Ecography (in press, 2018)

Box-Cox-chord transformations for community composition data prior to beta diversity analysis

Pierre Legendre[1,2] and Daniel Borcard[1]

[1] Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Qc, Canada H3C 3J7

E-mail addresses: Pierre.Legendre@umontreal.ca, Daniel.Borcard@umontreal.ca

[2] Corresponding author: Pierre.Legendre@umontreal.ca. Orcid ID: 0000-0002-3838-3305

Running title: Box-Cox-chord transformations for community data

Abstract — In studies of spatial or temporal beta diversity, community composition data, often containing many zeros, must be transformed in some way before they are analysed by multivariate methods of data analysis. Data are transformed to reduce the skewness of species distributions and make dissimilarities double-zero asymmetrical. Criteria have recently been proposed to determine which dissimilarity functions (or the corresponding data transformations) can be used for beta diversity assessment. The chord transformation is often used as the preliminary transformation for frequency data. When the Euclidean distance is computed on chord-transformed data, a chord dissimilarity matrix **D** is produced, which obeys the proposed criteria. The Hellinger transformation, i.e. the chord transformation applied to square-root transformed frequencies, is also often used with community composition data prior to multivariate analyses; it leads to the Hellinger dissimilarity, which is another widely used **D** function in beta diversity studies. Among the data transformations often used in simple or multivariate data analysis, the Box-Cox method provides a useful series of transformations to make data distributions more symmetrical, where exponent 1 is the absence of a transformation, exponent 0.5 is the square-root, exponent 0.25 is the fourth-root, and the log transformation is the limit of the Box-Cox function corresponding to exponent 0. Combining the two previous ideas, this paper proposes to combine any transformation of the Box-Cox family with exponent in the [0,1] range with the chord transformation. In particular, one can compute the $\log_e(y+1)$ transformation of a community composition (or other frequency) data table and follow with a chord transformation. A **D** matrix can be computed from the doubly-transformed data. The transformations and **D** functions in that family inherit the properties of the chord dissimilarity, and this ensures that they all follow the necessary criteria for beta diversity assessment that have recently been proposed.

Keywords: Box-Cox-chord family, double-zero asymmetrical coefficients, skewness reduction

Introduction

Before analysis by linear methods such as PCA, RDA and *k*-means clustering, community composition data, which often contain lots of zeros, should be transformed to satisfy two objectives. The first is to reduce the skewness of the data distributions. The second objective is to make the dissimilarity, preserved during the analysis, double-zero asymmetrical. These two conditions are explained in the first two sections of this paper.

Satisfying these two conditions can be achieved by a family of transformations that we are proposing to call the Box-Cox-chord family, which includes some of the transformations widely used by ecologists. These transformations are the first step before computing dissimilarity indices. While investigating this family of transformations, a new approach emerged, the log-chord transformation, which fills a gap in the array of transformations that ecologists can use for the analysis of community data.

Log-chord transformed data can be used to compute a new dissimilarity function, the log-chord *D* coefficient. It combines the log transformation, which belongs to the Box-Cox family, and the chord distance. The interest of this new *D* index for ecological analysis will be discussed.

Skewness reduction

Community composition data are notorious for their highly skewed distributions. Artificial data with lognormal distributions, rounded to integers, have long been used to represent community composition data in simulation studies (e.g. Gauch and Whittaker 1972) because their skewness is comparable to that often encountered in real community data (Preston 1948). This is the reason why real community data are often log-transformed before analysis, especially when using linear

models such as principal component analysis (PCA) and canonical redundancy analysis (RDA). Reducing data skewness helps disperse clouds of points in ordination and other methods of multivariate analysis, thus facilitating visualization and interpretation.

Not all ecological data are as strongly skewed as lognormal data, but for data distributions that are increasingly skewed to the right, the Box-Cox method (Box and Cox 1964) proposes a continuum of transformations that goes from exponent 1 (no transformation) to 0.5 (square root) to 0.25 (fourth root) to 0 (log transformation). The complete equation of the Box-Cox transformation is: $f(y, \lambda) = (y^{\lambda} - 1)/\lambda$; the limit of this equation as λ approaches 0 is $\log_e(y)$; (Box and Cox 1964).

Double-zero asymmetrical *D* coefficients

Ecologists have long known that untransformed (site × species) community composition data should not be analysed using the Euclidean distance, or through methods of multivariate analysis that preserve the Euclidean distance, such as principal component analysis and redundancy analysis, where calculations are carried out in Euclidean space. This is the case, for example, with *k*-means clustering, where the sums of squared deviations from the group centroids are computed using Euclidean distances. Clear demonstrations of the difficulties encountered when using the Euclidean distance, or methods that preserve it, for the analysis of community composition data have been given by Orlóci (1978), Legendre and Gallagher (2001) and Legendre and Legendre (2012). These authors based their demonstrations on the fact that the Euclidean distance was not a *double-zero asymmetrical* coefficient, which led to contradictory results and aberrant distance matrices when applied to real or simulated community composition data sets. For example, the Euclidean distance may produce *D* values indicating that sites with no species in common are closer (i.e. less dissimilar) to each other than sites that share most or all of their species.

On the contrary, in coefficients that have the property of *double-zero asymmetry*, the dissimilarity *D* does not change with the addition of double-zeros to the comparison of two sites, but it *decreases* when double-X are added, where X is any value of equal abundances other than zero. This is a formalization of the idea, known for a long time by ecologists, that the absence of a species from two sites cannot be interpreted as an indication of similarity between the sites, whereas equal abundances for a species that is present at two sites should make the dissimilarity smaller.

The Euclidean distance is also missing other properties that are necessary for sound assessment of beta diversity. Besides the double-zero asymmetry property described above, the missing necessary properties are the following (Legendre and De Cáceres 2013): (1) pairs of sites without species in common should have the largest *D* value; (2) species replication invariance, i.e. if the community composition data are repeated in two or several copies, the **D** matrix should be identical to the original **D** matrix; (3) invariance to measurement units; for example, the **D** matrix should be the same if biomass data are expressed in g or in kg; (4) a *D* function should have a fixed upper bound, D_{max}. The 14 properties investigated by Legendre and De Cáceres (2013) over 16 dissimilarity coefficients are described in Appendix 3 of their paper.

Examples of double-zero asymmetrical coefficients, listed in Legendre and De Cáceres (2013), are the chord (especially important in this paper, see next section), Hellinger, percentage difference (aka Bray-Curtis), Canberra, Whittaker, divergence, Wishart, and Kulczynski dissimilarity coefficients, plus their binary counterparts, which are the Jaccard, Sørensen and Ochiai dissimilarity indices. Legendre (2014) also found that the quantitative Ružička dissimilarity, whose binary form is the Jaccard *D* coefficient, had this property. The *double-zero asymmetry* property depends on the presence of a denominator in the formula of a coefficient,

which induces an upper bound for the *D* value and ensures that double-zeros (that do not change the denominator) and double-X values (that do) are treated differently. The Euclidean and Manhattan distance formulas, for example, have no denominator; for that reason, they do not have fixed upper bounds and they are double-zero symmetrical.

The Box-Cox-chord family of transformations

The chord distance has been known for a long time. The chord of a circle was described in Book III of Euclid's *Elements* (in ancient Greek: Στοιχεῖα, Stoicheia) about 300 BC. The chord distance was used by Hipparchus of Nicaea (190 to 120 BC) and, later, by Claudius Ptolemy (100 to 170 AD) to compute the first trigonometric tables, which were applied by these astronomers to calculate the motion of celestial bodies. The chord distance was first applied to community composition data by Orlóci (1967).

The same year, Cavalli-Sforza and Edwards (1967) proposed a variant of the chord distance for the analysis of genetic data, where the relative frequencies are square-rooted before being used in the chord distance formula. The result is not the chord distance of Orlóci (1967) but the Hellinger distance, often used by ecologists, as shown below. To our knowledge, this difference has never been mentioned in the literature. During the past 50 years, ecologists and geneticists may have thought that the chord distances they were using were the same, but that was not the case. In the version of the chord distance described in their paper, Cavalli-Sforza and Edwards (1967) further divide the computed distances by $\pi/2$. This transformation facilitates interpretation of the distances in terms of gene substitution.

The Box-Cox family of transformations gave us the idea to compute the chord distance on community composition data that have been transformed with various exponents of the Box-Cox family. Here is the reasoning:

1. Untransformed community data are the same as data transformed with exponent 1. The chord distance computed on untransformed data is thus equivalent to computing that distance on data transformed with exponent 1.

2. The Hellinger distance is the chord distance applied to square-root transformed data, where the square root is exponent 0.5 in the Box-Cox family of transformations. The demonstration is the following: (1) the chord transformation formula is $y'_{ij} = y_{ij} / \sqrt{\sum_{j=1}^{p} y_{ij}^2}$; (2) let us replace y_{ij} in the formula by square-rooted data $z_{ij} = \sqrt{y_{ij}}$; (3) the chord transformation formula becomes $y''_{ij} = z_{ij} / \sqrt{\sum_{j=1}^{p} z_{ij}^2}$; hence (3), $y''_{ij} = \sqrt{y_{ij}} / \sqrt{\sum_{j=1}^{p} y_{ij}} = \sqrt{y_{ij} / \sum_{j=1}^{p} y_{ij}}$, which is the formula of the Hellinger transformation. The chord and Hellinger dissimilarities are obtained by computing the Euclidean distance from chord and Hellinger transformed data, respectively. Taking the square root is thus a possibility before computing the chord distance. The square root reduces the asymmetry of modestly asymmetric data distributions before subjecting data to linear methods of analysis.

3. Following this line of reasoning, why not compute the chord distance on community composition data transformed with any exponent in the [0,1] interval? This paper describes the resulting log-chord transformation and distance.

The log-chord transformation and distance

For data that are strongly asymmetrical, the Box-Cox family suggests that one could log-transform the data before computing the chord distance. The log transformation is the limit of the Box-Cox function when the exponent tends to 0. The novel idea of the present note is to use a chord transformation, or compute the chord distance, after having log-transformed the data.

Ecologists often apply the $y' = \log_e(y+1)$ transformation to community composition data after adding the constant 1, because species frequency data contain zeros and log(0) = –Infinity. If $y = 0$, the $y' = \log_e(y+1)$ function returns the value $y' = 0$. The log base is of no importance as long as all data in the matrix are transformed using logarithms with the same base. However, when subjecting log(y+1) transformed data to multivariate linear methods of analysis that preserve the Euclidean distance, the analysis implements the Euclidean distance, which is an inappropriate dissimilarity for community composition data, as shown in the section on *Double-zero asymmetrical D coefficients*. The main problems with this distance are that it does not have an upper bound and is double-zero symmetrical even for $\log_e(y+1)$ transformed data.

The log-chord distance possesses the 9 necessary properties described by Legendre and De Cáceres (2013) and would thus be appropriate for beta diversity studies. This *D* coefficient combines two objectives: the log transformation makes the species distributions more symmetric, reducing the importance of the very abundant species, whereas the chord transformation produces a double-zero asymmetrical *D* coefficient, which can be used in beta diversity studies. This combination represents a fully justified use of log-transformed data in community analysis. In the computation of the log-chord distance matrix, since the chord distance is computed after the log transformation, the **D** matrix inherits all properties of the chord distance (see Legendre and De Cáceres 2013, Table 2) that these authors considered important for beta diversity studies.

When applied to presence-absence data, the chord, Hellinger and log-chord distances produce the Ochiai dissimilarity, or more precisely: $D_{Hellinger} = D_{\text{chord}} = D_{log\text{-}chord} = \sqrt{2}\sqrt{1 - S_{Ochiai}}$ where S_{Ochiai} is the Ochiai similarity index (Legendre & Legendre 2012). Transformation of presence-absence data using any exponent, followed by a chord transformation and calculation of the Euclidean distance, produces the same result. All these dissimilarities have a maximum value of $\sqrt{2}$, reached when two sites have completely different community compositions. By opposition, for untransformed presence-absence community data, the Euclidean distance produces $\sqrt{p(1 - Simple\ matching\ similarity)}$ where p is the number of species; it does not produce the Ochiai dissimilarity.

There were, up to now, four known ecological dissimilarities that could be computed in two steps, where the first step is a transformation of the data and the second is the calculation of the Euclidean distance. These transformations have been described by Legendre and Gallagher (2001) and are implemented in computer software, including the decostand() function of the *vegan* package in R. The chord transformation is one of them. It consists of dividing each frequency value in a row (site) vector by the norm, or length, of that vector. The Hellinger transformation is another; it is obtained by dividing each value in a row vector by the row sum, then taking the square root of the relative value. As shown above, the Hellinger transformation can also be obtained by taking the square root of the raw abundance data ($y^{0.5}$) and applying the chord transformation to the square-rooted values. The last two dissimilarities in that group are based upon the profile and chi-square transformations; they do not produce distances that are fully appropriate for beta diversity studies (Legendre and De Cáceres 2013, Table 2) and will not be discussed further here.

The log-chord transformation is obtained as follows:

1. Log-transform the raw abundance data, as explained in the previous section:

$$y'_{ij} = \log (y_{ij} + 1)$$

2. Compute a chord transformation of the log-transformed data y'_{ij}:

$$\textit{log-chord transformation}: y^{"}_{ij} = y'_{ij} / \sqrt{\sum_{j=1}^{p} (y'_{ij})^2}$$

This double transformation is computed by function box.cox.chord(); see the Discussion section below, as well as Appendix 1 (example 1) and Appendix 4. One can stop at this point and directly use the transformed data $y^{"}_{ij}$ as input into linear methods of analysis, such as principal component analysis (PCA), redundancy analysis (RDA) or *k*-means clustering, that preserve the Euclidean distance. The results of these analyses will preserve the log-chord distance among sites instead of the Euclidean distance. Note that the transformed data do not vary with the logarithm base because log-transformed values y'_{ij} are found in both the numerator and the denominator.

Else, if a dissimilarity matrix is needed, for example for clustering, one can compute the Euclidean distance from the log-chord-transformed data to obtain a matrix of log-chord dissimilarities. See Appendix 1, examples 1, 2 and 3.

Or else, in studies of beta diversity, one can pass the untransformed abundance data to the beta.div() function of the *adespatial* package and run it with method = "log.chord" to compute total beta diversity (BD_{Total}), which is then the variance of the transformed multivariate input data.

The chord distance, and hence also the new log-chord distance, are distances in the strict sense, i.e. metric dissimilarities, since these *D* indices have the four properties of a metric: minimum of 0 (if $\mathbf{x}_1 = \mathbf{x}_2$, then $D(\mathbf{x}_1,\mathbf{x}_2) = 0$), positiveness (if $\mathbf{x}_1 \neq \mathbf{x}_2$, then $D(\mathbf{x}_1,\mathbf{x}_2) > 0$), symmetry ($D(\mathbf{x}_1,\mathbf{x}_2) =$

$D(\mathbf{x}_2,\mathbf{x}_1)$), and triangle inequality ($D(\mathbf{x}_1,\mathbf{x}_2) + D(\mathbf{x}_2,\mathbf{x}_3) \geq D(\mathbf{x}_1,\mathbf{x}_3)$) for any three objects $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$. However, several of the measures used in ecology, including those that are appropriate for beta diversity studies, are not metric, so we use the more general term *dissimilarity* to refer to measures of resemblance between site vectors.

Discussion: the Box-Cox-chord transformation

1. Why would we need another dissimilarity function? A log transformation reduces the asymmetry of the species distributions, but it does not produce a distance that is double-zero asymmetrical. We need to go one step further and compute a chord transformation of the log-transformed data, as described in the previous section, to obtain a double-zero asymmetrical **D** matrix, which is one of the important conditions to carry out beta diversity assessment; see Legendre and De Cáceres (2013). The log-chord transformation represents a correct way of using the log transformation in beta diversity analysis of community composition data.

2. Different exponents of the Box-Cox series can be used before the chord transformation to produce different forms of transformed data and dissimilarity matrices that are all appropriate for beta diversity assessment. We call this combination the *Box-Cox-chord transformation*. Any exponent in the [0,1] interval can be used. The regular chord transformation corresponds to Box-Cox exponent 1 and leaves the frequency distributions of the data unchanged. The Hellinger transformation includes a square-root transformation of the data and corresponds to Box-Cox exponent 0.5. The double square root (exponent 0.25) is sometimes used by ecologists. The log-chord transformation described in this paper corresponds to Box-Cox exponent $\lambda = 0$. The log transformation offers the possibility of normalizing more strongly asymmetrical frequency distributions than the square-root or fourth-root transformations.

232 We wrote a function capable of implementing a transformation with any exponent, followed by
233 chord transformation of the data. That function, called box.cox.chord(), is available as a text file
234 in Appendix 4 of the Supplementary material. Examples are presented in Appendix 1. Ecologists
235 usually prefer using exponents in the series {0, 0.25, 0.5, 1} because these correspond to
236 conventional ways of transforming community composition data with various degrees of
237 asymmetry. We also modified functions dist.ldc() and beta.div() of *adespatial* and included the
238 log-chord transformation among the available methods. The functions can now compute the
239 chord, Hellinger and log-chord distances, which correspond, respectively, to exponents 1, 0.5,
240 and 0 of the Box-Cox transformation followed by computation of the chord transformation.

241 3. Users may wish to select the exponent that produces the largest probability of obtaining the
242 (untransformed or transformed) observed data as a sample from a statistical population with
243 multinormal distribution. We wrote another R function that transforms the data using a selection
244 of exponents (chosen by the user), then carries out Dagnelie's (1975) test of multinormality. The
245 function, called BCD(), is presented as a text file (Appendix 5). Dagnelie's test of multinormality
246 is described in Appendix 3. That Appendix also reports the results of a simulation study
247 indicating when the test has correct rates of type I error.

248 4. The Dagnelie test of multivariate normality requires that $n > (rank+1)$ where n is the number of
249 observations and *rank* is the rank of the column-centred data matrix, or equivalently the rank of
250 its covariance matrix. The rank of a covariance matrix is $rank = \min((n-1), p)$ unless collinearity
251 among the p variables further reduces the rank. When n is smaller than $(rank+2)$, the Dagnelie
252 test cannot be computed because all Mahalanobis distances to the multivariate centroid are equal.
253 In that case, one could determine empirically which transformation of the community
254 composition data produces the highest adjusted R-square, or the lowest AIC_c, in redundancy

255 analysis (RDA) of the species data **Y** against explanatory (e.g. environmental) variables **X** of
256 interest. The selected exponent could be used to transform community data prior to linear
257 analyses such as RDA. This approach does not aim at normalizing the community data but at
258 optimizing the linear relationships between the transformed data and the explanatory variables.

259 Appendix 2 presents transformation results for 7 data sets; six of them are multivariate data from
260 the community ecology literature and the 7th is a simulated data set. In each case, we transformed
261 the count data using exponents from 0 to 1 by steps of 0.1, plus exponent 0.25 which corresponds
262 to the double square root, a transformation occasionally found in the ecological literature. The
263 data raised to each exponent were then chord-transformed. All these data were submitted to a
264 Dagnelie test of multivariate normality to determine which version provided the largest
265 probability of obtaining the observed data as a sample from a statistical population with
266 multinormal distribution: data with exponent transformation only or exponent plus chord
267 transformation. Analysis shows that different data sets may be best transformed using any one of
268 the exponents in the [0,1] range under investigation here. Using this strategy will produce data
269 that are closer to multivariate normality. This may, in turn, lead to better analyses by Euclidean-
270 based linear methods like PCA, RDA and *k*-means clustering.

271 5. When the community composition data are analysed by RDA with explanatory variables, a
272 better approach would be to apply in turn various values of the Box-Cox exponent to the
273 community data, compute the RDA, and test the normality of the residuals. A computer function
274 could be written to carry out these analyses in a loop for a series of values of the Box-Cox
275 exponent. One could then select the transformation that produces RDA residuals closest to
276 normality.

Selecting the best normalizing transformation in a linear modelling situation is easy for univariate data, using function boxcox() of the MASS package in R. That function looks for the Box-Cox exponent that maximizes the log-likelihood described by Box & Cox (1964), yielding the best transformation of the model residuals to meet univariate normality. That function is only applicable to analyses conduced with the linear modelling functions lm() or aov() of R, though. It cannot be applied to multivariate data and to RDA.

Acknowledgements – We are thankful to Associate Editor Luis Mauricio Bini and two anonymous reviewers for interesting questions and suggestions during evaluation of the manuscript. This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant No. 7738 to P. Legendre.

References

Box, G. E. P. and Cox, D. R. 1964. An analysis of transformations. J. Roy. Statist. Soc. Ser. B 26: 211–243.

Cavalli-Sforza, L. L. and Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21: 550–570.

Dagnelie, P. 1975. L'analyse statistique à plusieurs variables. Les Presses agronomiques de Gembloux, Gembloux (Belgium).

Gauch, H. G. Jr. and Whittaker, R. H. 1972. Coenocline simulation. Ecology 53: 446–451.

Legendre, P. 2014. Interpreting the replacement and richness difference components of beta diversity. Global Ecol. Biogeogr. 23: 1324–1334.

297 Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data:
298 dissimilarity coefficients and partitioning. Ecol. Lett. 16: 951–963.

299 Legendre, P. and Gallagher, E. D. 2001. Ecologically meaningful transformations for ordination
300 of species data. Oecologia 129: 271–280.

301 Legendre, P. and Legendre, L. 2012. Numerical ecology, 3rd English ed. Elsevier Science BV,
302 Amsterdam.

303 Orlóci, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. 55:
304 193–206.

305 Orlóci, L. 1978. Multivariate analysis in vegetation research. 2nd edition. Dr. W. Junk B. V., The
306 Hague.

307 Preston, F. W. 1948. The commonness, and rarity, of species. Ecology 29: 254–283.

308

Supplementary material to:

Legendre, P. and Borcard, D. 2018. Box-Cox-chord transformations for community composition data prior to beta diversity analysis. *Ecography* (in press).

Appendix 1

Example of log-chord transformation of multivariate data

This Appendix uses an example to describe the transformation of a data matrix into log(y+1) followed by the chord transformation. The *log-chord transformed data* can then be used directly as input into linear methods of analysis such as PCA, RDA or k-means partitioning, or used to compute a log-chord dissimilarity matrix **D**. Three calculation methods are presented.

The example uses a small subset of the mite data (70 site × 35 morpho-species). The transformation is computed using the box.cox.chord() function found in the text file "box.cox.chord.R" (Appendix 4). The soil mite data, collected by D. Borcard, were first used in the paper of Borcard et al. (1992). They have been used as test data in a number of methodological papers and are available in the *vegan* R package. Package *adespatial*, available on CRAN, is required to run examples 2 and 3.

```
library(vegan)
data(mite)
### Select a small data set for the following calculations, so that
### the resulting D matrix will be small and easy to examine.
(mite.small <- mite[1:5,1:8])

### Example 1: transformation using function box.cox.chord() (App. 4)
### Proof of concept for computation of the log.chord dissimilarity:
(mite.new.tr <- box.cox.chord(mite.small, bc.exp=0))
###
### Compute the log.chord D matrix using function dist() of {base}
(mite.log.chord.D <- dist(mite.new.tr))

### Example 2: compute the log.chord D matrix using dist.ldc
library(adespatial)
(mite.log.chord.D.2 <- dist.ldc(mite.small, "log.chord"))

### Example 3: compute the log.chord D matrix in two steps
### Log-transform the data, then compute
### the chord dissimilarity matrix with dist.ldc() of adespatial
tmp <- log(mite.small + 1)
library(adespatial)
(mite.log.chord.D.3 <- dist.ldc(tmp, "chord"))
```

Reference

Borcard, D., Legendre, P. and Drapeau, P. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045–1055.

Appendix 2

Examples of search for best Box-Cox transformation

This Appendix shows examples of Box-Cox transformations, with and without subsequent chord transformation, of real and simulated multivariate data matrices, followed by tests of multivariate normality.

1. **Mite data** (70 × 35), *vegan* package: best transformations obtained with exponents 0.7 and 0.8.

2. **Mite data, panphytophagous species** (70 × 23), *vegan* package: best transformations obtained with exponents 0.6 and 0.8.

3. **Mite data, microphagous species** (70 × 12), *vegan* package: best transformations obtained with exponents 0.2 and 0.5.

4. **Spider data** (28 × 12) available from http://adn.biol.umontreal.ca/~numericalecology/data/: best transformations obtained with exponents 0.2 and 0.1.

5. **Ichtyo data** (32 × 9), *ade4* package: best transformations obtained with exponents 0.2 and 1.0.

6. **Baran95 data** (95 × 33), *ade4* package: best transformations obtained with exponent 0.1.

7. **Simulated random lognormal data**: best transformations obtained with exponents 0.1 and log.

These examples show that different data sets are best transformed using different exponents, followed by the chord transformation. Users should preferably choose the combination of "exponent plus chord transformation" that yields the most normal data when this is the data set that will be analysed by methods of multivariate analysis.

One may prefer to use the transformation that maximizes the multivariate normality of the data before or after the chord transformation.

The BCD.R function used in these analyses is displayed in Appendix 5.

Output details of the BCD.R function for the 7 data sets follow. The function output is a table showing the Box-Cox exponent in the first column of each row. In columns 2 and 3, one finds the Shapiro-Wilk *W* statistic (BC_W) of the Dagnelie test of multivariate normality and associated p-value (BC_p-val) after the exponent has been applied to the original data. Columns 4 and 5 show the same statistics (BC.chord_W and BC.chord_p-val) after the chord transformation has been applied to the Box-Cox transformed data.

In each table of results, the statistics corresponding to the most normal data sets after Box-Cox and Box-Cox+chord transformations are **in bold**.

1. Mite data, full data matrix

The mite data are available in package *vegan*

```
library(vegan)
data(mite)

( out1 = BCD(mite, chord=TRUE) )
      BC.exp      BC_W   BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9603061 0.02602953  0.9419801    0.002819179
 [2,]   0.10 0.9530738 0.01053773  0.9610871    0.028754177
 [3,]   0.20 0.9583455 0.02030527  0.9569133    0.016960910
 [4,]   0.25 0.9568211 0.01676630  0.9504914    0.007693903
 [5,]   0.30 0.9603525 0.02618375  0.9539955    0.011802819
 [6,]   0.40 0.9682930 0.07296080  0.9553200    0.013904805
 [7,]   0.50 0.9697667 0.08841214  0.9583125    0.020220975
 [8,]   0.60 0.9705510 0.09792945  0.9645633    0.044949363
 [9,]   0.70 0.9731090 0.13657654  0.9712284    0.106965366 <= Best Box-Cox
[10,]   0.80 0.9727516 0.13039029  0.9745991    0.165558765 <= Best Box-Cox+chord
[11,]   0.90 0.9695962 0.08646857  0.9741643    0.156542433
[12,]   1.00 0.9652421 0.04907666  0.9726766    0.129127979

dim(mite)
[1] 70 35
```

Note — n < 3*p, hence the test is too liberal. If H_0 is not rejected in this situation (e.g. when p > 0.05), the result is trustworthy; the hypothesis of normality cannot be rejected.

---

2. Mite data, 23 panphytophagous species

The mite data are available in package *vegan*

```
library(vegan)
data(mite)
```

The 23 panphytophagous species are:

```
panphyto <- c(2:6,12,13,16:22,24:29,31:33)

( out2 = BCD(mite[,panphyto], chord=TRUE) )
      BC.exp      BC_W    BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9707035 0.099896315  0.9554181     0.01407531
 [2,]   0.10 0.9509091 0.008092884  0.9677674     0.06813249
 [3,]   0.20 0.9619885 0.032267774  0.9668447     0.06042394
 [4,]   0.25 0.9671025 0.062484201  0.9646912     0.04569879
 [5,]   0.30 0.9669149 0.060978084  0.9613699     0.02981177
 [6,]   0.40 0.9665992 0.058526060  0.9549857     0.01333983
 [7,]   0.50 0.9743179 0.159671920  0.9569023     0.01693773
 [8,]   0.60 0.9790461 0.290016669  0.9610209     0.02851221 <= Best Box-Cox
 [9,]   0.70 0.9753870 0.183171576  0.9652367     0.04904228
[10,]   0.80 0.9700347 0.091556520  0.9654409     0.05035754 <= Best Box-Cox+chord
[11,]   0.90 0.9644316 0.044190770  0.9624732     0.03433710
[12,]   1.00 0.9567580 0.016634525  0.9582571     0.02008019

dim(mite[,panphyto])
[1] 70 23
```

Note — n > 3*p: the test has correct type I error. At the 0.05 significance level, if p > 0.05, the hypothesis of multivariate normality cannot be rejected.

---

3. Mite data, 12 microphagous species

The mite data are available in package *vegan*

```
library(vegan)
data(mite)
```

The 12 microphagous species are:

```
microphyto <- c(1,7:11,14,15,23,30,34,35)

( out3 = BCD(mite[,microphyto], chord=TRUE) )
      BC.exp      BC_W     BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9389853 0.0020038883  0.9525446   0.0098762913
 [2,]   0.10 0.9694429 0.0847579427  0.9137939   0.0001421749
 [3,]   0.20 0.9820120 0.4123702927  0.9389190   0.0019889337 <= Best Box-Cox
 [4,]   0.25 0.9791737 0.2945817873  0.9474531   0.0053454700
 [5,]   0.30 0.9768387 0.2203113404  0.9535511   0.0111742005
 [6,]   0.40 0.9542870 0.0122350834  0.9605179   0.0267410108
 [7,]   0.50 0.9473766 0.0052970920  0.9623839   0.0339459038 <= Best Box-Cox+chord
 [8,]   0.60 0.9433451 0.0033005076  0.9561124   0.0153456846
 [9,]   0.70 0.9377345 0.0017407543  0.9554642   0.0141560551
[10,]   0.80 0.9313931 0.0008663729  0.9558659   0.0148815865
[11,]   0.90 0.9255321 0.0004650655  0.9538361   0.0115731823
[12,]   1.00 0.9180240 0.0002159836  0.9489935   0.0064242512

dim(mite[,microphyto])
[1] 70 12
```

Note — n > 3*p and < 7.5*p: the test has correct type I error. At the 0.05 significance level, if p > 0.05, the hypothesis of multivariate normality cannot be rejected.

---

4. Spider data

Available from http://adn.biol.umontreal.ca/~numericalecology/data/

```
( out4 = BCD(spiders, chord=TRUE) )
      BC.exp      BC_W    BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9181606 0.031245851  0.9804109     0.86013321
 [2,]   0.10 0.9685690 0.542784991  0.9819122     0.89345046 <= Best Box-Cox+chord
 [3,]   0.20 0.9762135 0.752131082  0.9587561     0.32565132 <= Best Box-Cox
 [4,]   0.25 0.9715565 0.622891779  0.9544723     0.25610404
 [5,]   0.30 0.9470945 0.167369454  0.9708013     0.60223169
 [6,]   0.40 0.9046355 0.014714676  0.9700679     0.58240671
 [7,]   0.50 0.9265783 0.050572929  0.9470717     0.16714642
 [8,]   0.60 0.9319738 0.069146153  0.9433082     0.13416276
 [9,]   0.70 0.9321922 0.070030999  0.9378295     0.09734019
[10,]   0.80 0.9225437 0.040105550  0.9348596     0.08182024
[11,]   0.90 0.9084913 0.018187368  0.9314085     0.06690839
[12,]   1.00 0.8941633 0.008373844  0.9266658     0.05082874
```

```
dim(spiders)
[1] 28 12
```

Note — n < 3*p, hence the test is too liberal. If H_0 is not rejected in this situation (e.g. when p > 0.05), the result is trustworthy; the hypothesis of normality cannot be rejected.

---

5. ichtyo data

The data are available in package *ade4*

```
library(ade4)
data(ichtyo)

( res.ich = BCD(ichtyo$tab, chord=TRUE) )
      BC.exp      BC_W   BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9772254 0.71584034  0.9663527     0.40528267
 [2,]   0.10 0.9509547 0.15340171  0.9404767     0.07723569
 [3,]   0.20 0.9848224 0.92117555  0.9565073     0.21990253 <= Best Box-Cox
 [4,]   0.25 0.9830138 0.88086210  0.9606957     0.28695288
 [5,]   0.30 0.9726086 0.57430371  0.9607060     0.28713777
 [6,]   0.40 0.9700021 0.49945594  0.9586042     0.25144776
 [7,]   0.50 0.9664837 0.40842375  0.9496523     0.14088281
 [8,]   0.60 0.9482464 0.12849457  0.9407500     0.07862820
 [9,]   0.70 0.9300625 0.03931990  0.9438962     0.09661361
[10,]   0.80 0.9225076 0.02432857  0.9560140     0.21303551
[11,]   0.90 0.9212837 0.02252885  0.9638683     0.34916082
[12,]   1.00 0.9228731 0.02489464  0.9655391     0.38617609 <= Best Box-Cox+chord

dim(ichtyo$tab)
[1] 32  9
```

Note — n > 3*p and < 7.5*p: the test has correct type I error. At the 0.05 significance level, if p > 0.05, the hypothesis of multivariate normality cannot be rejected.

---

6. baran95 data

The data are available in package *ade4*

```
library(ade4)
data(baran95)

( res.baran95 = BCD(baran95$fau, chord=TRUE) )
      BC.exp      BC_W   BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9894593 0.65611464  0.9794748    0.141788600
 [2,]   0.10 0.9946292 0.97084927  0.9868997    0.468508483 <= Best Box-Cox
 [3,]   0.20 0.9938864 0.94596086  0.9806581    0.173362783    and Box-Cox+chord
 [4,]   0.25 0.9936129 0.93448592  0.9779096    0.108437094
 [5,]   0.30 0.9914625 0.80734280  0.9769826    0.092449482
 [6,]   0.40 0.9884609 0.57991662  0.9794200    0.140470587
 [7,]   0.50 0.9879013 0.53860631  0.9858741    0.402931113
 [8,]   0.60 0.9874182 0.50411127  0.9854637    0.378614164
 [9,]   0.70 0.9860602 0.41432926  0.9816065    0.203342446
[10,]   0.80 0.9836253 0.28347171  0.9752519    0.068613607
```

```
[11,]   0.90 0.9786790 0.12374654  0.9678170    0.019352968
[12,]   1.00 0.9723297 0.04153992  0.9602194    0.005604097

dim(baran95$fau)
[1] 95 33
```

Note — n < 3*p, hence the test is slightly too liberal. If H_0 is not rejected in this situation (e.g. when p > 0.05), the result is trustworthy; the hypothesis of normality cannot be rejected.

---

7. Random lognormal species-like data

Generate a (100×20) matrix of lognormally distributed integers:

```
n=100; p=20
mat2 <- matrix(round(exp(rnorm((n*p),mean=0,sd=2.5))),n,p)
```

How many zeros are found in that data matrix?

```
length(which(mat2==0))          # How many zeros = 803
length(which(mat2==0))/(n*p)    # Proportion of zeros = 0.402

( out6 = BCD(mat2, chord=TRUE) )
      BC.exp      BC_W     BC_p-val BC.chord_W BC.chord_p-val
 [1,]   0.00 0.9804151 1.431889e-01  0.9859730   3.723374e-01 <= Best Box-Cox+chord
 [2,]   0.10 0.9949069 9.726818e-01  0.9836121   2.511896e-01 <= Best Box-Cox
 [3,]   0.20 0.9805744 1.473243e-01  0.9831265   2.310064e-01
 [4,]   0.25 0.9774183 8.349209e-02  0.9855276   3.464196e-01
 [5,]   0.30 0.9686773 1.752362e-02  0.9822390   1.978718e-01
 [6,]   0.40 0.9460294 4.585089e-04  0.9695589   2.045315e-02
 [7,]   0.50 0.9264652 3.153650e-05  0.9602401   4.183005e-03
 [8,]   0.60 0.9063748 2.855237e-06  0.9496614   7.881005e-04
 [9,]   0.70 0.8867167 3.531224e-07  0.9382393   1.508004e-04
[10,]   0.80 0.8664618 5.060082e-08  0.9246952   2.519923e-05
[11,]   0.90 0.8471390 9.310615e-09  0.9058429   2.689940e-06
[12,]   1.00 0.8286615 2.082671e-09  0.8869547   3.617015e-07

dim(mat2)
[1] 100  20
```

Note — n > 3*p and < 7.5*p: the test has correct type I error. At the 0.05 significance level, if p > 0.05, the hypothesis of multivariate normality cannot be rejected.

==========

Appendix 3

The Dagnelie test of multivariate normality

Dagnelie (1975) proposed an elegant way of testing the multivariate normality of a set of multivariate observations. This Appendix describes the method.

The Dagnelie method is based on the Mahalanobis generalized distance. Generalized distances are computed, in multivariate space, between each object and the multivariate mean of all objects. The distance between object $\mathbf{y}_i$ and the mean point $\bar{\mathbf{y}}$ is computed as:

$$D(\mathbf{y}_i, \bar{\mathbf{y}}) = \sqrt{\mathbf{y}_{\mathrm{c}.i}\, \mathbf{S}^{-1} \mathbf{y}_{\mathrm{c}.i}{}'} \qquad \text{eq. 1}$$

where $\mathbf{y}_{\mathrm{c}.i}$ is row vector *i* in the matrix of column-centred data and **S** is the multivariate variance-covariance matrix. For standardized variables $\mathbf{z}_i$, eq. 1 becomes:

$$D(\mathbf{y}_i, \bar{\mathbf{y}}) = \sqrt{\mathbf{z}_i\, \mathbf{R}^{-1} \mathbf{z}_i{}'} \qquad \text{eq. 2}$$

where **R** is the correlation matrix. Dagnelie's approach is that, for multinormal data, the generalized distances should be normally distributed. He suggested to visually examine the cumulative frequency distribution and determine if the distribution of distances seemed normal. Actually, the generalized distances can be subjected to a Shapiro-Wilk test of normality, whose conclusions are applied to the multinormality of the original multivariate data. This is our improvement of Dagnelie's method.

The Dagnelie test of multinormality requires that $n > (rank+1)$, where n is the number of observations and *rank* is the rank of the column-centred data matrix, or equivalently the rank of its covariance matrix. The rank of a covariance matrix is $\min((n-1), p)$ unless collinearity among the p variables further reduces the rank. When the covariance matrix is not of full rank, its inverse can still be computed using generalized inversion through singular value decomposition (SVD). However, when n is smaller than $(rank+2)$, the Dagnelie test cannot be computed because all Mahalanobis distances to the multivariate centroid are all equal.

Numerical simulations conducted by one of us (D. Borcard) for type I error, using normal random deviates, showed the following:

• The Dagnelie test of normality, based on the Shapiro-Wilk test of Mahalanobis generalized distances, is not meant to be used with univariate data; in simulations conducted with univariate data, the type I error rate was higher than the significance level for all values of n. The Shapiro-Wilk test of univariate normality should be used in that case.

• The test had correct levels of type I error for values of n between $3p$ and $7.5p$, where n is the number of objects and p the number of variables in the data table (simulations with $1 \leq p \leq 50$).

• Outside that range of n values, the results were too liberal, meaning that the test rejected too often the null hypothesis of normality.

• For $p = 2$, the simulations showed the test to be valid for $6 \leq n \leq 11$ and too liberal outside that range of n values.

• If H_0 is *not* rejected in a situation where the test is too liberal, the result is trustworthy.

An R function to carry out this test, dagnelie.test(), is available in package ade4 (Dray et al. 2017) on CRAN. An alternative method is the Henze-Zirker test of multinormality, available in function hzTest() of the R package *MVN* (Korkmaz et al. 2014). Numerical simulations conducted by D. Borcard showed that the Henze-Zirker test has correct rates of type I error for a large range of combinations of number of objects (*n*) and variables (*p*).

Reference

Dagnelie, P. 1975. L'analyse statistique à plusieurs variables. Les Presses agronomiques de Gembloux, Gembloux (Belgium).

Dray, S., A.-B. Dufour and J. Thioulouse, with contributions from T. Jombart, S. Pavoine, J. R. Lobry, S. Ollier, D. Borcard, P. Legendre and A. Siberchicot. 2017. ade4: Analysis of ecological data: Exploratory and Euclidean methods in environmental sciences. R package version 1.7-10. https://cran.r-project.org/package=ade4.

Korkmaz, S., Goksuluk, D. and Zararsiz, G. 2014. MVN: An R package for assessing multivariate normality. The R Journal 6: 151–162.