REGULAR ARTICLE

The chiPower transformation: a valid alternative to logratio transformations in compositional data analysis

Michael Greenacre[1] (iD)

Abstract

The approach to analysing compositional data 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 allow the logarithmic transformation. An alternative new approach, called the 'chiPower' transformation, which allows data zeros, is to combine the standardization inherent in the chi-square distance in correspondence analysis, with the essential elements of the Box-Cox power transformation. The chiPower transformation is justified because it defines between-sample distances that tend to logratio distances for strictly positive data as the power parameter tends to zero, and are then equivalent to transforming to logratios. For data with zeros, a value of the power can be identified that brings the chiPower transformation as close as possible to a logratio transformation, without having to substitute the zeros. Especially in the area of high-dimensional data, this alternative approach can present such a high level of coherence and isometry as to be a valid approach to the analysis of compositional data. Furthermore, in a supervised learning context, if the compositional variables serve as predictors of a response in a modelling framework, for example generalized linear models, then the power can be used as a tuning parameter in optimizing the accuracy of prediction through cross-validation. The chiPower-transformed variables have a straightforward interpretation, since they are identified with single compositional parts, not ratios.

Keywords Box-Cox transformation · Chi-square distance · Correspondence analysis · Isometry · Logratios · Procrustes analysis · Subcompositional coherence · Tuning parameter

Mathematics Subject Classification 6204 · 62H25 · 62H30 · 62P10

✉ Michael Greenacre
michael.greenacre@gmail.com

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

1 Introduction

Compositional data are non-negative data carrying relative, rather than absolute, information. Often these data have a constant-sum constraint on each sample's set of values, for example, proportions summing to 1 or percentages summing to 100%. Such data are found in many fields, notably biochemistry, geochemistry, ecology, linguistics, as well as all the "omics" fields of genomics, microbiomics, transcriptomics, metabolomics, etc.. In most cases, such data are originally observed as counts, abundances or intensities, where the totals in the samples, usually the row totals of the original data matrix, are irrelevant. Consequently, the sample values can be divided by their respective totals to give vectors, called compositions, with sums equal to 1. This operation of dividing by the total is called closing, sometimes referred to as normalization.

It has long been recognized that such data need special statistical treatment, since the values in the compositions would change if some compositional parts were excluded and the data re-closed 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 that would change if an extended set of parts were observed. For example, in geochemistry, some studies use only major oxide elements, others treat trace elements, while others treat the full lithogeochemical spectrum of major, minor, trace and rare elements. Thus, in this last case, the compositional proportions of the major oxides would be different than those when the major oxides were studied alone. Similarly, in the study of fatty acid compositions in biochemistry, the set of fatty acids identified and analysed in any study is always a subcomposition of a much larger set, not only due to the focus of the research but also on the sophistication of the measuring instruments (e.g., gas chromatographs). The same is true for microbiome studies, for example, where the set of bacteria is never the full set of possibilities. One of the few contexts where a full composition is observed is in daily time use in behavourial studies, where all activities are recorded over a full 24-hour period—here the time budget is compositionally complete since no more time can be added to a day.

To deal with this dependency of compositional data on the particular set of parts that are included, the use of ratios of parts as the basis for statistical analysis was proposed by John Aitchison (Aitchison 1982, 1986), 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 subcompositionally coherent (simply referred to here as coherent), whereas any analysis of the original compositional data is incoherent. But ratios are awkward to handle statistically – their distributions are generally skewed and there is an asymmetry between the numerator and the denominator so that, for example, the variance of A/B is not equal to the variance of B/A. The logarithmic transformation reduces the skewness, the variance of $\log(A/B)$ equals the variance of $\log(B/A)$, and either $\log(A/B)$ or $\log(B/A)$ can be used in linear modelling, since they are just a change of sign. Because of the logarithmic transform, additive changes in logratios are thus multiplicative changes in the ratios, as in logistic regression, for example, where a logratio, the log-odds, is modelled as an additive model of

explanatory variables, and additive effects back-transform to multiplicative effects on the odds.

Hence, logarithms of ratios, called logratios, have become the preferred transformation for those following the tradition of Aitchison, and once this transformation is made, regular statistical methods applicable to interval-scale data can continue as before. This approach is exemplified by Grunsky et al. (2024), who present a comprehensive workflow, called GeoCoDA, for using logratio transformations in both unsupervised and supervised learning in geochemistry, with accompanying R code. For an in-depth review and reappraisal of Aitchison's ideas and legacy in the 40 years since his 1982 JRSS discussion paper (Aitchison 1982), see Greenacre et al. (2023). Aitchison's 1982 paper and legacy is further discussed by Coenders et al. (2023).

Coherence is the main advantage of the logratio approach, but its main disadvantage is the problem of data zeros, as well as the interpretation of results involving logratios. Data zeros need to be replaced before logratios can be computed, and there have been many proposals to do so—for a review, see Lubbe et al. (2021). It may be that alternative transformations, with simpler interpretations and natural handling of data zeros, are close enough to this ideal property of coherence for all practical purposes. To quantify this "closeness" to coherence, a possible measure of incoherence has already been proposed by Greenacre (2011), using a concept from multidimensional scaling called stress. In the present paper, an alternative measure will be used based on the Procrustes correlation, a by-product of Procrustes analysis (see Appendix 2), since this will unify the treatment of coherence and another concept called isometry.

Whereas coherence is a property of the compositional parts, isometry is a property of the samples. If the logratio approach is taken as a favourable reference for CoDA, then the sample structure using an alternative transformation can be checked against the sample structure using the logratio transformation. Here the Procrustes correlation will again be used to measure closeness to isometry, by which is meant closeness to the logratio sample structure. This idea of using Procrustes analysis, inspired by Krzanowski (1987), has already been used for logratio variable selection by Greenacre (2019). Such diagnostic measures of similarity between part structures (coherence) and between sample structures (isometry) allow practitioners to judge whether simpler alternative transformations are close enough to coherence and isometry to allow valid statistical analysis. As mentioned before, the benefit of these alternative transformations will be that they are easier to interpret and also cope naturally with zeros in the data without the need for replacement or imputation.

The objective of this paper is to demonstrate how the intrinsic standardization in correspondence analysis (Benzécri 1973; Greenacre 1984, 2016), combined with a Box-Cox power transformation (Box and Cox 1964), can be successfully used as an alternative to logratio transformations. This alternative is underpinned by the fact that correspondence analysis's chi-square distances computed on Box-Cox transformed compositions tend to logratio distances as the power parameter tends to zero (Greenacre 2009, 2010). This close theoretical connection holds for strictly positive data, and clearly not for data that include zeros. However, in the presence of zeros, it turns out that a power transformation can be identified that is optimal in approximating logratio distances (i.e., as close to isometry as possible), and the validity of the resulting transformation can be additionally checked using the measure of coherence,

for comparing various subcompositions to the full composition. Because the proposed transformation combines the ideas of chi-square standardization (i.e., division of the part values by the square roots of their respective mean values) and power transformation, the new transformation is termed the *chiPower* transformation, to be defined explicitly in Sect. 2.3 below.

Moreover, if the compositional variables serve as independent variables in a supervised learning context, then the value of the power can be used as a tuning parameter to optimize prediction of the response variable. In this particular situation isometry is no longer important, but coherence is still an issue and will need to be investigated in each case.

To illustrate this alternative approach, a "wide" compositional data set is first considered with almost 4000 compositional parts (microbial genes) (Martínez-Álvaro et al. 2022; Greenacre et al. 2021). This is a typical data set in the burgeoning field of "omics" research: genomics, microbiomics, metabolomics, proteomics, etc. A second data matrix with much fewer parts but many more samples, i.e. a "narrow" but "long" data set, is considered where there is a categorical response to be predicted from the compositional variables. In the both applications the issue of data zeros is considered.

2 Material and methods

2.1 Data sets "Rabbits" and "Crohn"

To demonstrate the suitability of the chiPower approach proposed here, two data sets are considered:

1. Data set "Rabbits", used by Greenacre et al. (2021): a "wide" data set of counts of $J = 3937$ microbial genes observed on a sample of $I = 89$ rabbits. The advantage of this data set is that it has no zero values, so logratio transformations are valid on all the data. By simulating a large percentage of small counts to be zeros, the behaviour of the chiPower transformation, which can handle zero values without any problem, can be studied in comparison with the original logratio-transformed data.

2. Data set "Crohn", used by Calle et al. (2011) and available in the R package `coda4microbiome` (Calle et al. 2023). This is a "narrow" matrix of counts of bacterial species aggregated into $J = 48$ genera on $I = 975$ human samples. In addition, each sample has been classified as having the digestive ailment called Crohn's disease (662 samples) or not (313 samples). A curiosity of this data set is that it has been published in two different versions, with the same name: first, the original one with many data zeros (totalling 13474, i.e., 28.8% of the data set), in the original `selbal` R package—this version was analysed by Rivera-Pinto et al. (2018) (see Supplementary Material Section S1), who explicitly state that the "replacement of zeros by positive numbers is performed under the assumption that the observed zeros represent rounded zeros"; and second, a modified version published in the `coda4microbiome` package, with the same data set name `Crohn`, where the value of 1 has been added to all the counts, no doubt to avoid

the zero problem when computing logratios. As of the date of writing, no warning or explanation in the `coda4microbiome` package is given that the data set has been changed in this way, where $975 \times 48 = 46800$ counts have effectively been added to the original data set.

Nevertheless, the advantages of considering both versions of these data are two-fold. First, thanks to the large number of samples, a machine-learning approach can be applied to both versions for predicting the disease, where cross-validation can be implemented to estimate prediction accuracy; and second, the original data set, without zero replacement, can be used to show how well the chiPower approach, applied to the original data with zeros, compares to the logratio approach applied to the modified data set without zeros. Since other papers may have used the original `Crohn` data handling the zeros in different ways, the issue of the effect of these zero replacement strategies on the data variance is dealt with in Supplementary Material Section S1. The two versions of the data will be referred to as "the original Crohn data, with zeros" and "the modified Crohn data, without zeros".

2.2 Logratio transformations

Because the new chiPower transformation will be compared with the logratio approach, a short summary of the most relevant logratio transformations is given here (Aitchison 1986). Suppose $\mathbf{X}$ is an $I \times J$ samples-by-parts (closed) compositional data matrix, and $[x_1 \, x_2 \, \cdots \, x_J]$ is a general row of $\mathbf{X}$, that is, a J-part composition, where $\sum_{j=1}^{J} x_j = 1$. A specific row, for example the i-th row of $\mathbf{X}$, is denoted $[x_{i1} \, x_{i2} \, \cdots \, x_{iJ}]$.

The basic logratio transformation is the pairwise logratio transformation, denoted by PLR, of two parts j and j'

$$\text{PLR}(j, j') = \log(x_j/x_{j'}) \tag{1}$$

There are $J(J-1)/2$ unique PLRs, but only $J-1$ linearly independent ones are needed to generate all the others by linear combinations (Greenacre 2018). Thus, for I compositional samples, the $I \times J(J-1)/2$ matrix of PLRs has rank $J-1$.

A special case of PLRs are the additive logratios (ALRs), where the denominator part (also called the reference part, ref) is fixed.

$$\text{ALR}(j|\text{ref}) = \log(x_j/x_{\text{ref}}), \quad j = 1, \ldots, J, \; j \neq \text{ref} \tag{2}$$

There are J choices for the reference part, each of which gives $J-1$ ALRs. Any $I \times (J-1)$ data matrix of ALRs has rank $J-1$, and the choice of the reference part is determined either (i) by domain knowledge, or (ii) based on a statistical criterion such as the one that gives a transformed matrix closest to being isometric, or (iii) the one with lowest variance of its log-transform (Greenacre et al. 2021). In the last case, if the variance of $\log(x_{\text{ref}})$ is low, i.e. $\log(x_{\text{ref}})$ is nearly constant, then the ALR $\log(x_j/x_{\text{ref}}) = \log(x_j) - \log(x_{\text{ref}})$ is an approximate constant shift from the $\log(x_j)$ values themselves, in which case the ALRs can be more easily interpreted as close to the logarithm of the numerator parts.

The centered logratio (CLR) transformation is the log-transform of each part divided by the geometric mean of all the parts:

$$\mathrm{CLR}(j) = \log(x_j/(x_1 x_2 \cdots x_J)^{1/J}), \quad j = 1, \ldots, J$$

There is only one set of J CLRs and the j-th one is the average of all the PLRs $\log(x_j/x_{j'})$, for $j' = 1, 2, \ldots, J$, one of which, $\log(x_j/x_j)$, is zero. The $I \times J$ data matrix of CLRs also has rank $J - 1$, due to a linear relationship amongst them (they sum to 0). They are generally not used as variables representing the individual parts, although it is tempting to do so, but rather as representing all the PLRs by their differences: $\mathrm{PLR}(j, j') = \mathrm{CLR}(j) - \mathrm{CLR}(j')$. For example, to construct the sample logratio geometry, by which is meant the Euclidean distance structure of the samples with respect to all PLRs, it is not necessary to work with the $I \times J(J - 1)/2$ matrix of all PLRs, but just with the $I \times J$ matrix of CLRs (Aitchison and Greenacre 2002). The logratio distances between samples using the CLRs are identical to those using all the PLRs (Greenacre 2018, 2021).

Transforming by logratios takes the compositions inside the simplex out into real vector space, where regular interval-scale statistical analysis, both univariate, bivariate and multivariate, can be performed. The problem, however, is with data zeros, which need replacement before such transformations can be made.

2.3 The chiPower transformation: chi-square standardization, with preliminary power transformation

In correspondence analysis (CA), usually applied to a matrix of counts, the rows are first divided out by their totals to get so-called row profiles, synonymous with compositions—see, for example, Greenacre (2016). In CoDA terminology, CA automatically closes the rows, and—if the analysis is considered column-wise – it symmetrically closes the columns to get column profiles. In a closed compositional data matrix, the compositions in the rows are already profiles, so closing in CA does not change them. The row profiles in CA are weighted proportionally to the original marginal row totals, but in the case of compositions these marginal sums are all equal, so there is uniform weighting on the rows. Finally, distances between profiles in CA are chi-square distances, which are Euclidean distances after standardizing each compositional value x_j by dividing by the square root of its expected value, the column (part) mean $\bar{x}_j$: $x_j/\sqrt{\bar{x}_j}$—this is called the chi-square standardization (see, for example, Greenacre and Primicerio (2010), chapter 4). In the chiPower transformation, the x_j will be raised to power λ and closed, again giving compositions (a standard CoDA operation called "powering" by Aitchison (1986)), and then divided by the square roots of their respective column means. Notice that, since the divisors $\sqrt{\bar{x}_j}$ are less than 1, the chi-square standardization takes the compositions outside the regular simplex, into a larger irregular simplex.

For the present purpose, the Box-Cox power transformation is defined for positive x as:

$$f(x \mid \lambda) = \begin{cases} \frac{1}{\lambda}\left(x^{\lambda} - 1\right) & \text{if } \lambda > 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{3}$$

(negative values of λ are not considered, and only values $0 < \lambda \leq 1$ are of present interest). Whereas the limiting result implicit in (3), that is, $f(x \mid \lambda) \to \log(x)$ as $\lambda \to 0$, is only valid for $x > 0$, the power transformation itself for $\lambda > 0$ is valid for nonnegative x, i.e. $x \geq 0$, which is the way it will be used in the present approach. The scale factor $\frac{1}{\lambda}$ corrects for the shrinking variance in the transformed (positive) data as λ decreases. As shown in Appendix 1, if one wants the chiPower transformation to converge directly to the CLR transform, then a scale factor of $\sqrt{J}$ needs to be introduced and the -1 of the Box-Cox transform needs to be retained.

The chiPower transformation is defined algorithmically in the following steps, where the determination of the power λ will be dealt with after the definition.

The chiPower transformation

1. For a given λ, power transform the compositional data matrix $\mathbf{X}$ to obtain $\mathbf{X}_{[\lambda]} = \left[x_{ij}^{\lambda}\right]$, where $0 < \lambda \leq 1$ (so the possibility of no power transformation is included, when $\lambda = 1$).
2. Close the rows of $\mathbf{X}_{[\lambda]}$ to obtain another matrix of compositions, $\mathbf{Y}_{[\lambda]}$
3. Compute the vector of column means $\bar{\mathbf{y}}_{[\lambda]} = \left[\bar{y}_{[\lambda]1} \ \bar{y}_{[\lambda]2} \cdots \bar{y}_{[\lambda]J}\right]$ of $\mathbf{Y}_{[\lambda]}$.
4. Divide the columns of the closed $\mathbf{Y}_{[\lambda]}$ by the square roots of their respective column means (i.e., the chi-square standardization) and apply the Box-Cox style of transformation as follows:

$$z_{ij[\lambda]} = \frac{1}{\lambda}\left(\sqrt{J}\,\frac{y_{ij[\lambda]}}{\sqrt{\bar{y}_{[\lambda]j}}} - 1\right) \tag{4}$$

   The inclusion of the scale factor $\sqrt{J}$ is related to the convergence to the CLR transformation and is shown in Appendix 1.
5. $\mathbf{Z}_{[\lambda]} = \left[z_{ij[\lambda]}\right]$ is the chiPower-transformed data matrix with power λ. Euclidean distances between the rows of $\mathbf{Z}_{[\lambda]}$ are called chiPower distances between the rows of $\mathbf{X}$, which for $\lambda = 1$ are the chi-square distances in a regular CA context. The set of all Euclidean distances between rows of $\mathbf{Z}_{[\lambda]}$, i.e. the Euclidean geometry of chiPower-transformed data, defines the chiPower geometry of the original matrix $\mathbf{X}$, corresponding to the power λ.

As shown in Appendix 1, the chiPower transformation converges in the limit, as λ tends to 0, to the CLRs that have been negatively shifted by the column means of $\mathbf{Z}_{[\lambda]}$. This can be corrected to give actual CLRs in the limit, if required, by simply adding the column means of $\mathbf{Z}_{[\lambda]}$. This is done by default in the R function `chiPower()`, provided as online supplementary material.

The way the power λ is chosen will depend on the statistical learning objective. In unsupervised learning, the power can be chosen to make the chiPower geometry of the samples be as close as possible to their logratio geometry (see Sect. 2.4).

This means that methods such as PCA and clustering of the samples can be validly performed on the chiPower-transformed data, as an alternative to logratio-transformed data. This alternative is particularly useful for compositional data with zeros, since no zero replacement is necessary, but it can also be useful for strictly positive data, since the interpretation is simplified, in terms of parts, not logratios.

In supervised learning where the compositions serve as predictors of a response, λ will be chosen to optimize model fit or predictivity, and if the sample is of sufficient size, the power can be chosen by cross-validation. In this case, not all the above steps are necessary—for example, steps 3 and 4 only change the scales of the predictors linearly and this does not affect their roles in modelling. In supervised learning where the compositions serve as responses, however, not only would closeness to logratio geometry be important, but also the predictability of the compositions by the explanatory variables—in this case a compromise would perhaps be desirable in choosing λ as a compromise between these competing objectives.

The idea to apply the Box-Cox style of power transformation to compositional data is not new—see Aitchison (1986), Rayens and Srinivasan (1991), Tsagris et al. (2016). Greenacre (2010) showed the connection between Box-Cox transformation prior to performing CA and logratio analysis (LRA, i.e. the PCA of CLR-transformed data). In the present work, however, we use this idea in a much wider context of analysing compositional data, both unsupervised and supervised. A recent paper by Erb (2023) also looks at estimating the power parameter of power-transformed compositions, considering this as a shrinkage problem, even proposing to estimate a different power for each sample. Estimating a different power for each compositional part is a further possibility, since each part has a different level of skewness.

Furthermore, Section S3 of the Supplementary Material shows how CA applied to a closed power-transformed data matrix, where the samples (rows) are equally weighted, reduces to a PCA of the chiPower-transformed data. The only difference between the two analyses is the treatment of the scalar factor $\frac{1}{\lambda}$, which is eliminated in CA and so has to be re-introduced into the final CA results.

2.4 Measuring closeness to isometry

Isometric means "the same metric", that is the same distance structure in multivariate space. In the present context, the term applies to the comparison with the sample geometry based on logratio distances, which are the Euclidean distances computed on the CLRs—see Section 2.2. Notice that the specific definition of logratio distance by Greenacre (2018, 2021) allocates weights to both the samples and the compositional parts, where equal weights are used in the present work for both rows and columns.

Hence, on the one hand, consider the logratio distances between all the samples as the reference, where any data zeros have to be replaced (see Section 2.6), and, on the other hand, the distances between the same samples based on chiPower-transformed data, where no zero-replacement is required. The closeness of the sample geometry of chiPower-transformed data to the sample logratio geometry can be measured by the Procrustes correlation between the respective sample configurations (Appendix 2 explains how this correlation is obtained). A convenient way to do this is to apply

PCA to the CLR-transformed data and to the chiPower-transformed data respectively, obtain the complete set of principal coordinates in each case, and then fit these two coordinate matrices to each other by Procrustes analysis. If the Procrustes correlation is close to 1, this means that the transformation is close to being isometric (always with respect to the logratio geometry, taken as the reference.)

Isometry is important in unsupervised learning, when the structure of the compositional data is being explored by methods such as dimension reduction and clustering, in which case it will be favourable to be close to the logratio geometry, which is known to be coherent. It can also be important in supervised learning when the compositions serve as responses to additional explanatory variables, since it is the complete compositional structure that is being modelled. This case is not considered in this paper, but see Yoo et al. (2022) for an application.

2.5 Measuring closeness to coherence

Whereas isometry is a property of the samples, coherence is a property of the compositional parts, usually the columns of the data matrix. Using PLRs and their special case, the ALRs, is a perfectly coherent strategy: for example, PLRs involving pairs of parts A, B and C are not affected if additional parts D and E are added to the composition.

There is nevertheless a relationship between the two concepts of coherence (of the parts) and isometry (of the samples). In Appendix 1, explicit convergence of the chiPower transformatio to the CLR transformation is shown. It follows that, since the logratio transformation is perfectly coherent, a transformation such as the chiPower is converging to isometry and coherence at the same time, as the power of the transformation tends to zero.

Notwithstanding this relationship, it is still useful to quantify the level of coherence in a particular application by comparing results for parts in subcompositions and the same parts in the "full" compositions of the given data. In each case the parts have been transformed in the same way (in this case, using the same chiPower transformation) but computed on different compositions due to the closing operation. This comparison does not involve the logratio transformation at all—it is confined to the chiPower transformation, or any other transformation that one wants to check for coherence. It is also useful to see how the lack of coherence (i.e., incoherence) is affected by the size of the subcompositions, since the subcompositional values will change more due to closing when there are less parts in the subcomposition than in larger subcompositions. The type of results to compare depends on the research problem, because coherence has a different meaning if the statistical analysis is unsupervised or supervised.

In CoDA there is the symmetric concept of the logratio geometry of the parts: logratios can be computed for each part pairwise across the samples (i.e., $I(I-1)/2$ logratios), and their structure is related in the same way to that of the CLRs of the parts (Aitchison and Greenacre 2002; Greenacre 2021). There is more than one way to quantify the geometry of the parts in the chiPower approach. One way is to simply transpose the data matrix and apply the chiPower transformation as before, in other words chiPower the columns (parts). Another way, which is adopted here, is to use the geometry of the column principal coordinates in the PCA of the chiPowered data.

This defines a distance geometry on the parts which is equivalent to the covariance structure of the transformed parts (Greenacre et al. 2022). For unsupervised learning, this chiPower geometry of the transformed compositional parts in many different random subcompositions will be compared to the chiPower geometry of the same parts, transformed in the same way, in the full compositional data matrix, again using the Procrustes correlation. So this is a similar measure as the one of isometry between the sample geometries, but between the same parts in the subcomposition and the composition. In other words, the coherence check is being made by measuring the isometry between the subset of parts.

The algorithm for assessing the coherence can be summarized in the following steps.

1. Transform the compositional data matrix $\mathbf{X}$ using chiPower, for the power λ of interest, resulting in $\mathbf{Z}_{[\lambda]}$
2. Perform the PCA of $\mathbf{Z}_{[\lambda]}$ using the SVD $I^{-\frac{1}{2}}\mathbf{Z}_{[\lambda]} = \mathbf{U}\mathbf{D}_\phi\mathbf{V}^\mathsf{T}$ (see Supplementary Material Section S3).
3. The part geometry of all the parts is defined by the coordinates $\mathbf{G} = \mathbf{V}\mathbf{D}_\phi$.
4. For any subcomposition $\mathbf{X}_s$, perform the same chiPower transform to obtain $\mathbf{Z}_{s[\lambda]}$.
5. Perform the PCA on $\mathbf{Z}_{s[\lambda]}$ (steps 2. and 3.) and define the geometry of the subcompositional parts from the results of this PCA in the same way as before, i.e., coordinates $\mathbf{G}_s$.
6. Compute the Procrustes correlation between $\mathbf{G}_s$ and the subset of rows of $\mathbf{G}$ corresponding to the same subset of parts in the subcomposition.

The above is repeated for many subcompositions of different sizes.

The previous approach by Greenacre (2009) to measure incoherence used a stress measure common in multidimensional scaling (Borg and Groenen 2010), applied to the distances between parts. This approach used a "worst-case scenario" of two-part subcompositions, which might be acceptable for small compositions but is too extreme and unrealistic for larger ones that are generally the case in practice. Here it is preferred to use a range of subcompositions in the range of 10–90% of the total number of parts, so that the lack of coherence can also be assessed for subcompositions of different sizes.

For supervised learning when compositions serve as predictors, this approach of comparing geometries of subsets of parts is no longer important, and coherence would rather be assessed by seeing how the model parameter estimates vary for the subcompositional parts compared to their compositional counterparts, all with the same chiPower transformation.

There are clearly very many possibilities to choose subsets of parts in order to create subcompositions and check for incoherence. Random subsets of parts can be selected, or it may be that subcompositions in particular applied contexts tend to include the more frequent parts more often 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 that they include. Similarly, in studies of fatty acid compositions, it is again the rarer fatty acids that might not appear in some

studies, depending on the sophistication of the laboratory equipment used in the data collection.

2.6 The problem of data zeros

With the chiPower transformation and measures of closeness to isometry and coherence in place, attention is now turned to compositional data with zeros. The problem of zeros has been called the "Achilles heel" of compositional data analysis (Greenacre 2021), 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 Lubbe et al. (2021) for a review. The presence of many zeros can cause problems in the analysis (te Beest et al. 2021).

Using the chiPower transformation provides an approach to avoid zero replacement, but as the power decreases, an incompatibility with logratios will develop. This is because the transformation of the original zeros leads to very large negative numbers as lambda tends to 0 and the transformed zeros approach minus infinity, with a resultant degradation of the metric properties of the transformed data. In the present approach, for data with zeros, the power of the chiPower transformation will be identified that leads to the transformed data having maximum isometry with the sample logratio geometry. However, zeros will have to be replaced to enable computations of the CLRs, which define the logratio geometry, so there is a slight disparity in the comparison between the chiPower-transformed data that have zeros and the logratio-transformed data that have zeros replaced. See Supplementary Material Section S1 for further discussion of this issue.

3 Results

3.1 Unsupervised learning: strictly positive compositions

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

Logratio analysis (LRA) 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. This is PCA applied to the CLRs. Then PCA is performed on the chiPower-transformed data, with powers λ descending from 1 in small steps to almost 0, where "almost" is $\lambda = 0.0001$. These analyses are effectively all CAs on closed power-transformed data, as explained in Supplementary Material Section S3.

Figure 1A shows a plot of the Procrustes correlations between the logratio geometry of the 89 samples and corresponding chiPower-transformed geometry, showing the
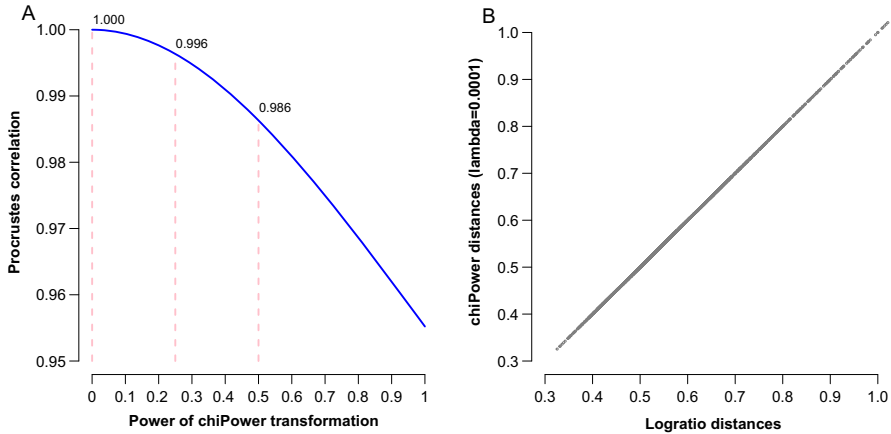
Fig. 1 **A** The Procrustes correlations for different powers of the chiPower transformation, measuring proximity to isometry between the exact logratio geometry and the geometry of chiPower-transformed data, showing the convergence to exact isometry close to 0. **B** For the power equal to 0.0001, the chiPower distances are practically identical to the logratio distances. In the limit as the power tends to 0, they are identical

convergence to 1 as λ tends to 0. In each case along the curve the 88-dimensional logratio geometry is compared to the 88-dimensional chiPower geometry. Values indicated are for square root, fourth root and ten thousandth root ($\lambda = 0.0001$) 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 chiPower distances for the $\lambda = 0.0001$ case, where the almost exact isometry is further shown.

To further illustrate the theoretical convergence of these geometries, Figure 2 shows the two-dimensional results of the CA for $\lambda = 1$ (original CA), 0.5 (CA on square-root data), 0.0001 (CA on ten thousandth-root data), and finally LRA. As shown in Supplementary Material S3, these CAs are identical to PCAs on chiPowered data. Figure 2C and D 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 Unsupervised learning: compositions with zeros

Here both the 'Rabbits' and the 'Crohn' data sets will be used to demonstrate how the chiPower transform can handle data zeros. To simulate a situation where zeros are present in the 'Rabbits' data, 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 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 closed to compositions, and analysed in a similar way as before. In order to compare the results using the chiPower and logratio transformations, the zeros were imputed using the function cmultRepl in the zCompositions R package (Palarea-Albaladejo and Martin-Fernandez 2015), which is one of the popular
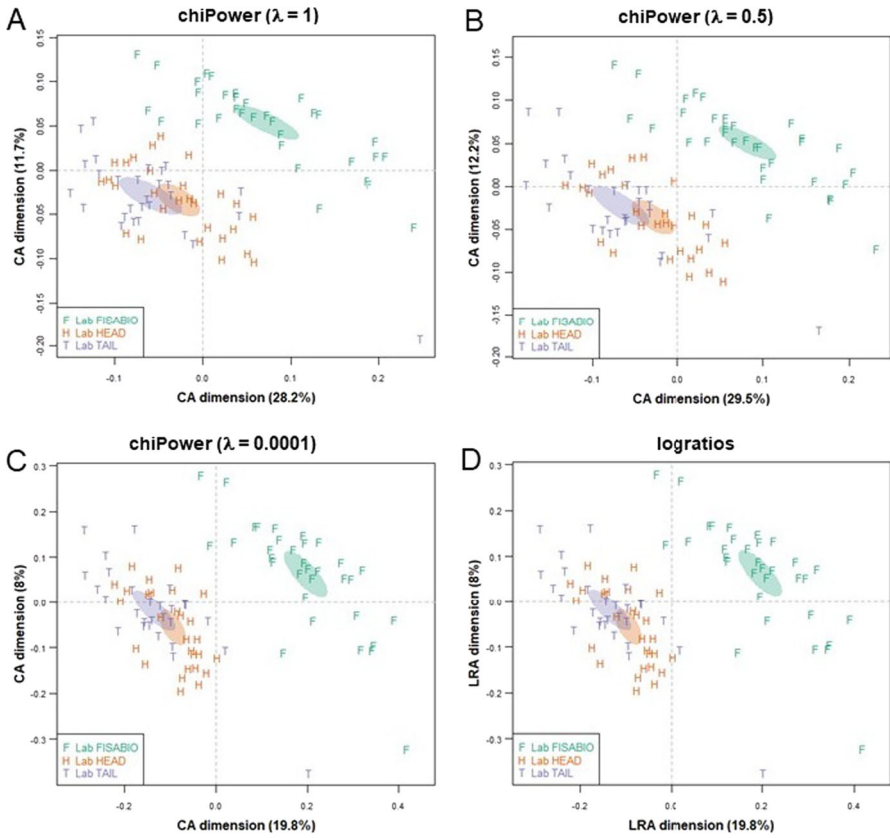
Fig. 2 Using the rabbits data set, three CAs, i.e. PCAs of chiPower-transformed compositions, with decreasing powers, and LRA analysis as the limit solution. **A** The regular CA with power 1. **B** CA with power 0.5 (square root). **C** CA with power 0.0001. **D** Logratio analysis (LRA). **C** and **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 corresponding to three testing laboratories

ways of zero replacement. The chiPower-transformed geometry of the data (with zeros) was then compared to the logratio geometry of the data matrix with zeros replaced.

The chiPower distances cannot reproduce exactly the logratio distances, because they are operating on slightly different data matrices, and thus convergence to logratio distances cannot be attained. However, the geometries can come very close to each other 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 $\lambda = 0.22$, which is close to a fourth-root transformation. The concordance of the chiPower and logratio distances can now be seen in Fig. 3B.

Since the chiPower-transformed data with $\lambda = 0.22$ are close to isometry, it is expected that they will also be close to coherence. This is assessed by taking many random subcompositions, as described in Section 2.5, each of which is reclosed and its subcompositional part geometry compared with that of the corresponding subset of
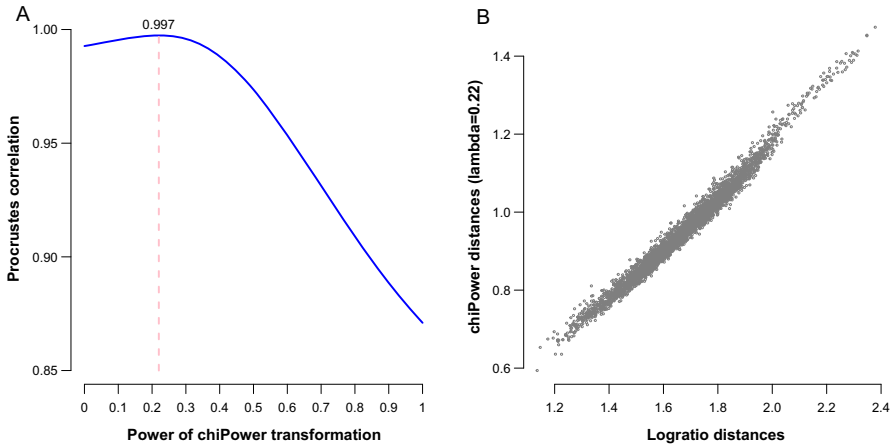
Fig. 3 A The Procrustes correlations, measuring proximity to isometry between the exact logratio geometry on the original data and the chiPower geometry with different powers, using the data with simulated zeros. The correlation is at an optimum value of 0.997 for a power of 0.22. **B** In the respective 88-dimensional spaces, the chiPower distances, with $\lambda = 0.22$, are quite similar to the logratio distances

parts in the full composition. Once again, Procrustes correlation is used to measure the degree of coherence. To contrast this with doing no change at all to the compositional data, the raw untransformed compositions were first assessed for isometry, which means that the regular Euclidean distance geometry on the raw compositions was correlated with the logratio geometry. The Procrustes correlation was computed as 0.891, and so it is expected that the coherence of the untransformed compositions will be worse than the quasi-isometric chiPower-transformed compositions with $\lambda = 0.22$. This is indeed how it turns out in the subcompositional coherence exercise, which does not involve a comparison with logratio-transformed data, shown in Fig. 4.

The same exercise was performed for the Crohn data set, and similarly successful results were obtained, given in Supplementary Material Section S5, where the optimal value of the power was $\lambda = 0.25$. The result turns out to be dependent on the zero replacement. Supplementary Material Section S1 further investigates the effect of using different zero replacements, for example adding 0.5 to the original data, or simply substituting the zeros by 0.5.

3.3 Supervised learning: use of power transformations

Compositions can serve as predictors of a response, or can form a multivariate response to other explanatory variables (e.g., Yoo et al. (2022)). In the latter case, isometry will still be relevant, since this affects the total compositional variance to be explained. Attention is restricted here to the former case, where the issue of isometry is no longer relevant but coherence certainly is, since the effect sizes and interpretation of the predictors should not depend on the particular (sub)composition they are part of—see Section 3.4. Since there are many parts in a composition, the question of variable
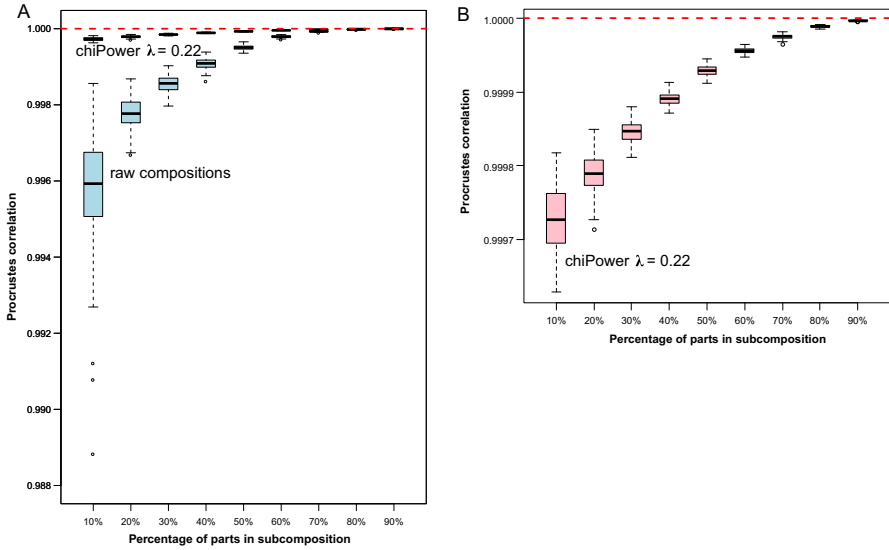
Fig. 4 Deviations from exact coherence (Procrustes correlation = 1) for 1000 random subcompositions of sizes 10% up to 90% of the 3937-part Rabbits data with simulated zeros. For each size of subcomposition, a boxplot of the Procrustes correlations is shown, for the untransformed raw compositions and for chiPowered compositions. **A** The lower sequence of boxplots is for the raw compositions, where no transformation is made at all, and deviations from coherence are much larger, especially for subcompositions with less parts. The upper sequence is for chiPower-transformed data, with power = 0.22 (the value obtained from the exercise on isometry), where deviations are very small, even in the smaller subcompositions. **B** The same boxplots as the upper sequence in A., i.e. the chiPowered data, with expanded vertical scale

selection is first addressed in this section, comparing the predictors that are either logratio- or chiPower-transformed.

The Crohn data set, with 975 samples and 48 bacteria, is used for this purpose since it has a dichotomous response $y =$ Crohn (patient with Crohn's disease), or $y =$ no (no disease), to be predicted from the compositions. Logistic regression models for predicting Crohn, using PLRs, have already been fitted in two different ways, by Coenders and Greenacre (2022) and Calle et al. (2023). Coenders and Greenacre (2022) proposed three forward stepwise algorithms for choosing PLRs, the first one being unrestricted choice from all possible PLRs, of which there are $\frac{1}{2} \times 48 \times 47 = 1128$. The available stopping criteria options were the Akaike information criterion (AIC), the stronger Bayesian information criterion (BIC) and the even stronger penalty on the number of variables in the model using the Bonferroni rule. This approach is implemented in the function `STEPR()` in the R package `easyCODA` (Greenacre 2018). For the present application, the BIC stopping criterion will be used.

Using a different approach, Calle et al. (2023) includes all the PLRs and imposes ElasticNet penalization on the predictors (Hastie et al. 2009), as implemented in the package `coda4microbiome`.

The above two approaches will be contrasted with simply using the power-transformed compositions, where the power is used as a tuning parameter to optimize the prediction. This third option using chiPower is the only one of the three that uses

Table 1 Results from three alternative ways of predicting Crohn's disease, based on different transformations of the compositional data: the first two using logratios on the modified version of the data set (with 1 added to all cells, from the `coda4microbiome` package), and the last using an optimized power transformation (power = 0.28) of the compositional parts from the original data set, with zeros

Method	Full data set				Cross-validation			
	PLRs	Parts	AUC	Acc'y	Sens'y	Spec'y	AUC	Acc'y
Forward stepwise	11	19	0.859	82.2%	0.826	0.704	0.831	79.2%
ElasticNet penalty	27	24	0.848	80.2%	NA	NA	0.824	NA
Power ($\lambda = 0.28$)	–	14	0.859	81.6%	0.819	0.697	0.826	78.6%

Different variable selection strategies are used: ElasticNet penalization, and the Bayesian information criterion (BIC) for the other two. (PLRs: the number of selected pairwise logratios. Parts: the number of parts included in the model. AUC: area under curve. Acc'y: accuracy (percent of correct predictions). Sens'y: sensitivity. Spec'y: specificity. NA: result not available in Calle et al. (2023))

the original version of the data with zeros. Notice that the chi-square standardization as well as the multiplication by $\frac{1}{\lambda}$ and subtraction of 1 in the Box-Cox transformation (3) are not necessary here, as such scale changes do not affect the predictions, just the values of the regression coefficients. Since Calle et al. (2023) uses the area under curve as a measure of prediction, and optimizes the variable selection using ten-fold cross-validation, the same approach is adopted here, to ensure comparability. The results are summarized in Table 1.

The performance of all three is similar, but the simpler power transformation of the compositions needs only 14 parts. The ElasticNet approach (Calle et al. 2023) chooses 27 logratios, involving 24 parts, while the forward stepwise approach (Coenders and Greenacre 2022) selects 11 logratios, involving 19 parts. Ten-fold cross-validation, using the same folds, evaluates the performance of each approach. Since the cross-validation AUC of the ElasticNet approach is an average of the AUCs of the ten folds, the mean AUC is also calculated for the other two methods. The power that is optimal in this supervised learning problem is $\lambda = 0.28$, slightly higher than the power of 0.25 that was optimal in the unsupervised objective reported in Supplementary Material Sections S4 and S5. The question of coherence and interpretation of the results of this third approach is dealt with in Section 3.4.

3.4 Coherence of the modelling with power-transformed compositions

In the previous subsection, a small subset of 14 parts, power-transformed, was identified as good predictors of the Crohn disease response. In this subsection the results and their interpretation are explained and it is investigated how the results would have changed if a subcomposition had been observed. Such a subcomposition would include the selected 14 parts, but would have different compositional values due to the closing of the subcomposition. For predictors in the form of PLRs or ALRs, their exact coherence ensures that the results remain the same—that is, a result for any subcomposition would be identical if any number of compositional parts were eliminated (or added) to the data set and the data reclosed to sum to 1. But for other transformations such as

Fig. 5 Scatterplot of standardized regression coefficients from models of same 14 predictors, fitted to data from the full 48-part composition, and to data from the 14-part re-closed subcomposition. The aspect ratio is 1 and the diagonal line at 45 degrees represents exact concordance
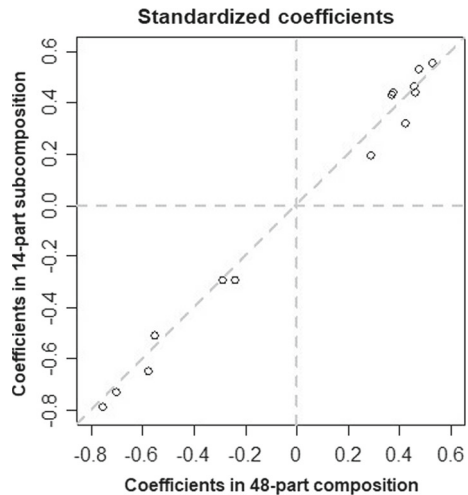
the present power-transformed one, a check is necessary on the extent of the lack of coherence in the results.

The first check is to isolate the 14 parts in a subcomposition, reclose, and then repeat the model. In order to compare the regression coefficients, because of changes of scale in the predictors, it is preferable to standardize the predictors in each case (i.e., mean 0, variance 1) in order to obtain standardized regression coefficients. This will also make the results invariant to performing a simple power transform or the chiPower transform. Figure 5 shows the coefficients to be almost in the same order and practically the same in value, whether the predictors are part of the original composition or in the subcomposition. This concordance between the two sets of coefficients shows that the power (or chiPower) transformation is very close to coherence in the sense of the modelling. Compared to the optimized results of Table 1, the AUC and accuracy, when the model is fitted to the closed 14-part subcomposition, both drop slightly from 0.859 to 0.847 and from 81.6 to 79.7%, respectively. This loss of predictivity might well be improved if the power was tuned specifically to optimizing coherence in the modelling, as opposed to the unsupervised objective of optimizing the isometry with respect to the sample logratio geometry.

To further investigate the coherence issue in the modelling, random subcompositions involving the same 14 parts but also additional parts, randomly selected and of random extents (from 1 to 33 additional parts), are added to the data set. For each of these, the subcomposition is closed and then power-transformed using $\lambda = 0.28$ (Table 1), and the logistic regression repeated using the 14 parts as predictors. Figure 6 shows the results for 1000 random subcompositions.

The original standardized regression coefficients are shown as vertical black lines, in the centre of a 95% confidence interval in light blue or pink, according to the margin of error $\pm 1.96\,\mathrm{SE}$ for each coefficient. The estimated coefficients in the subcompositions are shown as vertical blue or red lines (for negative and positive coefficients, respectively), where it can be seen that they all span the original estimates, and are well within the confidence intervals. For each of the 1000 subcompositions the accuracies
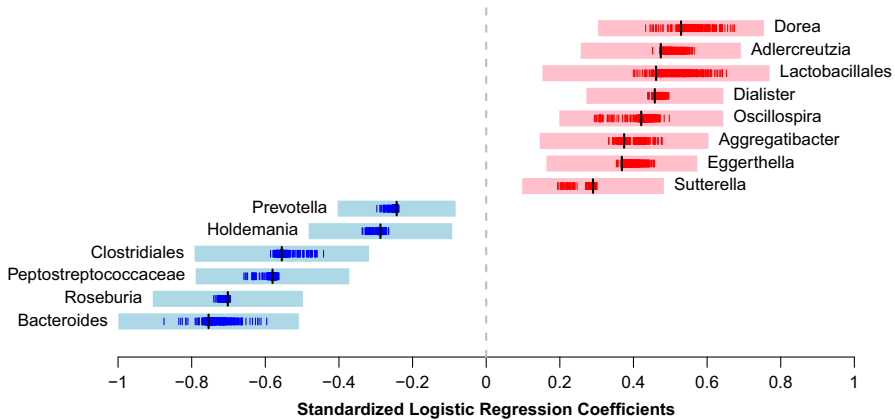
Fig. 6 Standardized logistic regression coefficients from 1000 analyses using different subcompositions of the Crohn data set. Each subcomposition contains the 14 parts used in the original model reported in the third row of Table 1 with random extents (1–33 additional parts), then each data set is closed and the parts are power-transformed using power = 0.28. The pale red and blue bars show 95% confidence intervals for the estimated coefficients in the original model and the vertical black lines are the point estimates, at the midpoints of the confidence intervals. For each variable there are 1000 vertical red or blue lines showing the estimates from the models using subcompositions. Positive coefficients increase the log-odds of Crohn's disease, negative coefficients decrease the log-odds

and AUCs are also computed, and 95% of the accuracies are between 80.0% and 81.4%, while 95% of the AUCs are between 0.848 and 0.859. This further demonstrates that the logistic regression results would be substantively the same for any subcomposition, so that the modelling using power-transformed compositions is coherent for all practical purposes, further supporting the good performance of these power-transformed predictors in Table 1.

Another diagnostic of coherence is to see how much the regression coefficients change as a function of the sizes of the chosen subcompositions. Random subcompositions of 10%, 20%, etc., up to 90% of the 33 remaining microbial taxa were taken, where the 14 parts in the original model are again always included, 100 subcompositions in each case. The dispersions of the AUC values (original value of 0.859 in the model—see Table 1) and for the regression coefficients of Roseburia (original standardized coefficient in the model equal to −0.702) are shown in Fig. 7, in the form of boxplots. For small subcompositions the AUCs are under-estimating, as already seen when just the 14-part subcomposition (with no others added) was analyzed. The standardized coefficients of Roseburia are more negative but both the model AUCs and these coefficients converge to the values in the original model as the subcompositional size increases. The dispersion of these coefficients should be judged against the margins of error of the estimates in the full composition. For example, Roseburia's coefficient estimate is −0.702, with a SE of 0.104, giving a 95% confidence interval of [−0.906, −0.498], much wider than the dispersions shown in Fig. 7B.

As for the interpretation, this is made directly on the part values (power-transformed), not on logratios, which is a considerable simplification. The standardized regression coefficients, shown in Fig. 6, give a model for log-odds of Crohn's disease
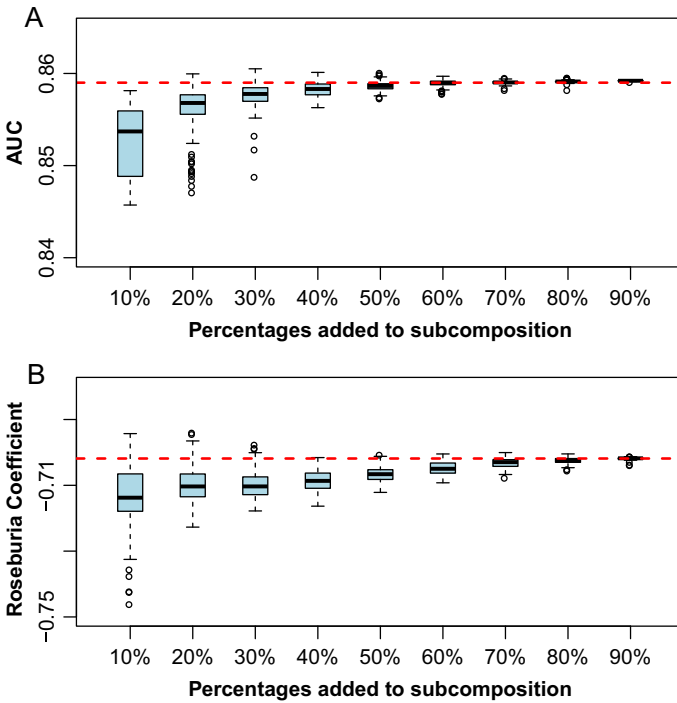
Fig. 7 Results from logistic regressions performed on 100 subcompositions for each of nine different percentages of the remaining parts added to the basic set of 14 power-transformed parts that were used in the original model predicting Crohn's disease. Boxplots show the dispersions for each subcomposition of increasing sizes. **A** AUC of prediction. **B** Standardized regression coefficients for Roseburia. The red dashed lines show the original values in the full composition (0.859 and − 0.702 respectively)

in terms of the 14 standardized predictors of the following form, showing only the extreme negative and positive terms:

$$\log(\frac{p}{1-p}) = 1.238 - 0.754\,\text{Bacteriode}^* - 0.702\,\text{Roseburia}^* \cdots$$
$$+0.474\,\text{Adlercreutzia}^* + 0.529\,\text{Dorea}^* \tag{5}$$

where $*$ indicates the standardized power-transformed variables. Alternatively, the equivalent model can be expressed in terms of the values in the original composition using power-transformed variables and no standardization, where the magnitude and order of the coefficients changes according to the ranges of the different predictors:

$$\log(\frac{p}{1-p}) = 3.197 - 5.027\,\text{Roseburia}^{0.28} - 4.866\,\text{Peptostreptococcaceae}^{0.28} \cdots$$
$$+5.208\,\text{Eggerthella}^{0.28} + 7.948\,\text{Adlercreutzia}^{0.28} \tag{6}$$

Whichever form is reported, it should be remembered that the effect sizes are applicable to infinitesimal (i.e., very small) changes in the predictors, and should not

be taken as linear effects as in regular regression. Like partial derivatives, these are measures of local changes. This is due to a change in one compositional value affecting all the others. For example, suppose all the predictors are at their mean values. The value of the regression equation in (6), including the constant is computed to be 2.583, back-transformed to a probability of Crohn's disease equal to 0.930 ($p = e^{2.583}/(1 + e^{2.583}) = 0.930$). Suppose the compositional mean value of Roseburia is multiplied by 20, which is still within the range of this bacteria's observed values. Simply making this increment and applying the model to the new set of power-transformed values results in the value 0.667. This back-transforms to a probability of 0.661, less than 0.930, as expected since Roseburia's coefficient in the regression is negative.

But one cannot simply change a compositional value as one would do with regular statistical variables, since the other compositional values are affected by the change. Hence, the increased value of Roseburia has to be compensated by a decrease in the compositional values of the other bacteria. Applying a proportional decrease to the other bacteria to obtain a composition that sums to 1 and applying the model formula leads now to a value of 0.607 and a back-transformed probability of 0.647, which would be a more accurate estimate of the effect of the Roseburia increase.

This issue of quantifying the correct effect sizes due to the nature of the compositional data, taking into account that a change in one part affects the others, is similarly present when logratios are used as predictors and the model is expressed as a log-contrast (Coenders and Greenacre 2022).

4 Conclusion

This paper demonstrates that an alternative pipeline is possible for analysing compositional data, using the chiPower transformation. This transformation combines a Box-Cox style of power transformation with the chi-square standardization that is inherent in correspondence analysis. The choice of the power gives the approach its flexibility. Unlike logratio transformations, this transform allows data zeros—notice that in the Crohn application, 28.8% of the original data matrix are zeros and need replacement in order to compute logratios. In an unsupervised learning context, where understanding of the data structure is sought, the power can be identified to maximize the proximity of the sample geometry of chiPower transformed compositions to the sample logratio geometry, using the Procrustes correlation as a measure of the closeness to isometry. In a similar way, the Procrustes correlation between the geometries of subsets of parts in a composition and the same parts in subcompositions gives a quantitative assessement of (subcompositional) coherence. For supervised learning where the compositions are predictors of a response, the power serves as a tuning parameter to optimize prediction of the response, preferably using cross-validation. In this case, where a subset of power-transformed predictors is selected and a model fitted, the coherence can be assessed by repeating the model fitting on subcompositions of different sizes and observing how the model estimates are affected.

Overall, in summary, the chiPower transformation, supported by diagnostics to assess the properties of isometry and coherence, can present a simpler and more easily

interpretable alternative to the logratio transformation, with the great advantage that no zeros need replacing.

These results give food for thought about the role that logratios play in compositional data analysis. When data are all positive, the logratio approach can be adopted, with its favourable property of exact coherence. But when there are data zeros, the user has a choice, either to use an algorithm to literally create data to replace the zeros for the sake of using logratios, or use an alternative approach that needs no change to the data and that can be shown to be almost isometric and coherent in terms of the research objective, either unsupervised or supervised. The different zero replacement methods can lead to different results (see Supplementary Material Section S1) and there is apparently no clear consensus about which is preferred in a specific context. Hence, it may be that an alternative approach, such as the chiPower transformation presented here, is preferable in the presence of data zeros, especially many data zeros. Notice that the investigation of the coherence of the chiPower transformation is achieved without any zero replacements.

In summary, transformations such as chiPower, which are highly coherent and needing no zero replacement, are proposed as a preferred first choice for analysing compositional data that have zeros. As Lundborg and Pfister (2023) state:

> "...we believe that it is generally preferable to modify the statistical procedure to fit the data rather than vice versa".

Then, if logratios are of specific interest, for whatever reason, the tables could be turned by choosing the zero replacement method that leads to logratio-transformed data that come closest (for example, in terms of isometry or model accuracy) to the data transformed by the preferred method that needs no zero replacement (e.g., chiPower).

For strictly positive data, both approaches are possible: (a) the purely logratio approach, where the final interpretation is in terms of logratios and log-contrasts, or (b) the chiPower approach where the interpretation is in terms of the original compositional parts, which may be easier for the practitioner, especially for supervised learning.

Appendix 1: Relationship between chiPower and CLR transformations

The fact that the chiPower transformation links directly to LRA, which is the PCA of the CLR-transformed positive compositional data, implies a direct link from the chiPower transform and the CLR transform. To show this, first consider this result, for the positive composition $\mathbf{x}_i = [\, x_{i1} \; x_{i2} \; \cdots \; x_{iJ} \,]$ in the i-th row of the compositional data matrix $\mathbf{X}$. Let $y_{ij[\lambda]} = x_{ij}^{\lambda} / \sum_k x_{ik}^{\lambda}$, i.e. the closed powered compositions. The convergence of $J y_{ij[\lambda]}$ to the CLR transformation, in Box-Cox formulation, is as follows

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \left(J \frac{x_{ij}^{\lambda}}{\sum_k x_{ik}^{\lambda}} - 1 \right) = \mathrm{CLR}(\mathbf{X})_{ij} = \log \left(\frac{x_{ij}}{g(\mathbf{x}_i)} \right) \qquad (7)$$

where $g(\mathbf{x}_i)$ is the geometric mean of the J elements of $\mathbf{x}_i$. To show this, divide the numerator x_{ij}^λ and denominator $\sum_j x_{ij}^\lambda$ both by $g(\mathbf{x}_i)^\lambda$.

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \left(J \frac{x_{ij}^\lambda / g(\mathbf{x}_i)^\lambda}{\sum_k x_{ik}^\lambda / g(\mathbf{x}_i)^\lambda} - 1 \right) \tag{8}$$

In the denominator $\lim_{\lambda \to 0} \sum_k x_{ik}^\lambda = J$ and $\lim_{\lambda \to 0} g(\mathbf{x}_i)^\lambda = 1$, hence the limit reduces to

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \left(\left[\frac{x_{ij}}{g(\mathbf{x}_i)} \right]^\lambda - 1 \right) = \log \left(\frac{x_{ij}}{g(\mathbf{x}_i)} \right) \tag{9}$$

using the Box-Cox theorem. A different proof of this result is given in the Appendix of Tsagris et al. (2016) using series expansions.

To prove the convergence properties of the chiPower transform, Tsagris el al's style of proof will be used here. The following results are used in the proof, both based on Taylor series expansions of these functions of λ, around the value $\lambda = 0$:

- $x^\lambda = 1 + \lambda \log(x) + O(\lambda^2)$
- $(1 + \lambda x)^a = 1 + a\lambda x + O(\lambda^2)$

In the proof, the terms in $O(\lambda^2)$ (including higher power) are written just the first time they occur in an expansion and then omitted since they will eventually disappear in the limit.

The basic chiPower transformation is $y_{ij}[\lambda]/\sqrt{\bar{y}_j[\lambda]}$ where $\bar{y}_j[\lambda] = (1/I) \sum_i y_{ij}[\lambda]$, the column means of the $y_{ij}[\lambda]$. The numerator and denominator are first handled separately.

The numerator is expanded as follows:

$$\begin{aligned}
y_{ij}[\lambda] &= \frac{x_{ij}^\lambda}{\sum_k x_{ik}^\lambda} \\
&= \frac{1 + \lambda \log(x_{ij}) + O(\lambda^2)}{\sum_k (1 + \lambda \log(x_{ik}) + O(\lambda^2))} \\
&= \left(1 + \lambda \log(x_{ij}) \right) \frac{1}{J} \left(1 + \frac{\lambda}{J} \sum_k \log(x_{ik}) \right)^{-1} \\
&= \frac{1}{J} \left(1 + \lambda \log(x_{ij}) \right) \left(1 - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \\
&= \frac{1}{J} \left(1 + \lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \tag{10}
\end{aligned}$$

From this result the inverse of the denominator is expanded as

$$1/\sqrt{\bar{y}_j[\lambda]} = \left(\frac{1}{I} \sum_i y_{ik}[\lambda]^\lambda \right)^{-0.5}$$

$$= I^{\frac{1}{2}} J^{\frac{1}{2}} \left(\sum_i \left(1 + \lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \right)^{-0.5}$$

$$= J^{\frac{1}{2}} \left(1 + \frac{1}{I} \sum_i \left(\lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \right)^{-0.5}$$

$$= J^{\frac{1}{2}} \left(1 + (-0.5) \frac{1}{I} \sum_i \left(\lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \right) \quad (11)$$

The division of numerator by denominator is thus the product of (10) and (11). Many products of terms are $O(\lambda^2)$ and the only ones remaining are those that are multiplied by the 1's in each bracket, reducing to

$$\frac{1}{J^{\frac{1}{2}}} \left(1 + \lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) \right.$$

$$\left. - 0.5 \frac{1}{I} \sum_i \left(\lambda \log(x_{ij}) - \frac{\lambda}{J} \sum_k \log(x_{ik}) + O(\lambda^2) \right) \right) \quad (12)$$

so that

$$\frac{1}{\lambda} \left(J^{\frac{1}{2}} \frac{y_{ij}[\lambda]}{\sqrt{\bar{y}_j[\lambda]}} - 1 \right) \to \text{CLR}(\mathbf{X})_{ij} - 0.5 \, \overline{\text{CLR}(\mathbf{X})}_j \ \text{ as } \lambda \to 0 \quad (13)$$

where $\text{CLR}(\mathbf{X})_{ij} = \log(x_{ij}) - \frac{1}{J} \sum_k \log(x_{ik})$ is the centered logratio defined in (7) and $\overline{\text{CLR}(\mathbf{X})}_j$ is the j-th column mean of $\text{CLR}(\mathbf{X})$.

Thus, the limit is the CLRs shifted negatively by half their column means. Since the column means of this limit in (13) are equal to (plus) half the column means, the negative shift can be cancelled by a translation that adds half the column means of the transformation $(1/\lambda)\left(J^{\frac{1}{2}} y_{ij}[\lambda]/\sqrt{\bar{y}_j[\lambda]} - 1\right)$. This "translated" version that converges to the CLR is the default in the R function chiPower(), but the "unadjusted" version can also be obtained as an option. Of course, these options that shift each part by a constant amount make no difference to computing distances, covariances, or models, since the column means are eliminated or just affect the constant terms in models. Notice that the similar proof by Choulakian (2023) is not for the chiPower transformation but for $y_{ij}[\lambda]/\bar{y}_j[\lambda]$, that is, dividing by $\bar{y}_j[\lambda]$ rather than by $\sqrt{\bar{y}_j[\lambda]}$. This ratio is a scalar multiple of the Pearson contingency ratio (Greenacre 2010) and converges to centered CLRs since the -0.5 in (13) above becomes -1 and the $J^{1/2}$ is eliminated, giving the result:

$$\frac{1}{\lambda} \left(\frac{y_{ij}[\lambda]}{\bar{y}_j[\lambda]} - 1 \right) \to \text{CLR}(\mathbf{X})_{ij} - \overline{\text{CLR}(\mathbf{X})}_j \ \text{ as } \lambda \to 0 \quad (14)$$

In fact, it is clear from the proof in Eqs. (10)–(13) that the general result for any power ϕ in the division $y_{ij}[\lambda]/\bar{y}_j[\lambda]^\phi$ can be obtained as follows:

$$\frac{1}{\lambda}\Big(J^{1-\phi}\,\frac{y_{ij}[\lambda]}{\big(\bar{y}_j[\lambda]\big)^\phi} - 1\Big) \rightarrow \text{CLR}(\mathbf{X})_{ij} - \phi\,\overline{\text{CLR}(\mathbf{X})}_j \quad \text{as } \lambda \rightarrow 0 \qquad (15)$$

of which (7), (13) and (14) are special cases for $\phi = 0, 0.5$ and 1 respectively.

These results are illustrated in the following R code, applied to the modified Crohn data without zeros, named `Crohn1` here. This data set is in the R package `coda4microbiome`. For the original Crohn data set with zeros, simply subtract 1 from the modified Crohn data. See Supplementary Material Section S1 for the source of this original Crohn data set.

```
### The chiPower transformation
chiPower <- function(X, close=TRUE, power=1, chi=TRUE,
                     BoxCox=TRUE, CLR=TRUE)
{
#  X: the compositional data matrix (it is closed in case)
#  close: close the data after powering
#  power: power of the transformation
#  chi: apply chi-square standardization
#  BoxCox: apply Box-Cox style of transformation
#  CLR: translate columns so that convergence is to CLR transformation
  foo <- as.matrix(X)
  foo <- foo / rowSums(foo)
  foo <- foo^power
  if(close)    foo <- foo / rowSums(foo)
  if(chi)      foo <- sweep(foo, 2, sqrt(colMeans(foo)), FUN="/")
  if(BoxCox & !chi)  foo <- (1/power)*(ncol(X)*foo - 1)
  if(BoxCox & chi)   foo <- (1/power)*(sqrt(ncol(X))*foo - 1)
  if (BoxCox & chi & CLR) foo <- foo + rep(colMeans(foo), each=nrow(X))
  X.chiPower <- foo
  X.chiPower
}

### Use modified Crohn data without zeros, Crohn1
library(easyCODA)
X <- CLOSE(Crohn1)

### CLRs
X.CLR <- CLR(X, weight=FALSE)$LR
### First 6 rows and 3 columns of CLR matrix
head(X.CLR,1:3])
#                 g__Turicibacter g__Parabacteroides g___Ruminococcus_
# 1939.SKBTI.0175      -4.920831          2.1617171          2.801403
# 1939.SKBTI.1068      -3.176255         -0.3430419          4.380173
# 1939.SKBTI047        -2.686822          1.2384461          2.980757
# 1939.SKBTI051        -1.956917          1.9142845          3.956586
# 1939.SKBTI063        -3.719167         -0.8287950          4.228865
# 1939.SKBTI072        -2.683135          6.1129017          2.948077

### chiPower without chi-square standardization and confirming (7)
### (compare these values with the CLR values given above)
lambda <- 0.001
head(chiPower(X, power=lambda, chi=FALSE))[,1:3]
```

```
#                      g__Turicibacter g__Parabacteroides g___Ruminococcus_
# 1939.SKBTI.0175           -4.914989          2.1577662         2.799038
# 1939.SKBTI.1068           -3.175177         -0.3469550         4.385789
# 1939.SKBTI047             -2.686208          1.2362093         2.982195
# 1939.SKBTI051             -1.957958          1.9131512         3.961451
# 1939.SKBTI063             -3.716163         -0.8323668         4.233885
# 1939.SKBTI072             -2.683284          6.1278452         2.948661

### chiPower with chi-square standardization but no CLR shift adjustment
head(chiPower(X, power=lambda, CLR=FALSE))[,1:3]
#                      g__Turicibacter g__Parabacteroides g___Ruminococcus_
# 1939.SKBTI.0175          -3.6753284          1.2511554         1.336367
# 1939.SKBTI.1068          -1.9333492         -1.2512999         2.920804
# 1939.SKBTI047            -1.4437711          0.3304323         1.519257
# 1939.SKBTI051            -0.7146141          1.0067617         2.497085
# 1939.SKBTI063            -2.4750092         -1.7362725         2.769121
# 1939.SKBTI072            -1.4408430          5.2176429         1.485772

### plot showing that it's just a shift away from CLRs (Figure 8)
plot(X.CLR$LR, chiPower.new(X, power=lambda, CLR=FALSE),
     col=rep(rainbow(ncol(X)), each=nrow(X)))

### chiPower with chi-square standardization and CLR shift adjustment
### (this is the default of the chiPower funcion)
head(chiPower(X, power=lambda))[,1:3]
#                      g__Turicibacter g__Parabacteroides g___Ruminococcus_
# 1939.SKBTI.0175           -4.919562          2.1566333         2.797086
# 1939.SKBTI.1068           -3.177583         -0.3458220         4.381522
# 1939.SKBTI047             -2.688004          1.2359101         2.979976
# 1939.SKBTI051             -1.958847          1.9122396         3.957804
# 1939.SKBTI063             -3.719243         -0.8307946         4.229840
# 1939.SKBTI072             -2.685076          6.1231208         2.946490
```

Figure 8 shows the scatterplot in the above code, for power $\lambda = 0.001$. Figure 9 shows the comparisons of the CLR transformation (horizontal axis on each plot), and the chiPower transformation, with the shift adjustment, for decreasing powers 1, 0.25, 0.1 and 0.001.

Fig. 8 Scatterplot of unadjusted chiPower transformation and the CLR transformation, showing that each part is a shift away from the CLR. This can easily be corrected if required, as in the next Fig. 9. Each colour corresponds to one of the 48 compositional parts
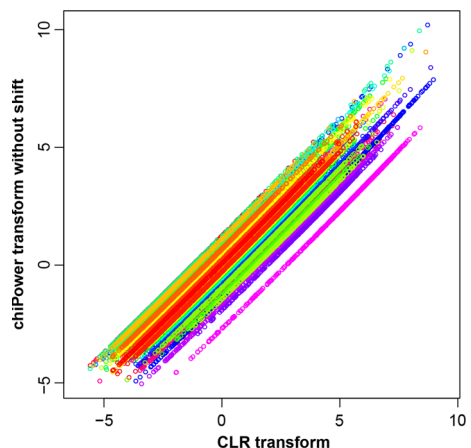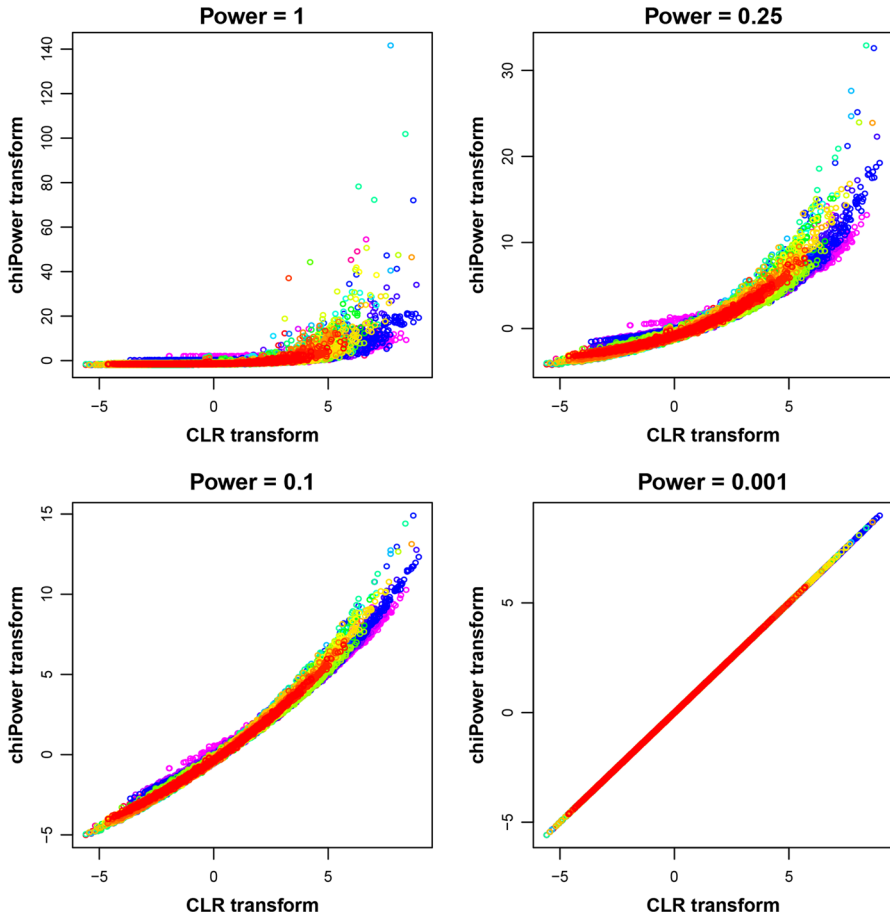
Fig. 9 Scatterplots of chiPower transformations for decreasing powers and the CLR transformation, with shift adjustment. Each colour corresponds to one of the 48 compositional parts

Appendix 2: the Procrustes correlation

Procrustes analysis (Gower and Dijksterhuis 2004) is a method for matching two multidimensional configurations by introducing translation, rotation and scaling operations to make them as similar as possible to each other. It is used here to measure how similar two data structures are (Krzanowski 1987), for example between the matrix $\mathbf{F}_1$ of principal coordinates from an LRA and the matrix $\mathbf{F}_2$ of principal coordinates from a PCA of a chiPower-transformed compositional data matrix. Both $\mathbf{F}_1$ and $\mathbf{F}_2$ are assumed to have already been column-centred, which takes care of the translation operation, since this makes their sample means identically equal to the zero vector. The first step is then to normalize the two configurations so they both have sums of squares equal to 1, which takes care of the scaling. It just remains to find the rotation of one configuration to agree as closely as possible with the other, in the sense of

least-squared differences between them, which is where the SVD is used. Procrustes analysis and the computation of the Procrustes correlation proceed as follows.

1. Normalize both matrices:
$$\mathbf{F}_1^* = \mathbf{F}_1/\sqrt{\text{trace}(\mathbf{F}_1^\mathsf{T}\mathbf{F}_1)}$$
$$\mathbf{F}_2^* = \mathbf{F}_2/\sqrt{\text{trace}(\mathbf{F}_2^\mathsf{T}\mathbf{F}_2)} \tag{16}$$

2. Compute cross-products:
$$\mathbf{S} = \mathbf{F}_1^{*\mathsf{T}}\mathbf{F}_2^* \tag{17}$$

3. Perform SVD:
$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \tag{18}$$

4. Rotation matrix:
$$\mathbf{Q} = \mathbf{V}\mathbf{U}^\mathsf{T} \tag{19}$$

5. Sum-of-squared errors:
$$E = \text{trace}[(\mathbf{F}_1^* - \mathbf{F}_2^*\mathbf{Q})^\mathsf{T}(\mathbf{F}_1^* - \mathbf{F}_2^*\mathbf{Q})] \tag{20}$$

6. Procrustes correlation:
$$r = \sqrt{1 - E} \tag{21}$$

The Procrustes correlation can be equivalently computed by vectorizing the two matrices $\mathbf{F}_1^*$ and $\mathbf{F}_2^*\mathbf{Q}$, and computing the Pearson correlation between them.

The function `protest()` (Peres-Neto and Jackson 2001) in the R package `vegan` (Oksanen et al. 2019) computes the correlation as follows:

```
protest(A, B, permutations=0)$t0
```

where A and B are the matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ to be fitted to each other.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-024-00600-x.

Data and code availability The Rabbit and original Crohn data are available on https://github.com/michaelgreenacre/CODAinPractice where some R code is available to reproduce several of the analyses.

References

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

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London

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

te Beest D, Nijhuis E, Möhlmann T et al (2021) Log-ratio analysis of microbiome data with many zeroes is library size dependent. Mol Ecol Resour 21(6):1866–1874. https://doi.org/10.1111/1755-0998.13391

Benzécri JP (1973) L'Analyse des Données. L'Analyse des Correspondances. Dunod, Paris, Tôme II

Borg I, Groenen P (2010) Modern multidimensional scaling: theory and applications, 2nd edn. Springer, New York

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

Calle M, Urrea V, Boulesteix AL et al (2011) Auc-rf: a new strategy for genomic profiling with random forest. Hum Hered 72:121–32

Calle M, Pujolassos M, Susin A (2023) coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. BMC Bioinform 24:82

Choulakian V (2023) Some notes on correspondence analysis of power transformed data sets. arXiv:2301.01364

Coenders G et al (2023) 40 years after Aitchison's article "The statistical analysis of compositional data". Where we are and where we are heading. SORT 47:1–22

Coenders G, Greenacre M (2022) Three approaches to supervised learning for compositional data with pairwise logratios. J Appl Stat 49:1–22. https://doi.org/10.1080/02664763.2022.2108007

Erb I (2023) Power transformations of relative count data as a shrinkage problem. Inf Geom 6:327–354

Gower J, Dijksterhuis G (2004) Procrustes problems. Oxford University Press, New York

Greenacre M (1984) Theory and applications of correspondence analysis. Academic Press, London

Greenacre M (2009) Power transformations in correspondence analysis. Comput Stat Data Anal 53:3107–16

Greenacre M (2010) Log-ratio analysis is a limiting case of correspondence analysis. Math Geosci 42:129–34

Greenacre M (2011) Measuring subcompositional incoherence. Math Geosci 43:681–93

Greenacre M (2016) Correspondence analysis in practice, 3rd edn. Chapman & Hall / CRC Press, Boca Raton

Greenacre M (2018) Compositional data analysis in practice. Chapman & Hall / CRC Press, Boca Raton

Greenacre M (2019) Variable selection in compositional data analysis using pairwise logratios. Math Geosci 51:649–82

Greenacre M (2021) Compositional data analysis. Annu Rev Stat Appl 8:271–99

Greenacre M, Primicerio R (2010) Multivariate analysis of ecological data. BBVA Foundation, Bilbao

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

Greenacre M, Groenen P, Hastie T et al (2022) Principal component analysis. Nat Rev Methods Prim 2:101. https://doi.org/10.1038/s43586-022-00192-w

Greenacre M, Grunsky E, Bacon-Shone J et al (2023) Aitchison's compositional data analysis 40 years on: a reappraisal. Stat Sci 38:386–410. https://doi.org/10.1214/22-STS880

Grunsky E, Greenacre M, Kjarsgaard B (2024) GeoCoDA: recognizing and validating structural processes in geochemical data. A workflow on compositional data analysis in lithogeochemistry. Appl Comput Geosci. https://doi.org/10.48550/arXiv.2307.11084

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

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

Lubbe S, Filzmoser P, Templ M (2021) Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemom Intell Lab Syst 210:104248. https://doi.org/10.1016/j.chemolab.2021.104248

Lundborg AR, Pfister N (2023) Perturbation-based analysis of compositional data. arXiv: 2311.18501

Martínez-Álvaro M, Auffret M, Duthie CA et al (2022) Bovine host genome acts on specific metabolism, communication and genetic processes of rumen microbes host-genomically linked to methane emissions. Commun Biol 5:350. https://doi.org/10.1038/s42003-022-03293-0

Oksanen J, Blanchet F, Friendly M, et al (2019) vegan: community ecology package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

Palarea-Albaladejo J, Martin-Fernandez J (2015) zCompositions–R package for multivariate imputation of left-censored data under a compositional approach. Chemom Intell Lab Syst 143:85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Peres-Neto P, Jackson D (2001) How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129:169–178

Rayens W, Srinivasan C (1991) Box-Cox transformations in the analysis of compositional data. J Chemom 5:227–239

Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V et al (2018) Balances: a new perspective for microbiome analysis. Systems 3:e00053-18

Tsagris M, Preston S, Wood A (2016) Classification for compositional data using the α-transformation. J Classif 33:243–261

Yoo J, Sun Z, Greenacre M et al (2022) A guideline for the statistical analysis of compositional data in immunology. Commun Stat Appl Methods 29:453–469