbf{a}_j - \mathbf{b}_j}$.
- Effect size for a given feature is the median the vector derived from $\Delta_{\mathbf{a}_j - \mathbf{b}_j}/\Delta_{\mathbf{a}_j \vee \mathbf{b}_j}$, and is thus a standardized difference between the distributions in \mathbf{a}_j and \mathbf{b}_j .

3.1 Simulated Data

RNA-Seq data was simulated for benchmarking purposes. Assemblies from *Saccharomyces cerevisiae* uid 128 and a complete reference genome of *S. cerevisiae* were drawn from GenBank. The

R package **polyester v1.10.0** (Frazee et al. 2016) was used to simulate an RNA-Seq experiment with 2 groups of 10 replicates with 20x average sequencing coverage across the simulation experiment. For the base dataset, forty genes were chosen at random to have 2-5 fold expression difference, and these were apportioned equally between the two groups. These 40 features serve as an internal control of true positives for each dataset as their fold changes are explicit and should always be displayed as differentially expressed. We used bowtie2 (Langmead and Salzberg 2012) to align the simulated reads to the *S. cerevisiae* reference genome. Labeling each group as A and B is arbitrary and hence the first 10 samples belong to condition A, and the final 10 sample belong to condition B. There are a total of 6349 features in these simulated data, but only the first 1000 genes by order were chosen for the majority of the figures.

An additional 98 datasets derived from the base dataset are generated in order to benchmark how well the interquartile log-ratio (IQLR) transformation supports the assumption that most features are invariant and unchanging. As the original dataset has approximately 6000 nonzero features, 60 features are incrementally removed from the samples of condition A in each simulated dataset for a resultant set of datasets with sparsity ranging from 0% to 98% sparse in condition A.

3.2 Four alternative methods

In its current implementation, ALDEx2 computes a per-sample geometric mean for the features and declares this as the baseline for feature comparisons. The ‘Symmetric dataset’ panel in Figure 1 is an effect plot demonstrating that the 40 internal control features are found to be both statistically significant and to have an effect size greater than 1 between the two groups, and the remainder of the features have very small difference, and correspondingly have an effect size much less than 1. The inset histogram shows the distribution of difference values between groups A and B, and it is clear that it is symmetric and has a location of 0. However, the introduction of small amounts of asymmetry strongly affect the results. The asymmetric 2% dataset is the base dataset modified by setting the count value to 0 for 20 features chosen at random from Group A, and likewise the asymmetric 6% dataset has 60 features from group A set to 0. It is apparent from the two right panels of Figure 1 that this low level of simulated asymmetry breaks the assumption that most features are invariant, and the location of the difference between groups is no longer at the origin. Supplementary Figure 1 shows compositional PCA biplots of the same data, and here it is obvious that the centre of the data is not at the origin. Thus, the small amount of asymmetry is shifting the geometric mean of the data, causing bias. Thus, if even a proportion of features in a sample do not follow the central tendency of the data, the geometric mean can be unreliable as a baseline. It is unlikely that the problem will be as easy to diagnose in real data as in simulated data.

As can be seen in Equation 1, the major determinant of the centre of a sample is the denominator, or basis, used to compute the CLR. Thus, one obvious approach to address the problem is to compute the geometric mean of a subset of features that are more representative of the central tendency of the data, and to use this value as the denominator in the equation. We examined four different approaches to identifying the features to include in the denominator.

The first approach was to identify those features that have variance which is most typical across all the samples. This was done by calculating the variance of each feature after CLR transformation of the data, then identify those features with a variance between the first and third quartiles of the dataset: this is referred to as the interquartile-variable feature *IQVF* set of features. Thus, Equation 1 becomes:

$$IQLR_x = \log\left(\frac{x_i}{g(IQVF)}\right)_{i=1 \dots D} \quad (6)$$

where $IQLR_X$ is the transformed composition, and $g(IQVF)$ is the geometric mean of the IQVF features of X.

The transformation in Equation 6 is termed the IQLR transformation. The results of this method

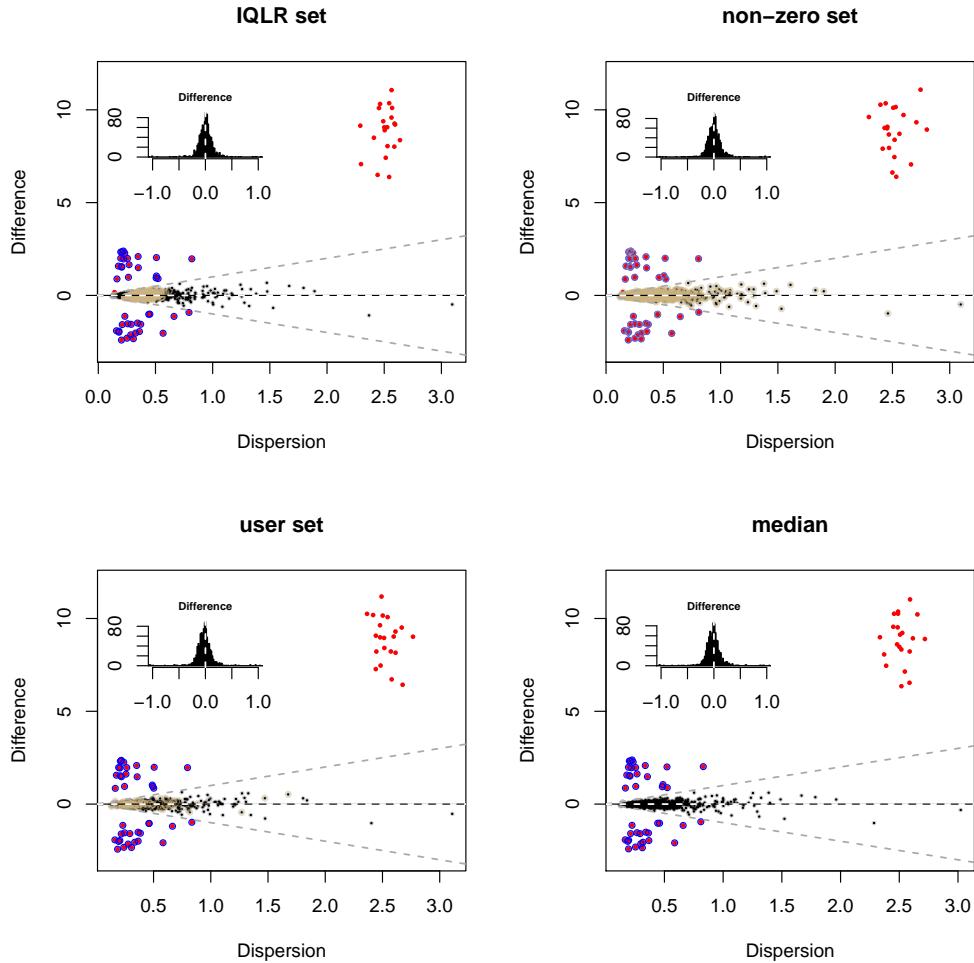


Figure 2: Effect plots of simulated asymmetric data with transformations that can result in a more accurate centring of the data. Points are coloured as in Figure 1, with the points used for the denominator in each case coloured in cyan.

are shown in Figure 2:IQLR on the Asymmetric 2% dataset. The IQLR transformation restores the centre of the dataset to the origin, and the proper set of features is identified as being both significant, and having an effect size greater than 1.

The second approach uses as the denominator the set of non-zero features in each group. Thus, in this case the geometric mean of group A and group B are based on different, but potentially overlapping, sets of features, this approach is called the no-zero log ratio (NZLR). As shown in Figure 2:non-zero, the NZLR method also restores the centre of the data to the origin and identifies the proper set of features as differential.

The third approach uses as the denominator a set of user-defined features, and is termed the ULR for user-defined log-ratio. Thus, the user could choose to use one feature, in which case the

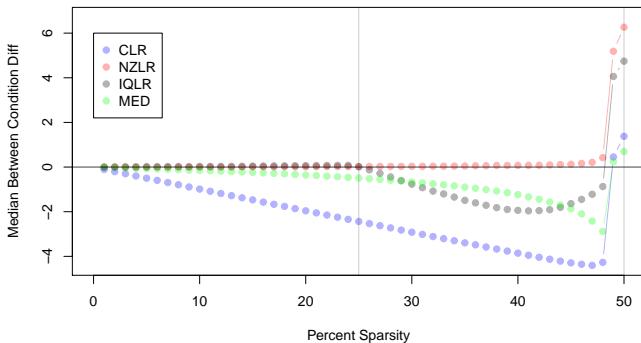


Figure 3: The behaviour of each transformations for datasets with varying sparsity. Each point represents the median between condition difference for a given transformation in a dataset with a specified sparsity. Points closer to the location $y=0$ are favourable. The CLR transformation fails as soon as asymmetric sparsity is introduced. The IQLR transformation is effective on datasets with up to 25% asymmetric sparsity from zeroes or extreme count features. The NZLR transformation is effective on datasets with up to 50% sparsity exclusively from zeroes. Replacing the geometric mean with the median in Equation 1, is an improvement, but results in a generally small shift in midpoint.

approach would be the same as the ALR, or all features, in which case the result would be the same as the CLR, or a subset chosen based upon prior information. In the case of RNA-seq, this could include the set of genes involved in translation as these have been shown to be relatively stable across multiple conditions (Scott et al. 2010), and can be presumed to represent a set of genes that are representative of the overall growth state of the cell. In principle, any set of features could be used although the investigator would need to present evidence for the appropriateness of those features chosen in any particular experiment. In the example shown Figure 2, the ‘user set’ also resulted in the location of the data being returned to 0.

The fourth approach replaces the geometric mean in Equation 1 with the median since this should be a robust estimate of the midpoint of the data.

3.3 Limitations of the approaches

We explored the limitations of these approaches in two ways. First, we examined how sparsity affected the ability of the approaches to properly centre the data when dealing with asymmetric sparse data. Figure 3 shows that the centre of the CLR-transformed data deviates from 0 when the data have even very small amounts of asymmetric sparsity. The deviation is much smaller when the median is used as the denominator, but is not, in general, the best solution. Both the IQLR and NZLR approaches are able to properly centre the data when large amounts of asymmetric sparsity are present. The breakdown point for the IQLR method is 25% sparsity in this dataset, and is approximately 45% sparsity for the NZLR. Both, are obviously better choices than the CLR, or the median choice when asymmetric sparsity is present.

Next, rather than modelling sparsity, we modelled low-count asymmetry by changing the asymmetry to a defined count of 1: note that any asymmetric count will behave similarly. In a biological context is entirely reasonable that asymmetry could occur because of low-counts rather than sparsity. For example, the default gene expression condition for many genes is low-level expression, and the inclusion of a transcriptional activator could increase expression of many genes from very low expression to very high expression. In the context of 16S rRNA gene sequencing study, it is possible for samples to be dominated by one very abundant organism but to contain many other

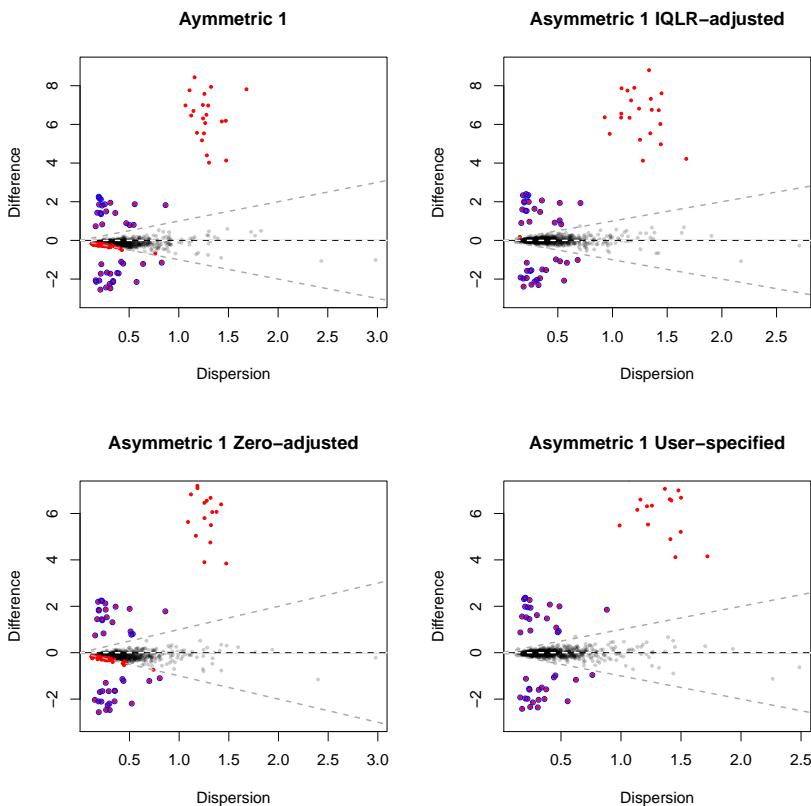


Figure 4: The behaviour of each transformations for datasets with varying 2% asymmetry where the asymmetric count value is 1. The top left panel shows that it is not sparsity that is the problem, but rather the asymmetry between groups. An asymmetry caused by a count of 1 also induces false positive identifications, shown in red below the dashed line. Both the IQLR and the User-specified transformations are able to centre the data appropriately. However, the zero-adjustment method fails because the problem is not sparsity but low counts.

taxa at low abundance. Thus, we would have a low-count asymmetry that is not necessarily based on sparsity. Furthermore, sparsity is strongly affected by read depth, the same samples derived from a sequencing dataset from an Illumina NextSeq run delivering a total count of 400M reads will be substantially less sparse than those derived from an Illumina MiSeq run delivering a total count of 25M reads, however any underlying asymmetry will be preserved.

The results, shown in Figure 4 show that an asymmetry where the asymmetric value is 1 again results in the location of the data being displaced from 0. However, when the IQLR and ULR methods are used the location of the data is restored to 0. Not surprisingly, the NZLR does not restore the data to the proper location since the asymmetry is not driven by sparsity. The median method was not tested. We suggest that the NZLR method be used only when the other approaches fail, and when the investigator is confident that sparsity is driving the asymmetry in the data.

3.4 Example of a meta-RNA-seq dataset

We finally introduce the example of an real RNA-seq dataset collected to determine the differences in gene expression of the vaginal microbial community in the healthy (H) and bacterial vaginosis

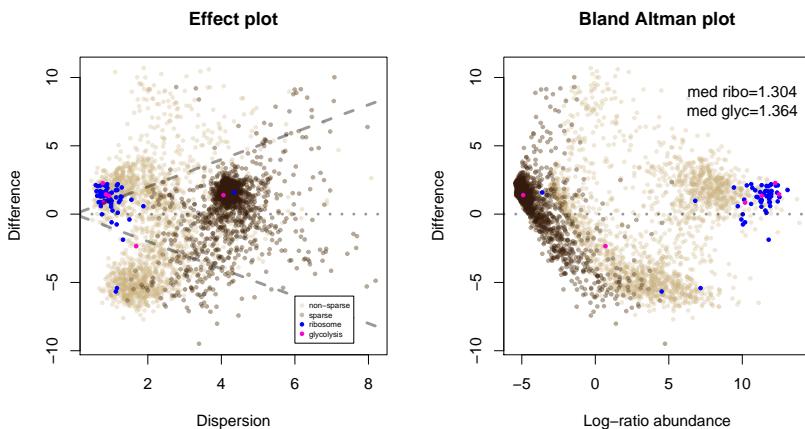


Figure 5: Effect plot and Bland-Altman plots summarizing gene expression in two different states.

(BV) states. The vaginal community can be dominated either by a few members of the *Lactobacillus* genus in the H state, or by a mixed group of anaerobic bacterial genera in the BV state (Ravel et al. 2010). In either state the members of the bacterial consortium from the other state are either very rare or absent. The data presented in Figure 5 show the distribution of between group difference, dispersion and relative abundance using an effect size plot and a Bland-Altman (BA) plot. There are 10 H samples and 12 BV samples. Each point represents the intersect of the two given summary statistics taken from the ALDEx2 output for an individual protein or enzymatic function in the dataset (Macklaim et al. 2013). These values were computed using the CLR, and we can observe the pathologic nature of the data when using this log-ratio transformation on such an asymmetric and sparse dataset.

We can see that there is a large asymmetry in distribution, the most striking of which are the functions in BV located below the midline on the y-axis; there are a large number of features centred at about -2,-5 on the Effect plot and at 5,-5 on the BA plot. This asymmetry is driven by the greater complexity of the BV microbial community, and the generally larger and more complex genomes in the set of bacteria found in BV (Macklaim et al. 2013). The asymmetry is composed of both presence-absence (sparsity) and large differences between groups. This can be seen with the sparsity overlay color, where functions that contain one or more zeros are coloured in dark brown. However, note that there appears to be two clusters of functions that are just above the y-axis midline. These are composed of functions expressed at very low relative levels, that consequently fluctuate at the level of detection in the two groups, and are centred around 4,1.5 and -5,1.5 on the Effect and BA plots respectively.

We also see a group composed of functions found in common between the H and BV group that is expressed at very high relative levels, centred around 1,1.5 and 8,1.5 on the two plots. These are functions that are central and required by all living organisms. Two sets of functions are highlighted: ribosomal protein functions in blue, and glycolytic functions in magenta. Many of the functions in these groups are often used as internal standards for comparison as it is assumed that their expression is invariant (Scott et al. 2010). The median offset of the two sets on the y-axis is given, and we can see that they are both observed to be substantially above the expected location of 0 when the CLR is used to determine differential abundance.

Finally, Figure 6 shows the result of applying the three adjustment approaches to the asymmetric meta-RNA-sea dataset. Both the IQLR and zero-adjustment methods centred the data somewhat better than when the CLR was used. We can see that the bulk of the expected invariant groups

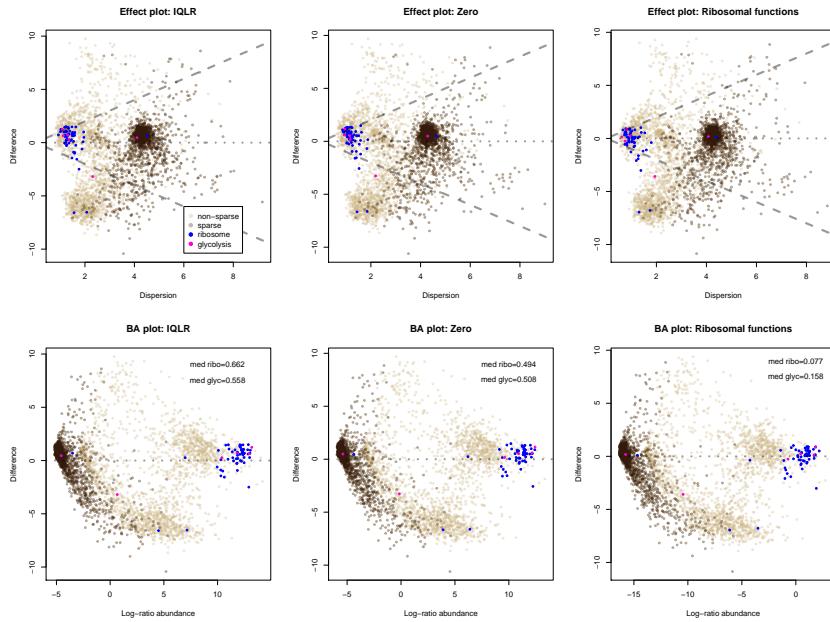


Figure 6: Effect plot and Bland-Altman plots summarizing gene expression in two different states.

are closer to the y-axis midline, and the offset of the ribosomal and glycolytic functions is reduced by half or more when compared to the CLR result. The zero-adjustment appears to be slightly better than the IQLR method in this dataset, likely because of the large amount of asymmetric sparsity. However, it is clear that the asymmetry is so extreme that neither unsupervised approach is appropriate. Centring on the geometric mean of the ribosomal functions provides a substantial improvement. Here the very rare and assumed invariant functions appear to be near the y-axis centre line. It would be a tautology to test the appropriateness using the ribosomal functions, but we can see that the glycolytic functions are nearly centred, being only slightly above the midline.

4 Conclusions

Biological data derived from high-throughput sequencing is rarely ideal and exhibits many pathologies. In particular, such data can be derived from asymmetric environments, where sets of genes, operational taxonomic units, or organisms can be present or abundant in one condition and absent or rare in another. Alternatively, an asymmetry in the data can arise because of a systematic failure in experimental design, for example, through improper blocking or the presence of outlier samples. In any of these instances the presence of an asymmetry may not be obvious.

We demonstrated that even a small number of asymmetric features can change the location of the dataset, leading to both false positive and false negative differences being identified. We showed that the asymmetry can be associated with sparsity or by differences near the margin; in either case, the pathology was similar. We tested four different methods to properly centre the data, and found that the IQLR and user-specified centring approaches were the most general purpose and recommended for use, although all are implemented in the ALDEx2 R package.

When the asymmetry is moderate, the IQLR correction is most appropriate. This correction makes the assumption that those features with variance that is found between the first and third quartile of variance, are a suitable proxy for the expected ‘typical’ variance of the data. This approach can

tolerate up to 25% asymmetry in the data when the geometric mean of these features are used as the denominator in a log-ratio normalization. In fact, we recommend that the IQLR be used as the default when performing differential abundance analysis, since this normalization makes no strong assumption about the data and appears to never perform worse than the CLR normalization.

More extreme asymmetry, as found in our vaginal transcriptome dataset, forces the investigator to make strong assumptions about the underlying data. These assumptions are similar to the assumptions made when performing qPCR: that there are one or more invariant features in the data. We showed that making the assumption that features encoding common core metabolic functions in either information processing and translation, or glycolysis, behave similarly and can be used as exemplars of ‘invariant’ features.

In either case, it must be remembered that the results of any analysis must be interpreted as *abundance relative to the chosen invariant part of the dataset*, and not as changes in absolute abundance.

Acknowledgements and appendices

Work in the lab of G. Gloor was supported by an NSERC discovery grant, RGPIN03878-2015. JMM was supported by a grant from Agrifood and Agriculture Canada.

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.
- Aitchison, J. and M. Greenacre (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51(4), 375–392.
- Altman, D. G. and J. M. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician)* 32(3), pp. 307–317.
- Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson (2013, Sep). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* 8(9), 1765–86.
- Auer, P. L. and R. W. Doerge (2010, Jun). Statistical design and analysis of RNA sequencing data. *Genetics* 185(2), 405–16.
- Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In *Proceedings of IAMG*, Volume 1, pp. 1–20.
- Cumming, G. (2008, Jul). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* 3(4), 286–300.
- Fernandes, A. D., J. M. Macklaim, T. Linn, G. Reid, and G. B. Gloor (2013, July). ANOVA-like differential expression (ALDEEx) analysis for mixed population RNA-seq. *PLoS ONE* 8(7), e67019.
- Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor (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, 15.1–15.13.
- Frazee, A. C., A. E. Jaffe, R. Kirchner, and J. T. Leek (2016). polyester: Simulate RNA-seq reads. R package version 1.10.0.
- Gajer, P., R. M. Brotman, G. Bai, J. Sakamoto, U. M. E. Schütte, X. Zhong, S. S. K. Koenig, L. Fu, Z. S. Ma, X. Zhou, Z. Abdo, L. J. Forney, and J. Ravel (2012, May). Temporal dynamics of the human vaginal microbiota. *Sci Transl Med* 4(132), 132ra52.

- Gierliński, M., C. Cole, P. Schofield, N. J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton (2015, Jul). Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics* 31(22), 3625–3630.
- Gloor, G. B., J. M. Macklaim, and A. D. Fernandes (2016). Displaying variation in large datasets: a visual summary of effect sizes. *Journal of Computational and Graphical Statistics* 25(3), 971–9.
- Gloor, G. B., J. M. Macklaim, M. Vu, and A. D. Fernandes (2016, September). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics* 45, 73–87.
- Gloor, G. B. and G. Reid (2016, Aug). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* 62(8), 692–703.
- Gloor, G. B., J. R. Wu, V. Pawlowsky-Glahn, and J. J. Egoscue (2016, May). It's all relative: analyzing microbiome data as compositions. *Ann Epidemiol* 26(5), 322–9.
- Halsey, L. G., D. Curran-Everett, S. L. Vowler, and G. B. Drummond (2015, Mar). The fickle p value generates irreproducible results. *Nat Methods* 12(3), 179–85.
- Hummelen, R., A. D. Fernandes, J. M. Macklaim, R. J. Dickson, J. Changalucha, G. B. Gloor, and G. Reid (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS One* 5(8), e12078.
- Jaynes, E. T. and G. L. Bretthorst (2003). *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press.
- Lang, K. S. and T. J. Johnson (2015, Jul). Transcriptome modulations due to a/c2 plasmid acquisition. *Plasmid* 80, 83–9.
- Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with bowtie 2. *Nature methods* 9(4), 357–359.
- Macklaim, J. M., J. C. Clemente, R. Knight, G. B. Gloor, and G. Reid (2015). Changes in vaginal microbiota following antimicrobial and probiotic therapy. *Microb Ecol Health Dis* 26, 27799.
- Macklaim, M. J., D. A. Fernandes, M. J. Di Bella, J.-A. Hammond, G. Reid, and G. B. Gloor (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* 1, 15.
- McMurdie, P. J. and S. Holmes (2014, Apr). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10(4), e1003531.
- McMurrough, T. A., R. J. Dickson, S. M. F. Thibert, G. B. Gloor, and D. R. Edgell (2014, Jun). Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. *Proc Natl Acad Sci U S A* 111(23), E2376–83.
- Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 35, pp. 2111–2245. Elsevier Science.
- Peng, J., B. Hao, L. Liu, S. Wang, B. Ma, Y. Yang, F. Xie, and Y. Li (2014). Rna-seq and microarrays analyses reveal global differential transcriptomes of mesorhizobium huakuii 7653r between bacteroids and free-living cells. *PLoS One* 9(4), e93626.
- Ravel, J., P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, R. M. Brotman, C. C. Davis, K. Ault, L. Peralta, and L. J. Forney (2010). Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A doi/10.1073/pnas.100611107*.
- Scott, M., C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa (2010, Nov). Interdependence of cell growth and gene expression: origins and consequences. *Science* 330(6007), 1099–102.

- Sun, J., T. Nishiyama, K. Shimizu, and K. Kadota (2013). TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics* 14, 219.1–219.13.
- Thellin, O., W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen (1999, Oct). Housekeeping genes as internal standards: use and limits. *J Biotechnol* 75(2-3), 291–5.
- Thorsen, J., A. Brejnrod, M. Mortensen, M. A. Rasmussen, J. Stokholm, W. A. Al-Soud, S. Sørensen, H. Bisgaard, and J. Waage (2016, Nov). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rrna gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4(1), 62.
- Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman (2002, Jun). Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. *Genome Biol* 3(7), RESEARCH0034.
- Weiss, S., W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight (2016, Jul). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 10(7), 1669–81.
- Zhao, H., C. Chen, Y. Xiong, X. Xu, R. Lan, H. Wang, X. Yao, X. Bai, X. Liu, Q. Meng, X. Zhang, H. Sun, A. Zhao, X. Bai, Y. Cheng, Q. Chen, C. Ye, and J. Xu (2013). Global transcriptional and phenotypic analyses of Escherichia coli O157:H7 strain Xuzhou21 and its pO157_Sal cured mutant. *PLoS One* 8(5), e65466.

Analysis of gamma radioactive isotopes content in surface soil layers near Longyearbyen, Spitsbergen

Z. Ziembik, A. Dołhańczuk-Śródka,
and T. Majcherczyk

University of Opole, Opole, Poland; *ziembik@uni.opole.pl*

Abstract

Arctic regions are valuable source of information regarding the global environment pollution. Because in Arctic regions there are no major local emission sources, the region is polluted mainly from emissions sources located in Europe, North America and Siberia.

Soil samples, from surface down to approx. 5 cm, were collected in the vicinity of Longyearbyen, administrative center of Svalbard Archipelago.

In the soil samples activity concentrations of 10 gamma radioactive isotopes were determined. Most of them were parts of radioactive decay series: uranium-actinium (U-235, Th-231), thorium (Ac-228, Pb-212, Bi-212), uranium-radium (Pb-214, Bi-214, Pb-210). Individual, natural K-40 and artificial Cs-137 were also determined.

For data analysis the activity concentrations of the radioisotopes were recalculated to their mass fractions. The calculated concentrations were used in the balances construction.

Different number of concentrations and different balance structures were used in the data analysis. The balances were grouped according to the concentrations number in numerator and denominator. Their actual variances were calculated and the balances were consequently sorted.

In the result's analysis the balances with the highest and the lowest variance were considered. Analysis of the balance's structures in relation to its variability enabled drawing conclusions regarding common relationships between concentrations.

Key words: radioisotope, soil, balance.

1 Introduction

Radioactive isotopes are omnipresent in the environment. A large part of them account for natural component, posing no threat to living organisms. However, some natural processes or human activity may lead to increase in their concentrations, and as a result, a threat can appear (Borrego et al., 2007; Peroni et al., 2012; Abdel Rahman et al., 2014).

As a result of failures in nuclear power plant facilities, a significant amount of radioactive isotopes can be released. An example would be accidents that occurred in the NPP in Chernobyl (1986) and in Fukushima (2011) (Aleksakhin et al., 2006; Akahane et al., 2012). Failure of various devices that use radioactive materials and disrespect to the procedures for radioactive materials use can lead to their uncontrolled release to the environment (Eisenbud and Gessel 1997; Strand et al., 1999; Copplestone et al., 2000; Pinder et al., 2009). Significant amounts of radioactive substances can also be leached from landfills of waste radioactive materials.

The main types of human activities which can cause permanent increase in the activity concentration of natural radionuclides in the environment comprises coal and metal ore mining, fossil fuels combustion in energy production and mineral phosphate fertilizer production.

In the environment the produced radioactive matter can be transported along big distances, even in scale of the Earth (Lauritzen and Mikkelsen 1999; Tsumune et al., 2011; Bossew et al., 2012).

In environment radioisotopic composition of its components can be changed in time. It is a result, for example, of radioactive matter deposition, washing or leaching. Some radioisotopes form decay chain as a result of parent – daughter relationship between them. In these decay chains a change in radioisotope concentration leads to break in radioactive equilibrium in fragment of a decay series. The way of changes is related to the disturbance character.

An analysis of temporal and spatial changes in composition, as well as relationships between concentrations in the environmental components, provides information about occurring processes.

In natural decay series 34 unique radioisotopes representing 10 chemical elements occur. Determination of their concentrations in components of the environment can provide valuable information. Analysis of joined data comprising chemical properties of an element, half-life time of radioisotope, and characteristic features of environmental component, deliver a base for studies of matter sources and transport routes, fate of dust deposited on ground and chemical compounds circulation between components of environment.

Results of investigations carried out in Arctic regions provide valuable information regarding the global environment pollution. Because in Arctic regions there are no major local emission sources, the region is polluted mainly from emissions sources located in Europe, North America and Siberia (Law et al., 2014; Kozak et al., 2013; Douglas et al., 2012; Ma et al., 2016).

In surface soil samples collected close to Longyearbyen, the largest settlement on Spitsbergen, activity concentrations of gamma radionuclides were determined. To study relationship between their concentrations a number of balances was constructed. In the result interpretation balance structure and its variance were considered.

2 Materials and methods

2.1 Materials

Surface soil samples, of the width 0-5 cm, were collected in the vicinity of Longyearbyen. It is a settlement in the Longyear Valley, on the shore of Adventfjorden, located on the west coast of Spitsbergen. The soil samples were collected at 7 sites located along Adventfjorden shore, close to Longyearbyen airport, and along Isfjorden shore. In this region higher than 0°C temperatures are observed from June to September. During samples collection, at the beginning of August, the temperatures were not lower than approx. 7-8°C, reaching 18-20°C.

2.2 Measurements results

In the soil samples activity concentrations of 10 gamma radioactive isotopes were determined. Some of them constitutes radioactive decay series: uranium-actinium (U-235, Th-231), thorium (Ac-228, Pb-212, Bi-212) uranium-radium (Pb-214, Bi-214, Pb-210). Activity concentrations of natural K-40 and artificial Cs-137 were also determined in soil the samples.

In Tab. 1 half-lives $t_{1/2}$ of the determined radioisotopes are shown.

Table 1. Half-lives of the radioisotopes determined in soil samples

Cs-137	K-40	U-235	Th-231	Pb-214	Bi-214	Pb-210	Ac-228	Pb-212	Bi-212
30.1	$1.23 \cdot 10^9$	$7.04 \cdot 10^8$	25.5	26.8	19.9	22.2	6.15	10.6	60.6
a	a	a	h	min	min	a	h	h	min

The half-lives are included in a wide range of values, from minutes to thousands of million years. If there are no rapid changes in soil composition, the short living radioisotopes indicate content of their long living ancestors. Content of the short living Pb-214 and Bi-214 is mainly related to Ra-226 ($t_{1/2} = 1.6 \cdot 10^3$ a) concentration, however they can also be affected by gaseous Rn-222 ($t_{1/2} = 3.8$ days). It could be expected that concentration of Th-231 is related to content of the parent U-235. Concentration of Ac-228 can be related to its parent Ra-228 ($t_{1/2} = 5.8$ a) or the grandparent Th-232 ($t_{1/2} = 1.4 \cdot 10^{10}$ a). Similar dependencies between concentrations can be expected also for other members of the thorium decay chain, i.e. Pb-212 and Bi-212. But an influence of Th-238 ($t_{1/2} = 1.9$ a), ancestor of these radioisotopes, should be also taken into account in data analysis.

2.3 Computations

Balances of different concentrations combinations were calculated. Starting from 2 up to 10 concentrations were used in the balance construction. Different numbers of variables were introduced in numerator and denominator of the balance.

In Fig. 1 the results of computations are illustrated in a boxplot. Each balance construction is described by a variable V with number of concentrations in numerator and denominator separated with a dot.

As it could be expected, balances variability increase with increase in concentrations number. Reasons of such effect appearance comprises, among others, the single measurement uncertainty associated with each result.

The lowest V1.1 values were observed for the radioisotopes from the thorium decay series and K-40. This observation supposes no (or only very low) action of processes which can disturb proportionality of concentrations in a decay chain. Proportionality of K-40 and thorium series radioisotopes could be a result of characteristic composition of a mineral containing these isotopes.

Explanation of the balances composed of more than 2 components is more complicated. A reasonable interpretation suggests a process in which concentrations are controlled by the law of mass action. For low variances of a balance an equilibrium state in a process can be supposed. High variances suppose highly differentiated stages of a process restricting concentration of the involved materials. But in current data analysis irreversibility of the transformation from one radioisotope to the other in radioactive decay chain has to be taken into account.

In 3 component balances, the ones with low variances contain in their structure Pb-214 and Bi-214 additionally to thorium series radionuclides. But position of corresponding radioisotopes from thorium and uranium-radium series is opposite, i.e. if concentration of an isotope from the first series appears in numerator then the isotope from the second series appears in denominator of balance. It should be noticed that common ancestor of these radioisotopes is radium, represented by Ra-226 and Ra-228 isotopes. Constant ratio of the daughter radioisotopes suppose also constant ratio of Ra-226 and Ra-228. Though these isotopes appear in different decay chains, they share the same chemical properties. As a result, for example, leaching would change radium content in soil in the same degree for both isotopes.

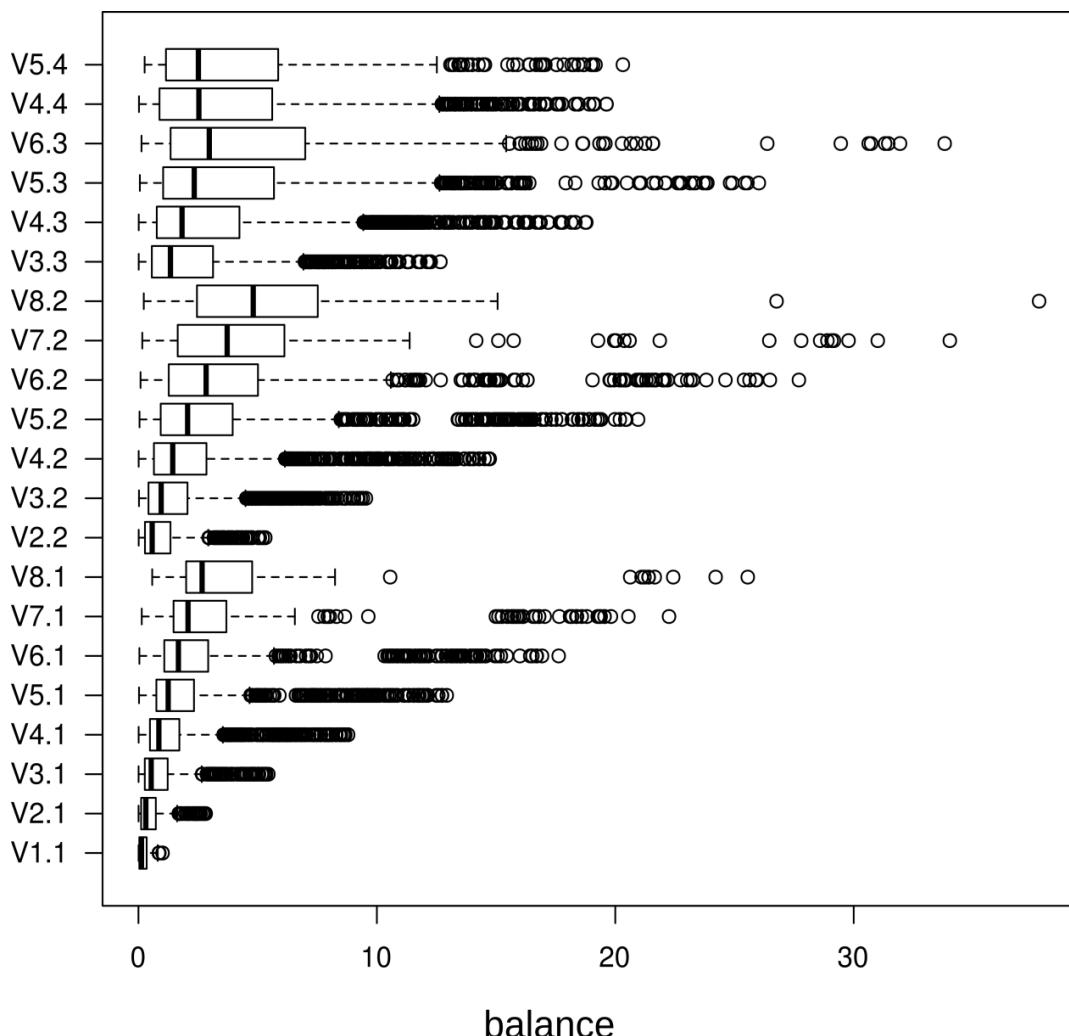


Figure 1. Boxplots of balances calculated for different combinations and number of concentrations.

The biggest V1.1 value was observed for the balances constructed from thorium series radionuclides and Cs-137. In opposite to Ra-228 daughters which are natural soil components, Cs-137 appeared in environment as a result of human activity. The most probable source of this isotope in Longyearbyen region is atmospheric deposition of global fallout. A uniform spatial distribution of such deposition could be expected, hence rather low variability in balances comprising Cs-137. The variances biggest among the calculated ones suggest negative correlation between a radioisotope belonging to thorium decay series and Cs-137. Cesium chemical properties seem to be not helpful in this phenomenon explanation.

The biggest variance was observed among V8.2 balances. It contains concentrations of K-40, U-235, Th-231, Pb-214, Bi-214, Ac-228, Pb-212, Bi-212 in numerator, and Cs-137 and Pb-210 in denominator. Though Pb-210 is a natural radioisotope with Pb-214 and Bi-214 parents, it introduces significant variability in the balance. Excessive abundance of this radioisotope are released during fossil fuels combustion. In the investigated area, heating systems and local power plant can be considered as such sources.

Conclusions

Analysis of balances can be regarded as valuable supplementary method in data exploratory analysis. Information extraction from the data involve balances construction and then analysis of their structure. However, presented in the article approach to the result interpretation requires essential improvements. Currently it depends too strong on intuitive selection of calculation results for analysis. Some information can go astray due to improper method applied for the data review.

References

- Abdel Rahman, R.O., Elmesawy, M., Ashour I., Hung Y.-T. (2014). Remediation of NORM and TENORM Contaminated Sites-Review Article. *Environmental Progress & Sustainable Energy* 33(2), pp. 588-596.
- Akahane, K., Yonai, S., Fukuda, S., Miyahara, N., Yasuda, H., Iwaoka, K., Matsumoto, M., Fukumura, A., Akashi, M. (2012). The Fukushima Nuclear Power Plant accident and exposures in the environment. *The Environmentalist* 32 (2), pp. 136–43. doi:10.1007/s10669-011-9381-2.
- Aleksakhin, R.M., Sanzarova, N.I., Fesenko, S.V. (2006). Radioecology and the accident at the Chernobyl nuclear power plant. *Atomic Energy* 100 (4), pp. 257–63.
- Borrego, E., Mas, J.L., Martín, J.E., Bolívar, J.P., Vaca, F., Aguado, J.L. (2007). Radioactivity levels in aerosol particles surrounding a large TENORM waste repository after application of preliminary restoration work. *Science of The Total Environment* 377 (1), pp. 27–35. doi:10.1016/j.scitotenv.2007.01.098.
- Bossew, P., Kirchner, G., De Cort, M., de Vries, G., Nishev, A., de Felice, L. (2012). Radioactivity from Fukushima Dai-ichi in air over Europe; part 1: spatio-temporal analysis. *Journal of Environmental Radioactivity* 114, pp. 22–34. doi:10.1016/j.jenvrad.2011.11.019.
- Copplesone, D., Johnson, M.S., Jones, S.R. (2000). Radionuclide behaviour and transport in a coniferous woodland ecosystem: The distribution of radionuclides in soil and leaf litter. *Water, Air, and Soil Pollution* 122 (3–4), pp. 389–404.
- Douglas, T.A., Loseto L.L., Macdonald R.W., Outridge P., Dommergues A., Poulin A., Amyot M., et al. (2012). The Fate of Mercury in Arctic Terrestrial and Aquatic Ecosystems, a Review. *Environmental Chemistry* 9 (4), pp. 321. doi:10.1071/EN11140.
- Egozcue, J., Pawlowsky-Glahn, V. (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology* 37 (7), pp. 795–828. doi:10.1007/s11004-005-7381-9.
- Eisenbud, M., Gessel, T. (1997). *Environmental Radioactivity. From Natural, Industrial and Military Sources*. San Diego, London, Boston, New York, Sydney, Tokyo, Toronto: Academic Press.
- Filzmoser, P., Hron K. (2008). Correlation Analysis for Compositional Data. *Mathematical Geosciences* 41 (8), pp. 905–19. doi:10.1007/s11004-008-9196-y.
- Kozak, K., Polkowska, Ź., Ruman, M., Koziol, K., Namieśnik, J. (2013). Analytical Studies on the Environmental State of the Svalbard Archipelago Provide a Critical Source of Information about Anthropogenic Global Impact. *TrAC Trends in Analytical Chemistry* 50, pp. 107–26. doi:10.1016/j.trac.2013.04.016.
- Lauritzen, B., Mikkelsen, T. (1999). A probabilistic dispersion model applied to the long-range transport of radionuclides from the Chernobyl accident. *Atmospheric Environment* 33 (20), pp. 3271–79. doi:10.1016/S1352-2310(99)00108-9.
- Law, K.S., Stohl, A., Quinn, P.K., Brock, Ch.A., Burkhardt, J.F., Paris, J-D., Ancellet, G., et al. (2014). Arctic Air Pollution: New Insights from POLARCAT-IPY. *Bulletin of the American Meteorological Society* 95 (12), pp. 1873–95. doi:10.1175/BAMS-D-13-00017.1.
- Ma, J., Hung, H., Macdonald, R.W. (2016). The Influence of Global Climate Change on the Environmental Fate of Persistent Organic Pollutants: A Review with Emphasis on the Northern Hemisphere and the Arctic as a Receptor. *Global and Planetary Change* 146, pp. 89–108. doi:10.1016/j.gloplacha.2016.09.011.
- Pawlowsky-Glahn, V., Buccianti, A., red. (2011). *Compositional Data Analysis. Theory and Applications*. United Kingdom: John Wiley & Sons, Ltd.

- Pawlowsky-Glahn, V., Egozcue, J.J.. (2006). Compositional data and their analysis: An introduction". W *Compositional Data Analysis in the Geosciences: From Theory to Practice - Special Publication no 264.* 264. Geological Society of London.
- Peroni, M., Mulas, V., Betti, E., Patata, L., Ambrosini, P. (2012). Decommissioning and Remediation of NORM/TENORM Contaminated Sites in Oil and Gas. *Chemical Engineering Transactions* 28, pp. 181–86. doi:10.3303/CET1228031.
- Pinder J.E., Hinton, T.G., Whicker, F.W., Smith, J.T. (2009). Cesium accumulation by fish following acute input to lakes: a comparison of experimental and Chernobyl-impacted systems. *Journal of Environmental Radioactivity* 100 (6), pp. 456–467.
- Strand, P., Brown, J., Drozhko, E., Mokrov, Y., Salbu, B., Oughton, D., Christensen, G., Amundsen, I. (1999). Biogeochemical behaviour of ^{137}Cs and ^{90}Sr in the artificial reservoirs of Mayak PA, Russia. *Science of The Total Environment* 241 (1–3), pp. 107–16. doi:10.1016/S0048-9697(99)00332-0.
- Tsumune, D., Aoyama, M., Hirose, K., Bryan, F.O., Lindsay, K., Danabasoglu, G. (2011). Transport of ^{137}Cs to the Southern Hemisphere in an ocean general circulation model. *Progress In Oceanography* 89 (1–4), pp. 38–48. doi:16/j.pocean.2010.12.006.



DST
Dipartimento di
Scienze della Terra



This volume contains the Proceedings of the 7th International Workshop on Compositional Data Analysis held at The Mine Museum of Abbadia San Salvatore (Siena), Italy. The increasing importance of research on compositional data analysis is testified by topics addressing different fields. In that regard, the proceedings are an excellent view of the recent state of the art.

isbn: 978-84-947240-0-8