SOFTWARE TOOL ARTICLE

Visualizing balances of compositional data: A new alternative to balance dendrograms [version 1; referees: 2 approved]

Thomas P. Quinn

Bioinformatics Core Research Group, Deakin University, Geelong, Victoria, 3220, Australia

v1 **First published:** 14 Aug 2018, **7**:1278 (doi: 10.12688/f1000research.15858.1)
Latest published: 14 Aug 2018, **7**:1278 (doi: 10.12688/f1000research.15858.1)

Abstract

Balances have become a cornerstone of compositional data analysis. However, conceptualizing balances is difficult, especially for high-dimensional data. Most often, investigators visualize balances with the balance dendrogram, but this technique is not necessarily intuitive and does not scale well for large data. This manuscript introduces the 'balance' package for the R programming language. This package visualizes balances of compositional data using an alternative to the balance dendrogram. This alternative contains the same information coded by the balance dendrogram, but projects data on a common scale that facilitates direct comparisons and accommodates high-dimensional data. By stripping the branches from the tree, 'balance' can cleanly visualize any subset of balances without disrupting the interpretation of the remaining balances. As an example, this package is applied to a publicly available meta-genomics data set measuring the relative abundance of 500 microbe taxa.

Keywords

compositional data, coda, balances, ilr, visualization, rstats, r

This article is included in the RPackage gateway.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 14 Aug 2018	✓ report	✓ report

1 **Vera Pawlowsky-Glahn**, University of Girona, Spain
Juan José Egozcue, Polytechnic University of Catalonia, Spain

2 **Marc Noguera-Julian**, IrsiCaixa AIDS Research Institute, Spain
Universitat de Vic-Universitat Central de Catalunya, Spain

Discuss this article

Comments (0)

Corresponding author: Thomas P. Quinn (contacttomquinn@gmail.com)

Author roles: Quinn TP: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Quinn TP. **Visualizing balances of compositional data: A new alternative to balance dendrograms [version 1; referees: 2 approved]** *F1000Research* 2018, **7**:1278 (doi: 10.12688/f1000research.15858.1)

First published: 14 Aug 2018, **7**:1278 (doi: 10.12688/f1000research.15858.1)

Introduction

A composition is a vector of positive measurements that sum to an arbitrary total[1]. Examples of compositions include measurements recorded in parts per million (ppm) or percentages, but also include measurements that are less obviously parts of the whole (e.g., count data generated by next-generation sequencing[2]). A component is one part of a composition. Compositional data analysis (CoDA) deals with the analysis of compositions. Compositional data, because they contain values bounded from zero to one, exist in a non-Euclidean space that render conventional statistical methods invalid. To deal with compositionality, CoDA typically begins with a log-ratio transformation that maps data into an unbounded space where conventional statistical methods can be used. The simplest transformations, the centered log-ratio transformation and the additive log-ratio transformation, use a simple reference as the denominator of the log-ratio. A more complex transformation, the isometric log-ratio transformation, transforms the composition with respect to an orthonormal basis[3]. Alternatively, one could analyze the log-ratio of each component to the other directly[4,5].

Balances use a sequential binary partition (SBP) to define an orthonormal basis that splits the composition into a series of non-overlapping groups[6]. This design allows for an interpretation of the data at the level of the isometric log-ratio coordinates[7]. This SBP contains a diverging set of contrasts that are each interpretable as a measure of "Group 1 vs. Group 2" (following an isometric log-ratio transformation). For a D-part composition, the SBP defines $D - 1$ balances that decompose the variance such the sum of the sample-wise variances for each balance in the tree equals the total sample-wise variance[6]. Balances (like the centered log-ratio transformation and the isometric log-ratio transformation) satisfy all properties required for compositional data analysis: scale invariance, permutation invariance, perturbation invariance, and sub-compositional dominance (reviewed in 8 and elsewhere).

Although balances have proved useful for the analysis of compositional data, their usual application depends on generating a meaningful SBP. Sometimes, this involves manually creating an SBP based on expert opinion, with or without the assitance of exploratory analyses[6]. However, using expertise to build an SBP is not always desirable, especially for high-dimensional data (where each composition can measure thousands of components). Principal balance analysis is a data-driven alternative that, similar to principal component analysis, seeks to identify an SBP whose balances successively explain the maximal variance of a data set (a computationally expensive procedure approximated with heuristics)[9,10]. In the field of meta-genomics, where next-generation sequencing is used to count the relative abundance of microbe taxa, scientists have applied balances of SBPs to summarize and classify microbiome samples[11]. One study defined the SBP by hierarchically clustering the microbe taxa based on the outcome of interest[12]. Another defined the SBP based on the phylogenetic relationship between microorganisms[13].

Once an SBP is generated, its balances can be visualized using a balance dendrogram[14]. The balance dendrogram illustrates (a) the distribution of samples across the balance, (b) the relationship between balances along the SBP tree, and (c) the decomposition of variance[6,15]. In addition, a balance dendrogram can show differences between sub-groupings of samples by coloring facets of the box plots. Although balance dendrograms capture a vast amount of data, the balance dendrogram may not provide the optimal visualization of balances. First, by building the figure around a tree, balance dendrograms place emphasis on the relationship between the balances, and not on the balances themselves. Second, each box plot has a unique scale positioned sporadically along the tree such that direct comparisons between one balance and all others become difficult. Third, the decomposition of variance uses lines that run parallel to the dendrogram branches, potentially confusing these concepts through use of a common symbol. In this software article, I present the R package balance for visualizing balances of compositional data. This package provides an alternative to the balance dendrogram that I hope will simplify balances for scientists less familiar with compositional data analysis.

Methods

Implementation

Within the R package universe, there are three standalone and well-documented tools for general compositional data analysis: `compositions`[16], `robCompositions`[17], and `zCompositions`[18]. The `compositions::CoDaDendrogram` function plots an archetypal balance dendrogram. There are also a number of domain-specific tools, tailored to next-generation sequencing data, and shown to work effectively[19,20]: `ALDEx2`[21,22] and `ANCOM`[23] for differential abundance analysis, `SparCC`[24] and `SPIEC-EASI`[25] for the correlation analysis of sparse networks, `propr`[26,27] for proportionality analysis, and `philr`[13] for the analysis of phylogeny-based balances. Of these, the `philr` package computes balances and visualizes them with dendrograms, but does not plot a balance dendrogram *per se*.

The `balance` package is available for the R programming language and uses `ggplot2`[28] to visualize the distribution of samples across balances of a sequential binary partition (SBP) matrix. Each balance is calculated by the formula:

$$b_i = \sqrt{\frac{|i_p||i_n|}{|i_p| + |i_n|}} \log\left[\frac{g(i_p)}{g(i_n)}\right]$$

for $b_i = [b_1, ..., b_{D-1}]$ balances where $g(x)$ is the geometric mean of x, i_p is the sub-composition of *positively*-valanced components, and i_n is the sub-composition of *negatively*-valanced components. Here, $|i_p|$ describes the norm, or length, of the sub-composition.

Operation

The `balance` package[29] computes and visualizes balances of compositional data. It requires few package dependencies, has negligible system requirements, and runs fast on a standard laptop computer (e.g., any modern budget CPU with 4GB RAM). To use `balance`, the user must provide a compositional data set (e.g., Table 1: samples as rows and components as columns) and a serial binary partition (SBP) matrix (e.g., Table 2: components as rows and balances as columns). Below, `balance` is shown for an example data set from `robCompositions`[17].

```
library(robCompositions)
library(balance)
data(expenditures, package = "robCompositions")
y1 <- data.frame(c(1,1,1,-1,-1),c(1,-1,-1,0,0),
                 c(0,+1,-1,0,0),c(0,0,0,+1,-1))
res <- balance(expenditures, y1)
```

Table 1. An example of a compositional data set with 20 sample compositions measuring 5 components each. As compositional data, the total expenditure for each individual is arbitrary. These example data are taken from `robCompositions`[17].

housing	foodstuffs	alcohol	other	services
640	328	147	169	196
1800	484	515	2291	912
2085	445	725	8373	1732
616	331	126	117	149
875	368	191	290	275
770	364	196	242	236
990	415	284	588	420
414	305	94	68	112
1394	440	393	1161	636
1285	374	363	785	487
1102	469	243	496	388
1717	452	452	1977	832
1549	454	424	1345	676
838	386	155	208	222
845	386	211	317	280
1130	394	271	490	386
1765	466	524	2133	822
1195	443	329	974	523
2180	521	553	2781	1010
1017	410	225	419	345

Optionally, users can color components or samples based on user-defined groupings. To do this, users must provide a vector of group labels for each component via the d.group argument (or for each sample via the n.group argument). The `boxplot.split` argument facets the box plots similar to the balance dendrogram[15].

```
group <- c(rep("A", 10), rep("B", 10))
res <- balance(expenditures, y1, n.group = group, boxplot.split = TRUE)
```

Figure 1 compares the balance dendrogram to its alternative using the `robCompositions` data[17].

Table 2. An example of a serial binary partition (SBP) matrix with 5 components split into 4 balances. These example data are taken from `robCompositions`[17].

z1	z2	z3	z4
1	1	0	0
1	-1	1	0
1	-1	-1	0
-1	0	0	1
-1	0	0	-1

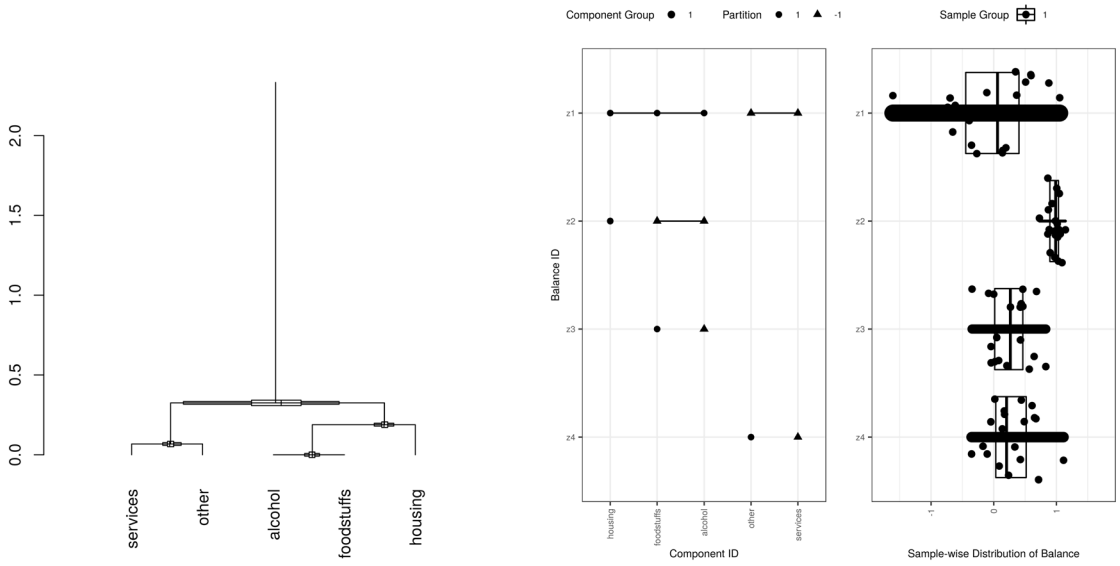

Figure 1. This figure shows a balance dendrogram and its alternative, both prepared using the data from Table 1 and Table 2. On the left, first branch of the balance dendrogram shows how the "services" and "other" components are contrasted against the remaining components. The box plot positioned at the branch shows the distribution of samples within this balance. The length of trunk shows the proportion of variance explained by this balance. On the right, this same information gets captured by a two-panel figure. The top balance in the left panel shows how the "services" and "other" components are contrasted against the remaining components. The top balance in the right panel shows the distribution of samples within this balance. In the right panel, the line length shows the range of the sample distribution, while its thickness shows the proportion of variance explained. Note that the median of this first contrast sits slightly positive, meaning that the most samples spend more on ["alcohol", "foodstuff", "housing"].

Use cases

As a use case, a publicly available microbiome data set is analyzed using balances. These data measure the abundance of microbe taxa in the feces of diabetics and their non-diabetic relatives[30], making it a true relative data set. Since these data contain many zeros that disrupt the log-ratio transformations, the zeros are first replaced through imputation by the `zCompositions` package. See the Supplementary Information for a demonstration of other pre-processing steps.

To identify balances for visualization, a serial binary partition (SBP) matrix is made by hierarchically clustering components based on their proportionality measure ϕ_s (used here as a dissimilarity measure[27]), thus joining together components that covary similarly across all samples. The `ape`[31] and `philr`[13] packages transform the tree object into an SBP ready for analysis and visualization.

```
# for compositional data with samples as rows
data.no0 <- zCompositions::cmultRepl(data, method = "CZM")
pr <- propr::propr(data.no0, metric = "phs")
h <- hclust(as.dist(pr@matrix))
phylo <- ape::as.phylo(h)
sbp <- philr::phylo2sbp(phylo)
# it is helpful to name the balances
colnames(sbp) <- paste("z", 1:ncol(sbp))
res <- balance::balance(data.no0, sbp, size.text = 4,size.pt = 1)
```

Supplementary Figure 1 visualizes all 499 balances and contains the same information that a balance dendrogram would contain: (a) the left panel dot plot shows the components being contrasted, (b) the right panel box plot shows the distribution of samples across each balance, and (c) the right panel line length shows the range of the balance (the range should cleanly approximate the decomposition of variance for purpose of exploratory visualization, though line width can optionally show the actual proportion of explained variance if desired). However, unlike a balance dendrogram, components and samples are projected on a common scale that facilitates direct comparisons and accommodates high-dimensional data. Yet, the main advantage of the `balance` package is that, by stripping the branches from the tree, it becomes possible to visualize any subset of balances without disrupting the interpretation of the remaining balances. In Figure 2, we subset the visualization to include only the top 10 most explanatory balances, ranked by the proportion of variance explained.

```
# full balances stored in results of balance plot
balances <- res[[3]]
vars <- apply(balances, 2, var)
rank <- order(vars, decreasing = TRUE)[1:10]
res <- balance::balance(data.no0, sbp[,rank], size.text = 4)
# then view for further study
sbp[,rank]
```

The `d.group` and `n.group` arguments offer a way to organize the results in a meaningful way. For example, the `d.group` can label microbes that most interest investigators, while the n.group can label patients based on clinical findings. Here, colored components (`d.group`) indicate the availability of supplemental metatranscriptomic data, while colored samples (`n.group`) indicate the presence or absence of type-1 diabetes. In Figure 3, we repeat the visualization of the top 10 most explanatory balances, with points colored by the user-defined groupings.

Summary

Compositional data measure parts of a whole such that the total sum of the composition is irrelevant and each part is only interpretable relative to others. The analysis of composition data requires interpreting the parts of the composition relative to the others. Log-ratio transformations offer a way to transform the data into an unbounded space where the analyst can apply conventional statistical methods. One transformation is the isometric log-ratio transformation which transforms the composition with respect to an orthonormal basis. Balances use a sequential binary partition (SBP) to define an orthonormal basis that splits the composition into a series of non-overlapping groups. Balances can help the investigator identify trends in relative data, and are

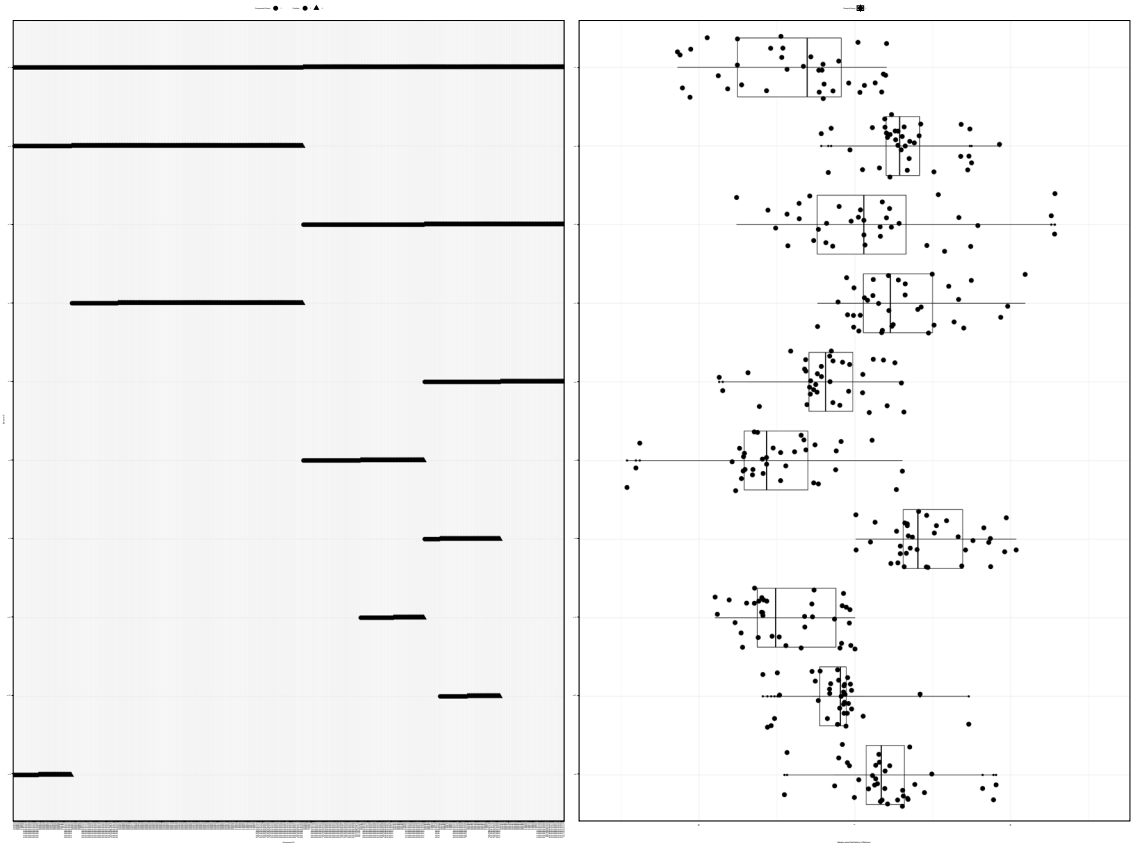

Figure 2. This figure shows the top 10 most explanatory balances, ranked by the proportion of variance explained. The left panel shows how select microbe taxa are contrasted against others. The right panel shows the corresponding distribution of samples within each balance, with the line length showing the range of the distribution. Many of the most explanatory balances occur toward the base of the serial binary partition (SBP) matrix. Yet, this subset visualization is not feasible with the balance dendrogram. Note that the order among the top 10 balances is determined procedurally to place the base of the tree at the top of the figure.

often visualized using a balance dendrogram. However, the balance dendrogram is not necessarily intuitive and does not scale well for large data. This paper introduces the `balance` package for the R programming language, a package for visualizing balances of compositional data using an alternative to the balance dendrogram. This alternative contains the same information coded by the balance dendrogram, but projects data on a common scale that facilitates direct comparisons and accommodates high-dimensional data. By stripping the branches from the tree, `balance` can cleanly visualize any subset of balances without disrupting the interpretation of the remaining balances.

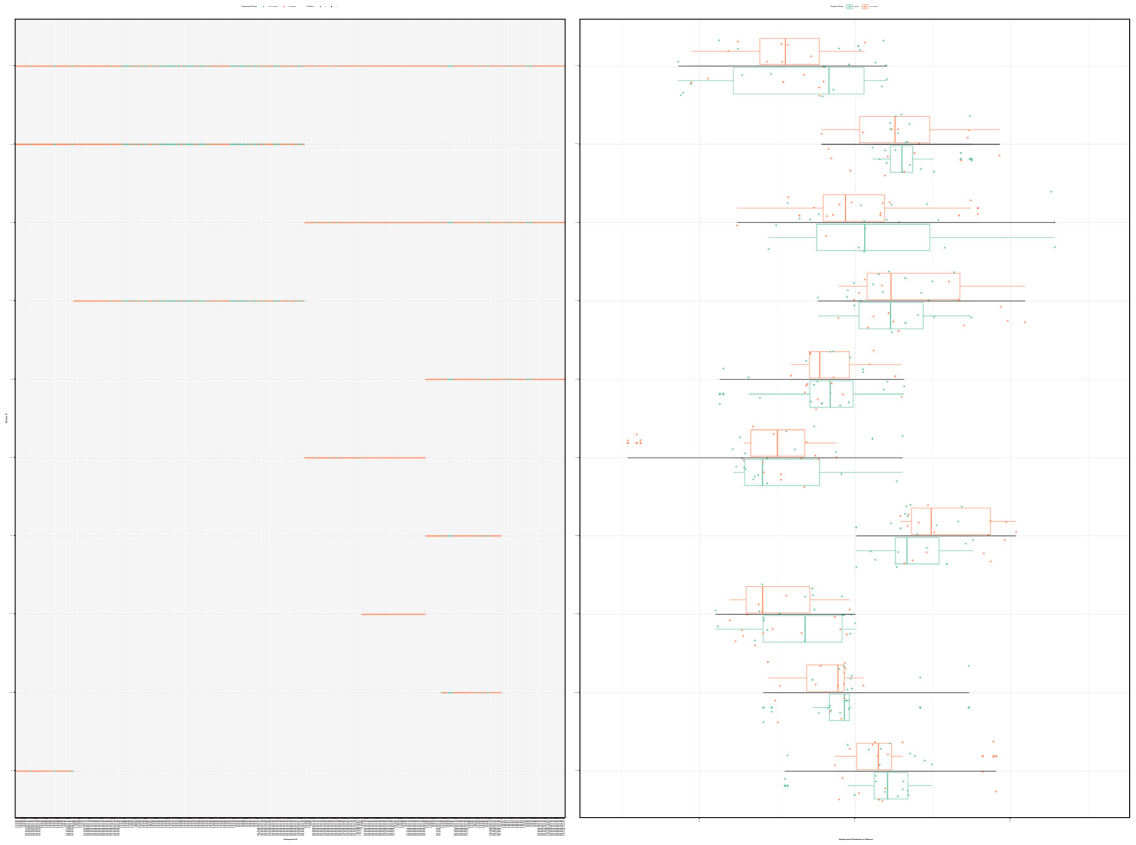

Figure 3. This figure shows the top 10 most explanatory balances, ranked by the proportion of variance explained, with points colored by the user-defined groupings. The left panel shows how select microbe taxa are contrasted against others. The right panel shows the corresponding distribution of samples for each group within each balance, with the line length showing the total range of the distribution. There is apparently a difference in the median values of diabetics and non-diabetics for some balances. One could test the significance of these differences using conventional statistical methods like the Student's *t*-test[32]. Note that the order among the top 10 balances is determined procedurally to place the base of the tree at the top of the figure.

Data availability

All data used for this analysis were acquired from the supplement of Heintz-Buschart *et al.*[30]. The supplement of this manuscript contains code to pre-process these data and reproduce the analysis.

Software availability

Software and source code available from: https://github.com/tpq/balance

Archived source code at time of publication: https://doi.org/10.5281/zenodo.1326860[29]

Software license: GPL-2

Author contributions

T.P.Q. designed the project, implemented the package, and wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

T.P.Q. thanks Sam Lee for his help rubber duck debugging.

Supplementary material

Supplementary Information. All data and scripts needed to reproduce the analysis.

Click here to get the data.

Supplementary Figure 1. Visualization of all 499 balances of the example microbial taxa data set. While this figure contains the same information as a balance dendrogram, it projects data on a common scale that facilitates direct comparisons and accommodates high-dimensional data.

Click here to get the data.

References

1. Aitchison J: **The Statistical Analysis of Compositional Data.** Chapman & Hall, Ltd., London, UK, UK, 1986.
Reference Source
2. Quinn TP, Erb I, Richardson MF, *et al.*: **Understanding sequencing data as compositions: an outlook and review.** *Bioinformatics.* 2018; **34**(16): 2870–2878.
PubMed Abstract | Publisher Full Text
3. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, *et al.*: **Isometric Logratio Transformations for Compositional Data Analysis.** *Math Geol.* 2003; **35**(3): 279–300.
Publisher Full Text
4. Greenacre M: **Towards a pragmatic approach to compositional data analysis**. Technical Report 1554, Department of Economics and Business, Universitat Pompeu Fabra, 2017.
Reference Source
5. Erb I, Quinn T, Lovell D, *et al.*: **Differential Proportionality - A Normalization-Free Approach To Differential Gene Expression**. *Proceedings of CoDaWork 2017, The 7th Compositional Data Analysis Workshop; available under bioRxiv*, 2018; 134536.
Publisher Full Text
6. Pawlowsky-Glahn V, Egozcue JJ: **Exploring Compositional Data with the CoDa-Dendrogram.** *Austrian J Stat.* 2011; **40**(1&2): 103–113.
Reference Source
7. van den Boogaart KG, Tolosana-Delgado R: **Descriptive Analysis of Compositional Data**. In *Analyzing Compositional Data with R*, Use R!, Springer, Berlin, Heidelberg, 2013; 73–93.
Publisher Full Text
8. van den Boogaart KG, Tolosana-Delgado R: **Fundamental Concepts of Compositional Data Analysis**. In *Analyzing Compositional Data with R*, Use R!, Springer Berlin Heidelberg, 2013; 13–50.
Publisher Full Text
9. Pawlowsky-Glahn V, Egozcue JJ, Tolosana Delgado R: **Principal balances**. *Proceedings of CoDaWork 2011, The 4th Compositional Data Analysis Workshop*, 2011; 1–10.
Reference Source
10. Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, *et al.*: **Advances in Principal Balances for Compositional Data.** *Math Geosci.* 2018; **50**(3): 273–298.
Publisher Full Text
11. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, *et al.*: **Balances: a New Perspective for Microbiome Analysis.** *mSystems.* 2018; **3**(4): pii: e00053-18.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Morton JT, Sanders J, Quinn RA, *et al.*: **Balance Trees Reveal Microbial Niche Differentiation.** *mSystems.* 2017; **2**(1): pii: e00162-16.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Silverman JD, Washburne AD, Mukherjee S, *et al.*: **A phylogenetic transform enhances analysis of compositional microbiota data.** *eLife.* 2017; **6**: pii: e21887.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Egozcue JJ, Pawlowsky-Glahn V: **Groups of Parts and Their Balances in Compositional Data Analysis.** *Math Geol.* 2005; **37**(7): 795–828.
Publisher Full Text
15. Thió-Henestrosa S, Egozcue JJ, Pawlowsky-Glahn V, *et al.*: **Balance-dendrogram. A new routine of CoDaPack.** *Comput Geosci.* 2008; **34**(12): 1682–1696.
Publisher Full Text
16. van den Boogaart KG, Tolosana-Delgado R: **"compositions": A unified R package to analyze compositional data.** *Comput Geosci.* 2008; **34**(4): 320–338.
Publisher Full Text
17. Templ M, Hron K, Filzmoser P: **robCompositions: an R-package for robust statistical analysis of compositional data.** John Wiley and Sons, 2011.
Publisher Full Text
18. Palarea Albaladejo J, Martín Fernández JA: **zCompositions - R package for multivariate imputation of left-censored data under a compositional approach.** *Chemometr Intell Lab Syst.* 2015; **143**: 85–96.
Publisher Full Text
19. Thorsen J, Brejnrod A, Mortensen M, *et al.*: **Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rRNA gene amplicon data analysis methods used in microbiome studies.** *Microbiome.* 2016; **4**(1): 62.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Quinn TP, Crowley TM, Richardson MF: **Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods.** *BMC Bioinformatics.* 2018; **19**(1): 274.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Fernandes AD, Macklaim JM, Linn TG, *et al.*: **ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq.** *PLoS One.* 2013; **8**(7): e67019.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Fernandes AD, Reid JN, Macklaim JM, *et al.*: **Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis.** *Microbiome.* 2014; **2**: 15.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Mandal S, Van Treuren W, White RA, *et al.*: **Analysis of composition of microbiomes: a novel method for studying microbial composition.** *Microb Ecol Health Dis.* 2015; **26**: 27663.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Friedman J, Alm EJ: **Inferring correlation networks from genomic survey data.** *PLoS Comput Biol.* 2012; **8**(9): e1002687.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Kurtz ZD, Müller CL, Miraldi ER, *et al.*: **Sparse and compositionally robust inference of microbial ecological networks.** *PLoS Comput Biol.* 2015; **11**(5): e1004226.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, *et al.*: **Proportionality: a valid alternative to correlation for relative data.** *PLoS Comput Biol.* 2015; **11**(3): e1004075.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Quinn TP, Richardson MF, Lovell D, *et al.*: **propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis.** *Sci Rep.* 2017; **7**(1): 16252.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.
Reference Source
29. Quinn T: **tpq/balance: balance-0.0.8 (Version balance-0.0.8).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1326860
30. Heintz-Buschart A, May P, Laczny CC, *et al.*: **Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes.** *Nat Microbiol.* 2016; **2**(1): 16180.
PubMed Abstract | Publisher Full Text
31. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics.* 2004; **20**(2): 289–290.
PubMed Abstract | Publisher Full Text
32. Student: **THE PROBABLE ERROR OF A MEAN.** *Biometrika.* 1908; **6**(1): 1–25.
Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 03 September 2018

doi:10.5256/f1000research.17311.r37158

✔ **Marc Noguera-Julian** [1,2]

[1] IrsiCaixa AIDS Research Institute, Badalona, Spain

[2] Universitat de Vic-Universitat Central de Catalunya, Catalonia, Spain

Manuscript by Quinn entitled “Visualizing balances of compositional data” introduces a new way to visualize balances and component partitions in compositional data. Typically, these are visualized using dendrograms where branches represent the component partitions (obtained through an external method and/or expert knowledge). The height of the branching point in the y-axis represents the proportion of variance explained, while intersection on the branching point represents the mean of the balance. Structure of the dendrogram relates to the partition hierarchy. In addition, boxplots can be defined on top of the dendrogram, at each branching point, depicting the distribution of balance values. These dendrograms are useful when a bunch of variables are analyzed but are hardly interpretable when in a high-dimensional space, which is often the case in omics derived data.

At present, the interpretation of CoDA results in terms of clinical/biological meaning is one of the bottlenecks for the adoption of such theoretical frameworks in omics data-based clinical research. In this context, new ideas on CoDA visualization are welcome and useful.

Quinn’s proposal splits the data that characterizes balances into two parts, one regarding the partition of the components into balances (left sub-plot) and the other to the distribution of the balances values when applied to the compositional data and the distribution of the variance of the same data (right sub-plot).This is similar to moving the boxplots aside from the dendrogram and turning the dendrogram into a simple sequential partition diagram. In addition, a grouping factor can be projected into the right sub-plot which facilitates to see the difference in one or more balances according to a response categorical variance, due to a common-scale axis for all balances.
The representation of balances is now “free” from a dendrogram structure and this presents some advantages such as ranking or selecting over the balances but also some inconveniences such as the need to check which components are in the “num” and “den” for each of the balances, which is now a moving target.

I think that this new presentation of balance data is useful in the low-dimensional setting. Unfortunately, the interpretation of these split diagrams appears still difficult in high-dimensional space as shown in sup info example and needs an inspection of each balance sub-space. While representing proportion of variance using line thickness is innovative, it is also difficult to compare between different balances.

The code to do this is based on ggplot2, available and easy to use and adapt to each user needs. Input is

a compositional data frame and a binary partition, depending on the role of the component in the balance. Thus, the input must be generated outside of the present code, which only represents the data, which allows for flexibility, and the code is only focused on calculation and representation of pre-specified balances.

I have some minor suggestions to improve the functionality of the code which may facilitate data interpretation:

- Selection: Author has added the possibility to plot multi-group boxplots. The author already mentions that this could be used for statistical testing, but it would be helpful to add an option select/highlight those balances that have statistically significant differences among groups and/or plot some statistical testing results on the right-hand-plot.
- Ranking: It looks like balances are ranked on the proportion of variance explained. However, in the provided code/examples, it is unclear whether the default ranking of the balances is the decreasing proportion of variance explained since this is not the case when the *weigh.var* option is set to TRUE. Also, it would also be useful to rank the balances according to their discrimination power over a response variable.
- With the standard balance dendrogram, when overlapping datasets, the variance for each of the subgroups could be represented (Pawlowsky-Glahn, Egozcue, Austrian Journal of Statistics,2011). I think this feature is lost in this representation.

The manuscript is well written and easy to follow by the expert reader. I'd like to highlight some minor points.

- The author states that these diagrams can accommodate high dimensional data, but it does that by focusing on sub-groups of the high dimensional data, and in the code, these sub-groups are selected previous to the *balance* function. Therefore, the proposed code/diagram can really accommodate subsets of high dimensional data.

Some details that caught my attention and that may be useful for future development:

- The name of the main function is *balance*. It overlaps with compositions::balance and ape::balance. While this is not a hard problem it may add confusion to the function namespace in this kind of analysis. I'd consider it changing it at this stage.
- When group-based boxplotting, data points are scattered over both (or more) boxplots, it would be clearer if data points were scattered over their own group boxplot.
- In the Figure S1 example. Some balances show zero values and near-zero variances, probably due to zero over-inflation and the zero-dominance(downstream imputed) balances. This is kind of a white noise in the diagram.
- Correlation: While it is outside the scope of this work, for the sake of utility, it would be helpful, when there is a continuous response variable, to plot points within the scatter plot within a y sub-axis for each boxplot in such a way that correlation between each of the balances and the response variable is visible and also be able to highlight/select those balances.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics, Data Analysis, Metagenomics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 28 August 2018

doi:10.5256/f1000research.17311.r37159

Vera Pawlowsky-Glahn [1], **Juan José Egozcue** [2]
[1] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain
[2] Department of Civil and Environmental Engineering, Polytechnic University of Catalonia, Barcelona, Spain

We have not been able to access the R-library “balance”. Consequently, we base our comments exclusively on the text presented. This circumstance motivates the answer “partly” to some of the previous questions. Thus, we cannot evaluate how well the software performs with particular datasets or how well it is documented.

The paper “Visualizing balances of compositional data” presents an R-package to visualize balances of compositional, high dimensional, data. The original visualization of balances was in the form of a dendrogram which represented in one figure the sequential binary partition, the mean of the corresponding balances, the variance corresponding to each balance, and a boxplot of each balance, if necessary separated by subgroup in several box-plots. A dendrogram is clearly difficult to visualize if a composition has several hundreds or even thousands of components. The tool described in the paper is thus a need in the field of compositional data analysis. The visualization strategy used can be summarized in a colloquial way by saying: “separate the dendrogram in two figures, one corresponding to the partition and grouping of parts, the other to the boxplots”. It is an interesting complement to the balance dendrogram, but we do not think it is an alternative. For high dimensional compositions it would be interesting to have three or more pictures, one corresponding to the partition, one to the box-plots, and the others to particular branches of the dendrogram, possibly summarizing groups of components by their common characteristic, if any. It would look like the three pictures depicted in Figure 1 of the paper.

One of the issues presented is the representation of the box-plots on a common scale, which is not completely new. It has been previously used at least in two papers, namely Lovell et al. (2013)[1] and Pawlowsky et al. (2015)[2]. Nevertheless, the additional features of visualizing the proportion of explained variance by the thickness of the segment covering the range and by the inclusion of the data as dots can be helpful in understanding the role of each balance. At the same time, in certain circumstances, like with high dimensional data observed in a large sample, it will probably be still difficult to visualize the mentioned features. In such a case, it might be interesting to represent the first principal balances, something already considered in the paper.

The most useful features of compositional dendrograms are (a) the visualization of the decomposition of the total variance into contributions of each balance for one or more populations in the sample; (b) the comparison of mean values between populations; (c) identifying groups of parts as they participate in balances defined by the partition. The proposed software solves point (b) efficiently by comparing box-plots in a homogenized scale. Point (c) can be supplemented including a tool able to enumerate parts in the numerator and denominator of the balance. This is important in high dimensional compositions where the labels in the partition panel are not identifiable (example Figure 2, left panel). Point (a) is more deficiently covered. A partial solution is suggested in section 3) of the paper by colouring some segments in the partition panel according to the value of the variance. However, it seems useful to have a tool allowing alphanumerical output of ordered variances or cuts of the partition tree. For instance, if somebody is looking for linear associations of groups of taxa (Egozcue, Pawlowsky-Glahn and Gloor 2018[3]) detecting balances which variance is smaller than a certain threshold, it is useful to visualize those balances by colour in the partition panel, but also to get a list of the parts involved in such balances.

We expect that the presented software, modified as suggested, becomes a useful tool for the analysis of high dimensional compositional data.

Minor issues that need to be revised in the paper are the following:

1. Introduction

   a) Nowadays, compositions are not defined as vectors of positive measurements that sum to a given total, arbitrary or not, but as representatives of equivalence classes in the positive orthant of D-dimensional real space (Barceló-Vidal et al. 2001[4]; Barceló-Vidal and Martín-Fernández 2016[5]).

   b) The usual representative of the equivalence classes is bounded between 0 and a given constant k. This defines a subset of real space which is not a subspace. Moreover, this set has a Euclidean space structure given by the operations that define the “Aitchison geometry” (Pawlowsky-Glahn and Egozcue, 2001[6]). Thus, it is not adequate to say that compositional data exist in a non-Euclidean space.

   c) The description of the additive (alr) and centred (clr) log-ratio transformations as “simple” is misleading. The alr defines coordinates in an oblique basis in the above mentioned Aitchison geometry, while the clr leads to coordinates in a generating system and changes with subcompositions. Thus, results of clr components are not subcompositionally coherent. Therefore, interpretation of results is very difficult, as users tend to interpret results in terms of the component in the numerator only, not taking into account the role of the denominator. Furthermore, in many cases results obtained with the alr are not permutation invariant, something that needs to be checked for each method. One of the most striking cases is e.g. regression, where the equation itself is permutation invariant, but not so the goodness-of-fit criteria.

d) The suggested alternative analysis in terms of simple log-ratios is also not simple at all, as they lead to the most general models, i.e. general log-contrast. The exponents involved in such a log-contrast are in general different for each part of the composition.

2. Methods

The equation given for computing balances is not correctly described. The term $|i_p|$ does not describe the norm or length of the sub-composition, but the number of parts in the sub-composition.

3. Use cases

a) The optional line width of the range of box-plots to illustrate the proportion of variance explained by a balance is not really informative in the case of high-dimensional data. In the low-dimensional case we think the dendrogram is more informative, as one can recognise easily if the balance that explains the largest proportion of variance corresponds to the first steps of the partition, involving thus a large number of parts or, on the contrary, involves only a small number of parts. This deficiency can be mitigated by colouring bars in the partition panel. For instance, plotting in red the lines corresponding to a given probability quantile of large variances and in blue for a probability quantile range of small variances.

b) Figure 2 shows two limitations of the proposed visualisation.
i) In the left panel it is clear which taxa are involved in each balance, but not which taxa are in the numerator, and which of those are in the denominator. Perhaps a good alternative would be to use different colours for each group, or to reorder the taxa in such a way that those in the numerator are always in the left hand side and a vertical bar indicates the dividing point.

ii) A numeration of the balances according to the larger (smaller) explained variance would be helpful in recognising rapidly which balance is the most (the less) informative in this sense.

4. Summary
a) It is not always true that “log-ratio transformations offer a way to transform the data into an unbounded space where the analyst can apply conventional statistical method”. For this to be true you need at least the transformation to be an isometry. For example, the alr is not an isometry, and thus conventional statistical methods should not be applied blindly.

b) Balances is a particular case of isometric log-ratio transformation. Another example is given by general log-contrasts obtained as coordinates in compositional principal component analysis.

References

1. Lovell D, Pawlowsky-Glahn V, Egozcue JJ: Have you got things in proportion? A practical strategy for exploring association in high-dimensional compositions. *Proceedings of the 5th International Workshop on Compositional Data Analysis, CoDaWork*. 2013.
2. Pawlowsky-Glahn V, Monreal-Pawlowsky T, Egozcue JJ: Representation of species composition. *Proceedings of the 6th International Workshop on Compositional Data Analysis, CoDaWork*. 2015.
3. Egozcue J: Linear Association in Compositional Data Analysis. *Austrian Journal of Statistics*. 2018; **47** (1). Publisher Full Text
4. Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V: Mathematical foundations of

compositional data analysis. *Proceedings of IAMG'01 - The 7th Annual Meeting of the International Association for Mathematical Geology*. 2001.
5. Barcelo-Vidal C, Martín-Fernández J: The Mathematics of Compositional Analysis. *Austrian Journal of Statistics*. 2016; **45** (4). Publisher Full Text
6. Pawlowsky-Glahn V, Egozcue JJ: Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*. 2001. 384-398

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Statistics - compositional data analysis

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.