Universitat Pompeu Fabra Barcelona

Department of Economics and Business

Economics Working Paper Series
Working Paper No. 1857

The chi-square standardization, combined with Box-Cox transformation, is a valid alternative to transformin to logrations in compositional data analysis

Michael Greenacre

January 2023

The chi-square standardization, combined with Box-Cox transformation, is a valid alternative to transforming to logratios in compositional data analysis

Michael Greenacre[1,1*]

[1*]Department of Economics and Business, and Barcelona School of Management, Universitat Pompeu Fabra, Ramon Trias Fargas, 25–27, Barcelona, 08272, Spain.

Corresponding author(s). E-mail(s): michael.greenacre@gmail.com;

Abstract

The approach to analysing compositional data with a fixed sum constraint has been dominated by the use of logratio transformations, to ensure exact subcompositional coherence and, in some situations, exact isometry as well. A problem with this approach is that data zeros, found in most applications, have to be replaced to permit the logarithmic transformation. A simpler approach is to use the chi-square standardization that is inherent in correspondence analysis. Combined with the Box-Cox power transformation, this standardization defines chi-square distances that tend to logratio distances for strictly positive data as the power parameter tends to zero, and can thus be considered equivalent to transforming to logratios. For data with zeros, a value of the power can be identified that brings the chi-square standardization as close as possible to transforming by logratios, without having to substitute the zeros. Especially in the field of high-dimensional "omics" data, this alternative presents such a high level of coherence and isometry as to be a valid, and much simpler, approach to the analysis of compositional data.

Keywords: Box-Cox transformation, chi-square distance, correspondence analysis, isometry, logratios, Procrustes analysis, subcompositional coherence

1 Introduction

Compositional data are non-negative data with a fixed sum constraint, usually 1 or 100%. Such data are found in many fields, notably biochemistry, geochemistry, ecology, linguistics, as well as all the "omics" fields of genomics, proteomics, microbiomics, and metabolomics, to name but a few. In most cases, such data are observed as counts, abundances or intensities, where the totals in the samples, usually the rows of the data matrix, are irrelevant. Consequently, the sample values can be divided by their respective totals to give compositional vectors, with sums equal to 1. This operation of dividing by the total is called closure, or normalization.

It has long been recognized that such data need special statistical treatment, since the vectors of sample data would change by different scalar multiples if some compositional parts were excluded and the data re-normalized with respect to their new totals, giving so-called subcompositions. In reality, in almost all applications the observed compositions are themselves subcompositions of a larger set of potentially observable parts, with proportional values out of a total of 1 that would change if an extended set of parts were observed. To deal with this problem in compositional data, the use of ratios of parts as the basis for statistical analysis was proposed by John Aitchison [1, 2], who laid the foundation for a field of statistics often referred to as compositional data analysis, or CoDA. Ratios are invariant with respect to deleting parts from or adding parts to a composition, and are thus described as being (subcompositinally) coherent. Because ratios are compared multiplicatively, logarithmic transforms of the ratios, called logratios, have become the preferred transformation for CoDA. Thus, the idea has been to transform to logratios and then continue with regular statistical methods applicable to interval-scale data, but taking appropriate care in the interpretation of the subsequent results in terms of the original compositional parts.

The logratio approach has not been without controversy, especially since the introduction of the isometric logratio transformation [3]. Several authors have insisted that the logratio transformation has to be isometric, a requirement that was never in Aitchison's vision of CoDA [4]. Isometry means that the geometry of the transformed data, which can be thought of as the multivariate distances between all pairs of samples, must be exactly the same as the logratio distances, which are the Euclidean distances between samples based on all their pairwise logratios. Criticisms of this insistence on the type of "permitted" data transformation, which severely limits practitioners in their choice, can be found in several papers, for example [5], [6], [7], [4].

To counter the rigid approach to CoDA, bordering on dogmatic, a more flexible, pragmatic approach to CoDA has been suggested by, for example, [4, 8–12], where simple pairwise logratios are used, for example the additive logratio transformation first proposed by [2]. Here the CoDA principles of coherence and isometry are regarded as ideals and transformations are allowed that come very close to these ideals. The idea is similar to the approach in statistics where an ideal mathematical model for the data is assumed, for example the normal distribution, and then methods are validly applied to data that are found to be closely following the model, for example approximately normally distributed.

The challenge is thus to find alternative transformations of compositional data that approximate closely the logratio approach. To support such an approach, quantitative measures are needed to compare the properties of a transformed data set to that of a logratio-transformed one, just like the distribution of observed data are quantitatively compared to the ideal of a normal distribution.

The two ideal properties considered here are coherence and isometry. A possible measure of coherence (or, conversely, lack of coherence, or incoherence) has already been proposed by [13]. Similarly, if isometry is deemed to be important, a measure of isometry (or, conversely, lack of isometry, or anisometry) can also be defined [4]. This allows practitioners to judge whether alternative transformations are quasi-coherent, or quasi-isometric, which means that they are close enough to coherence and isometry, for all practical purposes.

The objective of this paper is to demonstrate the circumstances in which the chi-square standardization of compositional data, which is intrinsic to correspondence analysis [14–16], can be used as a valid alternative to the logratio transformation, especially when the standardization is preceded by the Box-Cox power transformation [17]. This result is not coincidental, for it is known that chi-square distances on Box-Cox transformed compositions tend to logratio distances as the power parameter tends to zero, where the Box-Cox transformation tends to the log-transformation [18, 19]. This close theoretical connection holds for strictly positive data, and clearly not for data that include zeros where the logarithm in the limit can not be attained. However, in the presence of zeros, it turns out that a power transformation can be identified that is optimal in approximating the logratio distances, and the validity of the resulting chi-square transformation can be checked using the coherence and isometry measures mentioned above. This approach has the important advantage that zeros do not have to be substituted, as is the case for the logratio approach.

To illustrate this more flexible approach, a large "wide" compositional data set is considered from a study of the intramuscular fat of 89 rabbits, where the compositional parts are 3937 microbial genes [12, 20]. This is a typical data set in the burgeoning field of "omics" research: genomics, microbiomics, metabolomics, proteomics, etc. The original data set is of counts, so these data are first expressed as compositions by dividing each sample's counts by its total count. This data set has only positive counts, so that these closed (or normalized) values are strictly positive. Hence, it is a good example to demonstrate the chi-square transformation's approximation to the logratio transformation. And then, by simulating a large percentage of small counts to be zeros, the behaviour of the chi-square approach, which can handle zero values without any problem, can be further studied. In this case, the logratio transformation is not possible unless the zeros are replaced with imputed positive values.

2 Methods

2.1 Measuring coherence and isometry

Coherence is a property of the compositional parts, usually the columns of the data matrix, whereas isometry is a property of the sampling units, usually the rows. Measures of coherence and isometry can be unified under the umbrella of statistical distance in both cases.

Isometry means "the same metric", that is the same distance structure or geometric configuration in multivariate space. Hence, consider the ideal logratio distances between all the samples, on the one hand, and the distances between the same samples based on any other transformation of the compositional data, on the other hand. Even the identity transformation can be considered, that is no transformation at all. There are various measures of concordance of these two sets of distances in the literature on multidimensional scaling [21]. For the present purpose we prefer a measure of the similarity of the multivariate configurations that are engendered by these two sets of distances, called the Procrustes correlation [22], also proposed by [23] in the context of variable selection. One advantage of a correlation measure is that it has a familiar scale, so that for alternative transformations Procrustes correlations very close to 1 would be desirable.

Perfect (subcompositional) coherence of the compositional parts means that the results for subsets of parts forming subcompositions, after renormalization, remain exactly the same as when they formed the "full" composition. This principle is important because the set of observed parts is almost always a subset of a larger set of parts, and so the results should not change. Coherence can have a different meaning depending on whether the statistical analysis is unsupervised, that is identifying patterns in the data, or supervised, that is using compositional variables to explain one or more response variables. In an unsupervised context, coherence would be when the relationships between parts remain the same in subcompositions, where these relationships can be quantified by the multivariate configuration of the parts, similar to the distance-based approach to measuring isometry of the samples, described previously. In a supervised context, coherence would be when the predictive models estimated using certain compositional parts remain the same irrespective of the parts being in a subcomposition or the original composition.

To assess alternative transformations that are not coherent, subcompositions of parts need to be chosen, and results compared with those in the full composition. There are clearly very many possibilities to choose subsets of parts and create subcompositions. Random subsets of parts can be selected, or it may be that subcompositions in particular applied contexts tend to include the more frequent parts than the less frequent ones. For example, in microbiome research, the more frequent bacteria would always be present across different studies, whereas they would vary in the rarer bacteria they include.

Whichever way a subcomposition is chosen, the multivariate configuration of the parts in the subcomposition can be compared to the configuration of the parts in the full composition, and their similarity can again measured by the Procrustes correlation. Notice that in this case, the comparison is not between the logratio configuration and the alternative one based on a different transformation, but between configurations both based on the same alternative transformation. For the logratio transformation, which is coherent, there would be no difference in the configurations and the Procrustes correlation would be exactly equal to 1. For alternative transformations, high correlations consistently near 1, whatever the subcomposition selected, would indicate that the transformation is very close to being coherent, in other words that the renormalizing of the subcompositions has minimal effect on the relationships between the parts. In the supervised context emphasis would not be focused on the relationships between the parts themselves but rather on their performance in explaining response variables.

There is an obvious relationship between the two concepts of coherence (of the parts) and isometry (of the samples). Showing that a transformation is very close to being isometric, means that it is close to the logratio transformation. But the logratio transformation is perfectly coherent, which implies that alternative transformations close to being isometric must also be coherent.

2.2 Theoretical link between correspondence analysis and logratio analysis

It has been known for some time that the principal component analysis (PCA) of all pairwise logratios, called logratio analysis (LRA) [8], and correspondence analysis (CA) [16] have a theoretical connection via the Box-Cox power transformation:

$$f(x) = \begin{cases} \frac{1}{\alpha}\left(x^{\alpha} - 1\right) & \text{if } \alpha > 0 \\ \log(x) & \text{if } \alpha = 0 \end{cases}$$

The connection is that if the compositional data in the matrix $\mathbf{X}$ are power-transformed as above, where the -1 can be omitted thanks to the double-centring in CA, then the CA of the power-transformed data converges in the limit to LRA as α tends to 0. This result only applies to strictly positive data, although we will soon treat the case when zeros are present. The multiplicative rescaling by $1/\alpha$ is important to correct for the decreasing variance as α decreases to 0. Notice that LRA has a weighted and unweighted version [8], and for the purposes of the present study the unweighted version is used, as in [24].

The important implication of the above result is that, for strictly positive data, one can perform a CA with a very strong power transformation (i.e., tiny power α), and come as close as required to performing LRA based on all the pairwise logratios. Equivalently, the chi-square distances on power-transformed compositional data are converging to the logratio distances as the power tends to 0. In other words, the chi-square standardization on power-transformed compositional data is converging to an isometric transformation. And, furthermore, in coherence terms, the chi-square transformation is increasingly coherent as the power decreases and tends to exact coherence.

The chi-square standardization is very simple: each part x_{ij} in a compositional data matrix $\mathbf{X}$ is divided by the square root of its respective mean value $\sqrt{\bar{x}_j}$: $x_{ij}/\sqrt{\bar{x}_j}$. Since all the means $\bar{x}_j$ are less than 1, this has the effect of stretching out the regular simplex of the compositions into an irregular simplex, where the vertices of the simplex corresponding to rarer parts are stretched out more than those corresponding to abundant parts. Since the logratio transformation takes the compositional data into real space, it is noteworthy that the power transformation combined with the chi-square standardization is stretching the simplex, and consequently the samples in the simplex, to exactly the positions of the samples in logratio space as the power of the transformation tends to 0.

2.3 The problem of data zeros

The problem of zeros has been called the "Achilles heel" of compositional data analysis [25], since data have to be strictly positive to be able to compute logratios. Because zeros are usually present in compositional data, and often in large quantities, a number of zero replacement strategies have been developed — see [26] for a review. The presence of many zeros can cause problems in the analysis [27]. The CA approach combined with a power transformation, provides an approach to avoid zero replacement, but it is clear that, as the power of the Box-Cox transformation decreases when applied to data that have many zeros, there will be a problem as the transformation approaches the logarithm. In the present approach, for data with zeros, the power of the Box-Cox transformation will be identified that leads to chi-square standardized data having maximum isometry with the logratio-transformed data, where zeros will have to be replaced in the latter case to enable its computation.

3 Results

3.1 Strictly positive compositional data

The compositional "Rabbits" compositional data set (89 samples, 3937 genes) has strictly positive values, which is rather atypical, but it is useful here to illustrate the good properties of the chi-square standardization combined with power transformations. The next subsection treats the case with data zeros.

Logratio analysis (LRA, the "ideal") is first performed on the data and the configuration of the 89 samples established in 88-dimensional multivariate space, one less than the number of samples for this wide data set. Then CA is performed on the original untransformed compositions, and then successively on versions of the data transformed by powers α descending from 1 to almost 0, where "almost" is $\alpha = 0.0001$.

Figure 1A shows a plot of the Procrustes correlations between the logratio geometry of the 89 samples and corresponding power-transformed chi-square geometry, showing the convergence to 1 as α tends to 0. In each case along the curve the 88-dimensional logratio geometry is compared to the 88-dimensional chi-square geometry. Values indicated are for square root, fourth root and ten thousandth root ($\alpha = 0.0001$) transformations.

Figure 1B plots the $89 \times 88/2 = 3916$ logratio distances between pairs of sample points in the full 88-dimensional space against the corresponding distances for the $\alpha = 0.0001$ case, where the isometry is further shown.

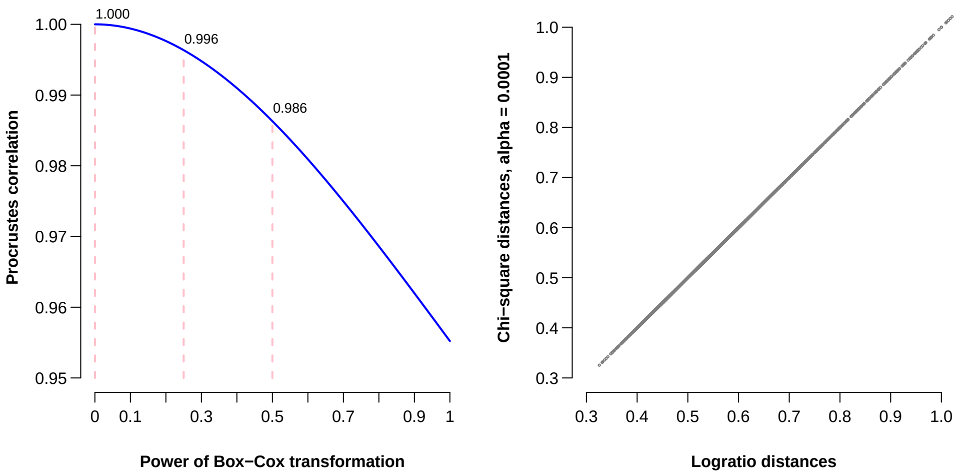

Fig. 1 A. The Procrustes correlations, measuring proximity to isometry between the exact logratio geometry and the geometry of correspondence analysis on compositional data subjected to Box-Cox transformation with different powers. B. For the power equal to 0.0001, the chi-square distances are practically identical to the logratio distances. In the limit as the power tends to 0, they are exactly equal.

To further illustrate the theoretical convergence of these geometries, Figure 2 shows the two-dimensional results of the CA for α =1 (original CA), 0.5 (CA on square-root data), 0.0001 (CA on ten thousandth-root data), and finally LRA. Figures 2C and 2D are identical in their coordinates up to four decimals — the maximum absolute difference over all coordinate values is 0.00006. The three groups of points correspond to three different laboratories which performed the testing, where it can be seen that one was quite different from the other two.

3.2 Compositional data with zeros

To simulate a situation where zeros are present, using the same rabbits data set, a count of 20 was temporarily regarded as the detection limit and all values less than 20 in the original matrix of microbial gene counts were set

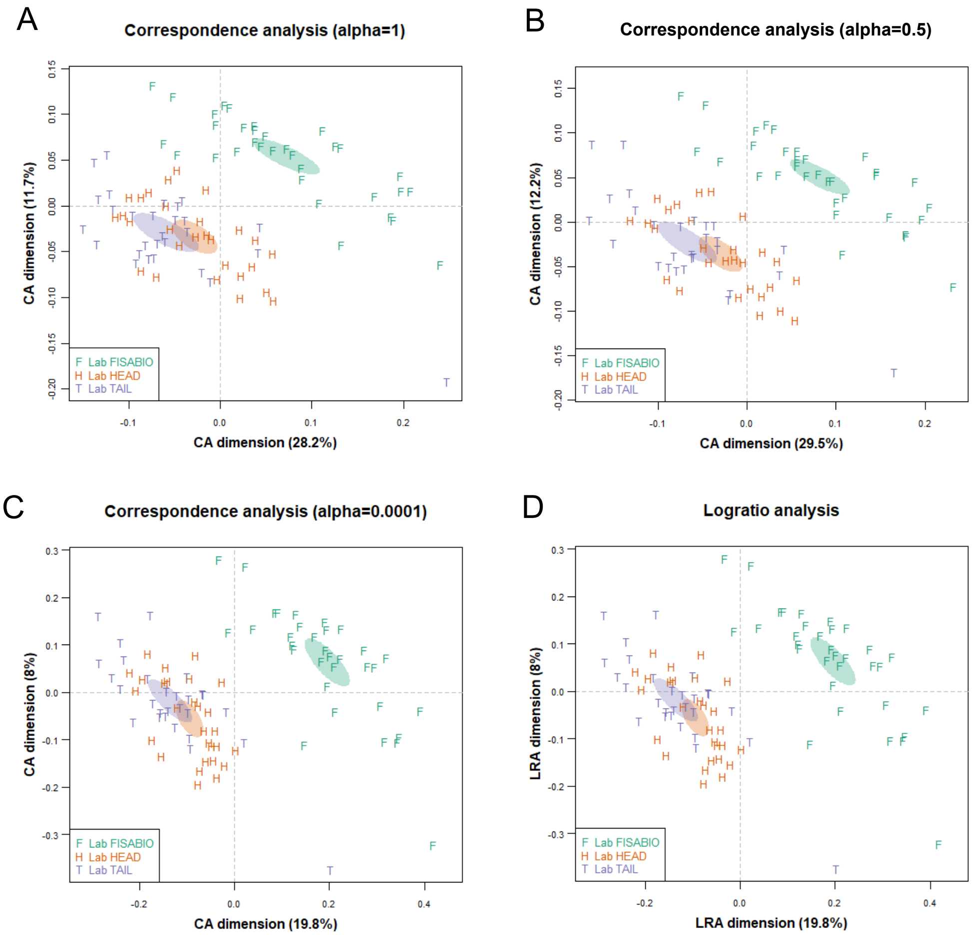

Fig. 2 Using the rabbits data set, three correspondence analyses of Box-Cox transformed data with decreasing powers, and logratio analysis as the limit solution. A. The regular CA with power 1 (no transformation) B. CA with power 0.5 (square root). C. CA with power 0.0001. D. Logratio analysis (LRA). C and D are identical in their coordinate values to the fourth decimal. The ellipses are 95% bootstrap confidence regions for the means of the three groups of points.

to 0. This resulted in a data matrix with 25035 zeros, which is 7.1% of the 89×3937 data matrix. This matrix was then normalized to compositions, and analysed in a similar way as before. In order to compare the CA results with logratio ones, the zeros were imputed using the function `cmultRepl` in the `zCompositions` R package [28], which is one of the popular ways of zero replacement. The chi-square distance structure of power-transformed data (with zeros) was compared to the logratio distance structure of the data matrix with zeros replaced.

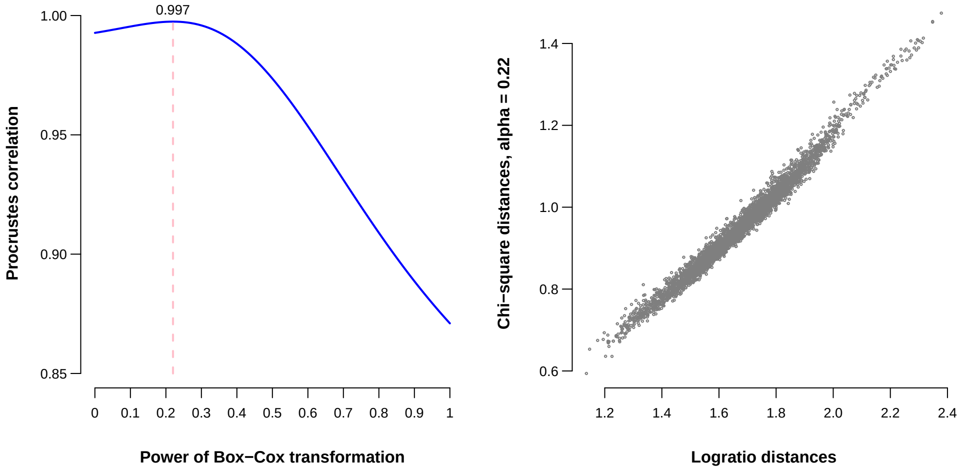

Fig. 3 A. The Procrustes correlations, measuring proximity to isometry between the exact logratio geometry and the geometry of correspondence analysis on compositional data subjected to Box-Cox transformation with different powers. The correlation is at an optimum value of 0.997 for a power of 0.22. B. In the respective 88-dimensional spaces, the chi-square distances are very similar to the logratio distances.

Figure 3 shows the results, where now the CA obviously cannot reproduce exactly the logratio structure, but it can come very close depending on the power transformation selected. Figure 3A shows that, as the power decreases, an optimal value of the Procrustes correlation is reached, equal to 0.997, at $\alpha = 0.22$, which is close to a fourth-root transformation. Hence the level of isometry is almost perfect, and the concordance of the chi-square and logratio distances can be seen in Figure 3B. Since we know the true values of the constructed zeros in the data, the comparison could also be made directly between the CA applied to the data matrix with zeros and the LRA on the original data. The highest Procrustes correlation in this case is 0.990, only slightly less than 0.997 above, but for a power of $\alpha = 0.59$, which is close to a square root transform. All these results point again to the successful use this type of transformation of compositional data in the context of unsupervised learning, and without the need for zero replacement.

3.3 Power transformations for supervised learning

When it comes to compositional variables serving as predictors of a continuous or categorical response, it will be essential to perform some type of variable selection, especially in the present wide example where the number of variables considerably exceeds the number of samples. In this case the issue of isometry is no longer relevant but coherence certainly is. For example, suppose a small subset of parts of a composition is identified as good predictors of a response, how would the results change if only a subcomposition had been observed, with different compositional values?

Using logratios, because of their coherence, a result for any subcomposition would be identical if any number of genes were eliminated (or added) to the data set and the data reclosed to sum to 1. The level of incoherence of the chi-square approach can be investigated by taking random subcompositions of different sizes to see how model estimates in logistic regression are affected. From Figure 3 the optimal Box-Cox transformation is for power $\alpha = 0.22$, so we apply this transformation, without any need to apply the chi-square normalization since it would only affect the scale of the predictors. Likewise, the compositional data can be transformed simply as $x_{ij}^{0.22}$, without the need for the division by 0.22, which affects the scale, nor the subtraction of 1/0.22, since this only affects the constant of the regression. Standardized regression coefficients, however, should be computed for the subsequent comparisons on subcompositions. A model with two power-transformed variables, microbial genes MG2249 and MG3907, was found to be highly significant ($p = 0.0001$ and $p = 0.001$ respectively), with estimated model

$$\log\left(\frac{p}{1-p}\right) = 0.116 + 1.077\,\mathrm{MG2249}^{0.22} + 1.023\,\mathrm{MG3907}^{0.22} \qquad (1)$$

The acid test of coherence is how much the regression coefficients in (1) change when subcompositions are chosen. Random subcompositions of 10%, 20%, etc., up to 90% of the microbial genes were taken, 100 in each case, the only condition being that MG2249 and MG3907 were always included. Figure 4 plots the dispersions of the estimated standardized regression coefficients in all the subcompositions in the form of boxplots.

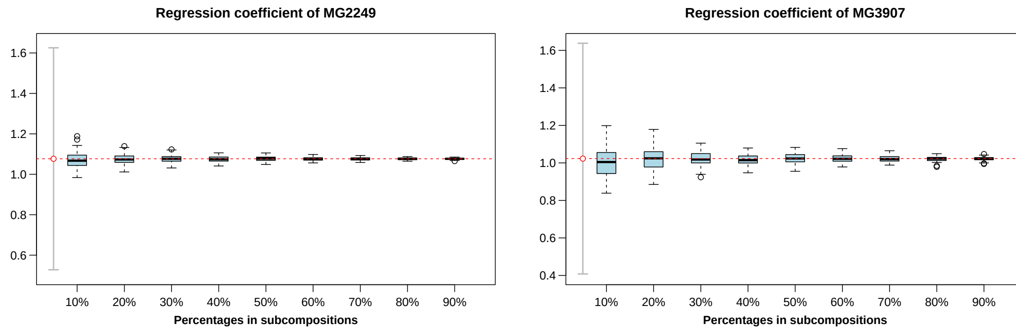

Fig. 4 Boxplots of the standardized regression coefficients for each predictor, computed for sets of 100 random subcompositions of different sizes, from 10% to 90%. The horizontal red lines show the values of the two coefficients, 1.077 and 1.023, which were estimated in the full composition. The vertical grey bars on the left of the plots show the 95% confidence intervals for the original estimates, using values in the full composition.

The models applied to the simulated subcompositions are all giving coefficient estimates distributed closely around the respective values estimated in the full composition. As expected, the smaller subcompositions have estimated coefficients more variable than the larger ones. But the variability of these coefficients should be judged against the margins of error of the estimates in the full composition. The standard errors are $\mathrm{SE}_{MG2249} = 0.280$ and $\mathrm{SE}_{MG3907} = 0.314$, implying confidence intervals, shown in Figure 4, which are much wider than the narrow dispersions of the subcompositional estimates, for all sizes of subcompositions.

4 Conclusion

An alternative pipeline is possible for analysing compositional data, using the chi-square standardization preceded by a Box-Cox power transformation. The power can be deduced from a study of the proximity to isometry, as illustrated in Figure 3, which is useful for unsupervised learning. The same approach can be used for supervised learning, where the closeness to coherence can be evaluated by observing the effect of taking subcompositions on the model estimates. Overall, this approach, which uses simple measures to assess its validity, presents a simpler and more easily interpretable alternative to the logratio transformation in compositional data analysis, without any need for zero replacement.

References

[1] Aitchison, J.: The statistical analysis of compositional data (with discussion). J R Stat Soc Ser B **44**, 139–77 (1982)

[2] Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (1986)

[3] Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math Geol **37** (2005)

[4] Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., Quinn, T.: Aitchison's compositional data analysis 40 years on: a reappraisal. *arXiv*. arXiv (2022). https://doi.org/10.48550/arXiv.2201.05197

[5] Scealy, J.L., Walsh, A.H.: Colours and cocktails: Compositional data analysis. Aust N Z J Stat **56(2)**, 145–169 (2014)

[6] Greenacre, M., Grunsky, E., Bacon-Shone, J.: A comparison of amalgamation and isometric logratios in compositional data analysis. Comput Geosc **148**, 104621 (2020)

[7] Smithson, M., Broomell, S.B.: Compositional data analysis tutorial. Psych Meth **27** (2022)

[8] Greenacre, M.: Compositional Data Analysis in Practice. Chapman & Hall / CRC Press, Boca Raton, Florida (2018)

[9] Greenacre, M.: Variable selection in compositional data analysis using pairwise logratios. Math Geosc **51**, 649–82 (2019)

[10] Greenacre, M.: Discussion of "Compositional data: the sample space and its structure", by Egozcue and Pawlowsky-Glahn. TEST **28**, 20–24 (2019)

[11] Greenacre, M.: Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. Appl Comput Geosc **5**, 100017 (2020)

[12] Greenacre, M., Mártinez-Álvaro, M., Blasco, A.: Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. Front Microbiol **12**, 2625 (2021). https://doi.org/10.3389/fmicb.2021.727398

[13] Greenacre, M.: Measuring subcompositional incoherence. Math Geosc **43**, 681–93 (2011)

[14] Benzécri, J.-P.: L'Analyse des Données. Tôme II: L'Analyse des Correspondances. Dunod, Paris (1973)

[15] Greenacre, M.: Theory and Applications of Correspondence Analysis. Academic Press, London (1984)

[16] Greenacre, M.: Correspondence Analysis in Practice (3rd Edition). Chapman & Hall / CRC Press, Boca Raton, Florida (2016)

[17] Box, G.E.P., Cox, D.R.: An analysis of transformations. J R Stat Soc Ser B **26**, 211–52 (1964)

[18] Greenacre, M.: Power transformations in correspondence analysis. Comp Stat Data Anal **53**, 3107–16 (2009)

[19] Greenacre, M.: Log-ratio analysis is a limiting case of correspondence analysis. Math Geosc **42**, 129–34 (2010)

[20] Martínez-Álvaro, M., Auffret, M.D., Duthie, C.-A., Dewhurst, R., Cleveland, M., Watson, M., Roehe, R.: Bovine host genome acts on specific metabolism, communication and genetic processes of rumen microbes host-genomically linked to methane emissions. Commun Biol **5**, 350 (2022)

[21] Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling: Theory and Applications, 2nd Edition. Springer, New York (2010)

[22] Gower, J., Dijksterhuis, G.B.: Procrustes Problems. Oxford University Press, New York (2004)

[23] Krzanowski, W.: Selection of variables to preserve multivariate data structure, using principal components. J R Stat Soc Ser C (Appl Stat) **36**, 22–33 (1987)

[24] Aitchison, J., Greenacre, M.: Biplots of compositional data. J R Stat Soc Ser C (Appl Stat) **51**, 375–92 (2002)

[25] Greenacre, M.: Compositional data analysis. Annu Rev Stat Appl **8**, 271–99 (2021)

[26] Lubbe, S., Filzmoser, P., Templ, M.: Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chem Intell Lab Syst **210**, 104248 (2021)

[27] te Beest, D.E., Nijhuis, E.H., Möhlmann, T.W.R., ter Braak, C.J.F.: Log-ratio analysis of microbiome data with many zeroes is library size dependent. Molecular Ecology Resources **21**(6), 1866–1874 (2021) https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13391. https://doi.org/10.1111/1755-0998.13391

[28] Palarea-Albaladejo, J., Martin-Fernandez, J.: zcompositions – r package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems **143**, 85–96 (2015)