Size and shape in the measurement of multivariate proximity

Michael Greenacre

Department of Economics and Business

Universitat Pompeu Fabra

Ramon Trias Fargas, 25-27

08005 Barcelona

SPAIN

E-mail: michael.greenacre@upf.edu

Abstract: Most methods of multivariate analysis rely on a measure of proximity between individual cases or samples to quantify inter-sample differences. The choice of this measure is fundamental to the method and its subsequent results. For example, when data are abundance counts of a set of species at several sampling locations, some approaches rely on the Bray-Curtis dissimilarity measure between samples, while other approaches rely on the chi-square distance. A set of observed species abundances at a location has both size, in the form of the overall levels of the species counts, and shape, in the form of the relative values of the counts. The aim of this report is to clarify how much the chosen proximity measure is capturing differences in size between samples as opposed to differences in shape. After motivating the idea using physical morphometric data, the study is extended to nonnegative data in general, with special focus on abundance counts and biomass estimates, which are ubiquitous in ecological research.

Keywords: Bray-Curtis dissimilarity, chi-square distance, cluster analysis, correspondence analysis, multivariate analysis, ordination, visualization.

Introduction

Inherent in most methods of multivariate analysis is the definition of a particular proximity measure computed between multivariate vectors of observations on the samples: for example, species counts from a grab sample in benthic research, or from a quadrat in botanical research. Proximity, or distance, is the multivariate equivalent of a difference for univariate data, and is needed to define equivalent concepts in multivariate space such as the variance and deviations of samples from hypothesized models. The term "proximity" is used here to include measures of distance (also called "metrics"), which obey the mathematical axioms of distance, notably the triangle inequality, as well as measures of dissimilarity, which do not follow the triangle inequality. Well-known examples (amongst many others – see, for example, Gower and Legendre, 1986) of these two types of proximity are the chi-square distance, which is the basis of simple, multiple and canonical correspondence analysis, and the Bray-Curtis (or Sørenson) dissimilarity, which is used widely by ecologists as an easily understandable way of measuring difference between multivariate samples, where the data are species abundances or biomass observed in a fixed area or volume. The choice of proximity measure is also crucial in cluster analysis, because "closeness" of two samples determines their being placed in the same cluster. Hence, for all distance-based multivariate methods it is essential to understand clearly in what respect samples are being measured "close" to one another or "far" apart according to the chosen proximity measure.

In my experience, researchers' choices of a proximity measure, as well as possible transformations of the initial data, are usually governed by their prior education, the school of thought they happen to follow, and the literature they are emulating, rather than an insight into the properties of the measure itself. It is hoped that this report can at least clarify one crucial property of these proximity measures, namely whether, and how much, they capture differences in “size” in the multivariate observations or differences in “shape”.

Some simple examples serve as motivation for this study. Consider two multivariate observation vectors with the same shape but different sizes (the second is twice the first), shown in Figure 1. Thinking of these as abundance counts of five species, or measurements of five variables all on the same scale (e.g. centimeters), can be helpful. Since the shapes are identical, both the Euclidean distance and Bray-Curtis dissimilarity (defined below in Table 1) must be measuring only the difference in size. For example, if one vector is k times the other, then the Bray-Curtis measure is equal to $|k - 1|/(k + 1)$. Hence for $k = 2$, Bray-Curtis = 0.33, as in Figure 1, for $k = 3$, Bray-Curtis = 0.50, for $k = 4$, Bray-Curtis = 0.60, and so on, where the measure of difference increases as the size of one of the vectors increases. If, on the other hand, two vectors of observations have the same size but different shapes with the values shown in Figure 2, then the Euclidean distance and Bray-Curtis dissimilarity turn out to have values identical to those in Figure 1, showing that the same proximity values can be measuring differences in pure size or in pure shape. Once the values are converted to relative values in the form of so-called *profiles*, then any proximity measure is one of pure shape, equal to 0 in Figure 1, but positive in Figure 2.

Kendall’s (1977, 1989) definitions of size and shape apply to a physical object, which is a good starting point as an analogy: shape is all the geometrical information that remains when location, scale and rotation effects are filtered out from the object, while size-and-shape is all the geometrical information that remains when location and rotational effects are filtered out. In the present context where samples are characterized by a set of nonnegative data (e.g., a vector of species abundance counts) there is no question of location and rotational effects, so the vector of data inherently has

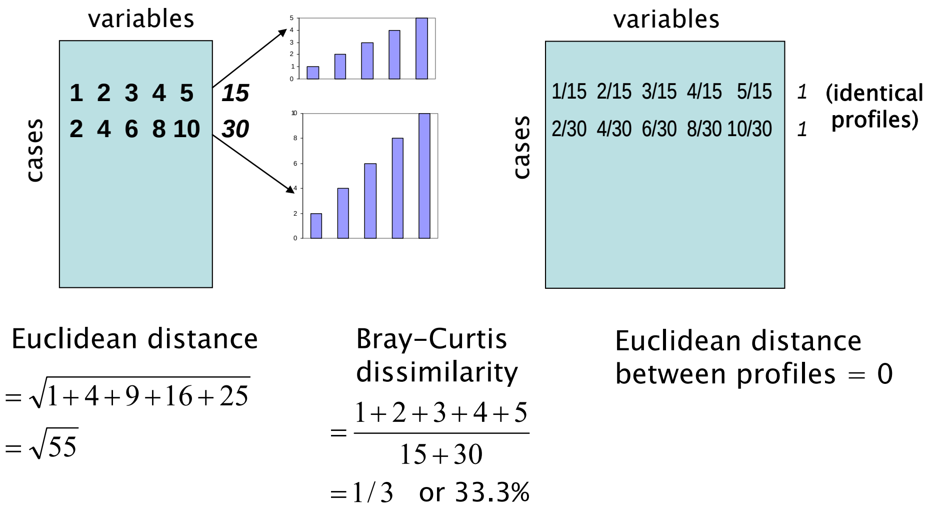

Figure 1: Multivariate observations on two cases (left-hand side) with the same shape but different sizes, summing to 15 and 30 respectively. Both the Euclidean distance and Bray-Curtis dissimilarity give positive values. If the totals are divided out to give profiles (right-hand side) any proximity measure gives a zero difference between the identical profiles.

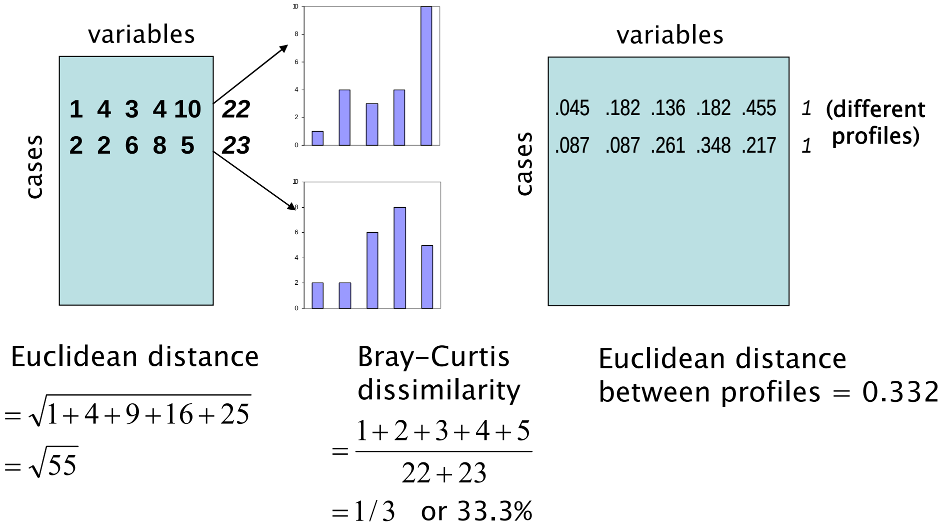

Figure 2: Multivariate observations on two cases (left-hand side) with different shapes but (almost exactly) the same size. The Euclidean distance and Bray-Curtis dissimilarity have values identical to those in Figure 1, while any measure of difference between the profiles (right-hand side) gives a positive value.

size and shape. For our purpose, size will be characterized by the absolute levels of all the values: some cases can have "larger size" when their values are generally higher than those of other cases, and some cases can have "smaller size" when their values are generally lower – what Kendall calls the "scale effect". Filtering out the size to get pure shape can be done in at least two ways: each vector of values can be expressed relative to its grand total, to obtain a profile of values summing to 1 (also called a "composition"). Alternatively, for strictly positive data, all pairwise ratios of values can be considered. Both these ways of "relativizing" the data are commonly found in the literature: the former in all variants of correspondence analysis (see, for example, Greenacre, 2007, 2010b), and the latter in the log-ratio approach to compositional data analysis (see, for example, Pawlowsky-Glahn and Buccianti, 2011).

The idea will be illustrated first by a consideration of some morphometric data, chosen specifically to have uncorrelated physical size and shape components. Then we shall extend the idea to the general case of nonnegative multivariate data, where correlation usually exists between size and shape, illustrated using abundance counts in fish ecology. Several measures of proximity will be studied, including the ones already mentioned above.

Size and shape in morphometric data

Morphometric data serve as a concrete example of the concepts of size and shape in the context of proximity measures. A subset of 200 specimens was extracted from the abalone data in the UCI machine learning repository (Bache and Lichman, 2013) – this reduced data set is provided as supplementary material. For each abalone the database gives length (L), depth (D) and height (H), as well as the abalone's weight (W). W can be used as a surrogate for size, or alternatively the sum L+D+H, or (in this specific case) the product LDH, with which W is highly correlated. The interesting property of this particular data set is that, although there are shape differences, these differences are uncorrelated with size. This can be demonstrated by regressing log(W) on either (i) the log-ratios log(L/D), log(L/H) and log(H/D), or (ii) the log-profile values, log($L/(L+D+H)$), log($D/(L+D+H)$) and log($H/(L+D+H)$). In both cases the variance explained (R^2) is approximately 0.005 (i.e. 0.5%). Alternatively, a redundancy analysis (RDA) of the log-ratios, for example, with log(W) as the constraining variable, gives only 0.1% of the shape variance accounted for by the size variable. This lack of correlation can be visualized in a principal component biplot of the variables that quantify size: W, L, D, H, and LDH, and those that quantify shape: the ratios L/D, L/H, H/D, and the profile values $L/(L+D+H)$, $D/(L+D+H)$ and $H/(L+D+H)$, all of which are log-transformed (Figure 3). The biplot, which shows how all these variables covary, clearly reveals the lack of correlation of W (similarly, LDH) with the shape variables (ratios and profile values). Notice too that the first principal axis, identified with the size variables, accounts for over 90% of the variance, which shows how dominant the size component is compared to the shape component in this data set.

A number of popular proximity measures, shown in Table 1, will be applied to these morphometric data. The first two are examples of unstandardized measures of difference between the values: the *Manhattan* (or *city-block*, or L_1) *distance*, and the *Euclidean* (or L_2) *distance*. Then there are three distances that involve standardizing the data in different ways. The *standardized Euclidean distance* involves dividing each value by its respective standard deviation (hence the division of each squared term by the variance), the *Gower distance* divides each value by the respective width of its range, while the chi-square distance divides the squared term by the expected value, or average. This form of the chi-square distance, applied to the raw data, has been introduced by Greenacre (2010a) to be able to incorporate size differences in comparing ecological samples based

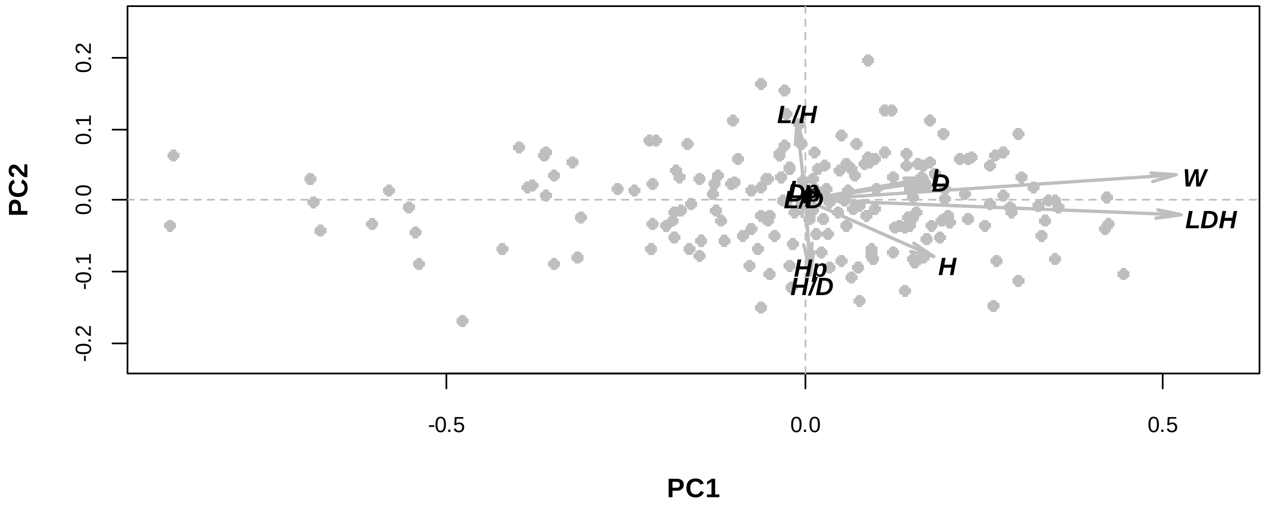

Figure 3: Principal component biplot of weight (*W*), the dimensions length (*L*), depth (*D*) and height (*H*), as well as the product *LDH*, the ratios *L*/*H*, *L*/*D*, and *H*/*D*, and the profile values *L*/(*L*+*D*+*H*), *D*/(*L*+*D*+*H*) and *H*/(*L*+*D*+*H*), labeled as *Lp*, *Dp* and *Hp*, all on a logarithmic scale. The points *L/D*, *Lp* and *Dp* are almost at the origin. The first axis explains 92.2% of the variance, and the second 6.1%. Individual cases are indicated by dots.

on equal areas or equal volumes. The usual chi-square distance formulation, inherent in correspondence analysis, is defined on profile values, and is thus a pure measure of shape differences. The last two measures in Table 1 involve a partial attempt to relativize the values in the observation vectors. The *Canberra distance* divides each absolute difference by its respective sum, whereas the *Bray-Curtis dissimilarity* divides the sum of absolute differences (i.e. the Manhattan distance) by the sum of all the values in the two observation vectors.

The object in comparing the proximity measures is to see how much variance in each matrix of inter-abalone differences is explained by the quantifiers of size and shape. The function `adonis` in

1. Manhattan (city-block) distance	$\lvert L_i - L_j \rvert + \lvert D_i - D_j \rvert + \lvert H_i - H_j \rvert$
2. Euclidean distance	$\sqrt{(L_i - L_j)^2 + (D_i - D_j)^2 + (H_i - H_j)^2}$
3. Standardized Euclidean distance	$\sqrt{(L_i - L_j)^2 / s_L^2 + (D_i - D_j)^2 / s_D^2 + (H_i - H_j)^2 / s_H^2}$
4. Gower distance	$\sqrt{(L_i - L_j)^2 / r_L^2 + (D_i - D_j)^2 / r_D^2 + (H_i - H_j)^2 / r_H^2}$
5. Chi-square distance	$\sqrt{(L_i - L_j)^2 / \overline{L} + (D_i - D_j)^2 / \overline{D} + (H_i - H_j)^2 / \overline{H}}$
6. Canberra distance	$(\lvert L_i - L_j \rvert / (L_i + L_j) / + \lvert D_i - D_j \rvert / (D_i + D_j) + \lvert H_i - H_j \rvert / (H_i + H_j))$
7. Bray-Curtis dissimilarity	$(\lvert L_i - L_j \rvert + \lvert D_i - D_j \rvert + \lvert H_i - H_j \rvert) / (\sum LDH_i + \sum LDH_j)$

Table 1: Proximity measures to measure the differences between two sets (indexed by *i* and *j*) of multivariate observations on three variables *L*, *D* and *H*. Standard deviation, range and mean are denoted by *s*, *r* and an overhead bar, respectively; $\sum LDH_i$ denotes $L_i + D_i + H_i$. Measures 1 and 2 apply as well to the log-ratios log(*L*/*D*), log(*L*/*H*), log(*H*/*D*). All measures can be applied to the relative values $L/\sum LDH$, $D/\sum LDH$, $H/\sum LDH$ of the observations (i.e., profiles). The usual chi-square distance applies to the profiles; when defined on the original values it is called the chi-square of raw data (Greenacre 2010a). The Bray-Curtis dissimilarity applied to the relative values is equal to the Manhattan distance divided by 2. When these measures are applied to more than three variables, the summations simply extend to the full number.

the **vegan** package in R (Oksanen et al, 2013; R Core Team, 2013) is the perfect tool to enable this comparison. This function takes an n×n inter-individual proximity matrix as a response and any set of variables on the n individuals as explanatory variables. As an explanatory variable quantifying size, the logarithm of weight will be used, whereas for explanatory shape variables either the set of profile values or the set of log-ratios will be used. Since size and shape are uncorrelated in this example, any measure of proximity based on the profile values or the ratios should have near zero variance explained by weight. In computing the proximity measures in Table 1, four different forms of the variables L, D and H will be used: the profile values, the log-ratios log(L/H), log(L/D), log(H/D), the original values of L, D and H, and their logarithms in appropriate cases. The so-called *Aitchison distance* (Aitchison et al, 2000) is the Euclidean distance applied to the log-ratios – in general, for p variables, there would be $\frac{1}{2}p(p-1)$ log-ratios.

Another transformation to be considered, which is a kind of standardizing transformation through discretization, is to code the variables into a pre-specified number of fuzzy catgories (Aşan and Greenacre, 2010; Greenacre, 2013). Fuzzy coding (for an introduction, see Greenacre and Primicerio, 2013, chaps 3 and 19) reduces each variable to a set of nonnegative values that sum to 1, quantifying the "possibility" of the variable to be in each category. Following Aşan and Greenacre (2011), triangular membership functions are used to map the original values to their fuzzy values, and the chi-square distance is applied to these values. Since the sums of the fuzzy values for each variable are constant, it makes no difference if one uses the "raw" version of the chi-square distance or the usual "relative" version.

For each variation of the proximity measure, Table 2 reports the variance explained by the logarithm of weight, describing size, and by the two alternative sets of variables describing shape, the profile

Variables used in proximity measure		Percentage variance explained		
		by size (log-weight)	by shape (profile/log-ratios)	residual
profiles (L/∑LDH, D/∑LDH, H/∑LDH)	*fuzzy coding (5 categories)*	*0.3*	*39.5/39.4*	*60.2/60.3*
	fuzzy coding (3 categories)	*0.2*	*66.9/66.6*	*32.9/33.2*
	chi-square (relative)	0.1	99.9/99.6	0.0/0.3
	Euclidean	0.1	99.9/99.7	0.0/0.2
	Manhattan	0.0	100.0/100.0	0.0/0.0
log-ratios (log(L/D), log(L/H), log(H/D)	*fuzzy coding (5 categories)*	*0.1*	*38.7/38.7*	*61.2/61.2*
	fuzzy coding (3 categories)	*0.1*	*66.3/66.0*	*33.6/33.9*
	Euclidean (=Aitchison)	0.1	99.6/99.9	0.3/0.0
	Manhattan	0.0	98.2/98.5	1.8/1.5
original (L, D, H)	Manhattan	90.8	2.5/2.5	6.7/6.7
	Gower	89.4	4.1/4.1	6.5/6.5
	Bray-Curtis	89.3	2.2/2.2	8.5/8.5
	Euclidean	89.1	4.0/4.0	6.9/6.9
	Canberra	87.4	4.0/4.1	8.6/8.5
	chi-square (raw)	87.0	6.2/6.2	6.8/6.8
	standardized Euclidean	84.0	9.4/9.4	6.6/6.6
	fuzzy coding (3 categories)	*55.1*	*6.3/6.3*	*38.6/38.6*
	fuzzy coding (5 categories)	*30.2*	*4.9/4.9*	*64.9/64.9*
log(original) (log(L), log(H), log(D))	Manhattan	90.8	4.1/4.2	5.1/5.0
	Gower	90.1	4.0/4.0	5.9/5.9
	Bray-Curtis	89.9	2.4/2.4	7.7/7.7
	Euclidean	85.8	9.1/9.1	5.1/5.1
	fuzzy coding (3 categories)	*54.1*	*6.6/6.5*	*39.3/39.4*
	fuzzy coding (5 categories)	*30.2*	*4.9/4.9*	*64.9/64.9*

Table 2: Percentages of variance explained by log(weight), as a measure of size, and two alternative quantifiers of shape, the profile values or the log-ratios, for different inter-individual proximity measures. The residual variances for the two alternative choices of shape variables are given in the last column. The measures are applied to different variants of the three variables: the profiles (relative values), log-ratios, original values and log-transformed original values. ∑LDH is short for (*L*+*D*+*H*). Computations are performed using function `adonis` in R package **vegan**.

values (values relative to their total) and the log-ratios. For example, in the first row, when the row profiles are fuzzy coded into 5 categories, the variance explained by size (log-weight) is 0.3%; 39.5% of the variance is explained by shape when the profile values are used as predictors, in which case the residual variance is 60.2%; or, alternatively, 39.4% is explained by shape when the log-ratios are used as predictors, in which case the residual variance is 60.3%. Within each of the four blocks of results the rows are in descending order of the first column, the variance explained by log-weight.

Firstly, notice that it is immaterial whether the profiles or the log-ratios are used as shape explanatory variables – the values in the last two columns of Table 2 are practically the same, varying in the first decimal in only a few cases.

Secondly, all proximity measures computed on the pure shape variables, either the profile values or the log-ratios, are explained close to 0% by log-weight, as expected, since size and shape are uncorrelated in this data set. With the exception of the proximity measures computed on fuzzy-coded variables, shown in italics in Table 2 and which will be dealt with later as a special case, the shape variables account for close to 100% of the variance.

Thirdly, when the original data or their log-transforms are used, the situation is the opposite: size explains between 84.0% and 90.8% of the variance of inter-individual proximities, depending on the chosen measure (again, the fuzzy-coded cases will be discussed separately below), while shape explains between 2.2% and 9.4%. There is no definitive measure of size and shape variance but an RDA of log(L), log(D) and log(H) , constrained by log(W) and either set of shape variables estimates 85.8% of the variance due to size, 9.1% due to shape and a residual of 5.1%. Notice that this RDA is equivalent to the analysis of the Euclidean distances of the log-transformed data reported in the third last line of Table 2. The measures in Table 2 that come closest to these first two percentages are, for the original data, the chi-square distance on raw data and the standardized Euclidean distance. By contrast, the Bray-Curtis dissimilarity has the smallest component of variance attributable to shape, only about 2% – this shows that the partial relativization of the absolute differences between the values by their total sums (see formula in last line of Table 1) under-estimates the shape component and over-estimates the size component. In fact, the size component estimated by Bray-Curtis is similar to the size components measured by the Manhattan and Euclidean distances, which operate on unrelativized data.

Finally in this section, we comment on the fuzzy coded versions of the variables, shown in italics in Table 2. Fuzzy coding is able to manifest nonlinear relationships between the variables, and the more fuzzy categories used the more complex the revealed nonlinearities. Hence, because Table 2 is established by "regressing" the inter-individual proximities linearly on the size and shape variables, the variances explained for the fuzzy-coded variables are quite different from the other proximity measures. For example, when the profiles or log-ratios are "fuzzified" (upper two blocks of Table 2), the variance explained by size is still near zero, as it should be, but the variance explained by shape is much lower than 100%. And when the original data or their log-transforms are fuzzified (lower two blocks of Table 1), the variances explained by the size variable are much lower than the other proximity measures, and decrease further when the number of categories increases. The residual variances are very high for all the fuzzy-coded options, corroborating the fact that they include nonlinear effects not captured by the linear modeling. It would be interesting, in further work, to investigate the form of the nonlinear components of size and shape that the fuzzy coding is intrinsically capturing. A further application of fuzzy coding, showing how nonlinear relationships with explanatory variables can be visualized in a canonical correspondence analysis, is given by Aschan et al (2013), Greenacre (2013) and Greenacre and Primicerio (2013, chap. 19).

Application to abundance counts

As an example of a data set where size and shape are correlated, which is the more common situation in practice, consider the data on abundance counts of 30 fish species in 89 samples from the Barents Sea, used by Greenacre (2013) in the context of canonical correspondence analysis (this data set is available as supplementary material in that article).   In contrast to the previous morphometric example, where shape differences are less pronounced than size differences, the major component of variation in this data set, as well as similar examples in community ecology, will be seen to be shape. Size can be quantified by the logarithm of the total abundance counts in each sample and shape can be quantified by the profile values. In an RDA of the profile values, with size as the constraining variable, 3.24% of the shape variance can be explained by size. This figure, based on Euclidean distance between profiles, increases slightly to 3.94% when chi-square distances are used.

In this example, the size and shape components cannot be separated exactly, as in the previous morphometric application. There is a component of variance due to size once the shape component is partialled out, a component due to shape once the size component is partialled out, and a component that is shared by both size and shape. Borcard, Legendre and Drapeau (1992) explain how to separate environmental and spatial components in a multivariate data set when there is a shared environmental-spatial component – here the same idea is used but applied to size and shape components.

Figure 4 shows the decomposition of variance for some selected proximity measures. The compositional barcharts of the percentages have been ordered from left to right in terms of the percentage of size-related variance that is uncorrelated with shape (the black-shaded part). On the left, the chi-square distances between samples, based on the abundance profiles and hence pure shape, are totally explained by the shape component (the profile values), as expected – there is thus no size component uncorrelated with shape. Since shape accounts for all the variance, the percentage due to size that is common with the shape component is 3.94% , the same figure mentioned previously. At the other extreme on the right is the Bray-Curtis dissimilarity computed on the original (raw) data. This has the highest component of size, 6.60% that is uncorrelated with shape as well as 5.13% that is confounded with shape. This shows once more how much the Bray-Curtis over-estimates the size component. Using the Bray-Curtis on log-transformed data, however,

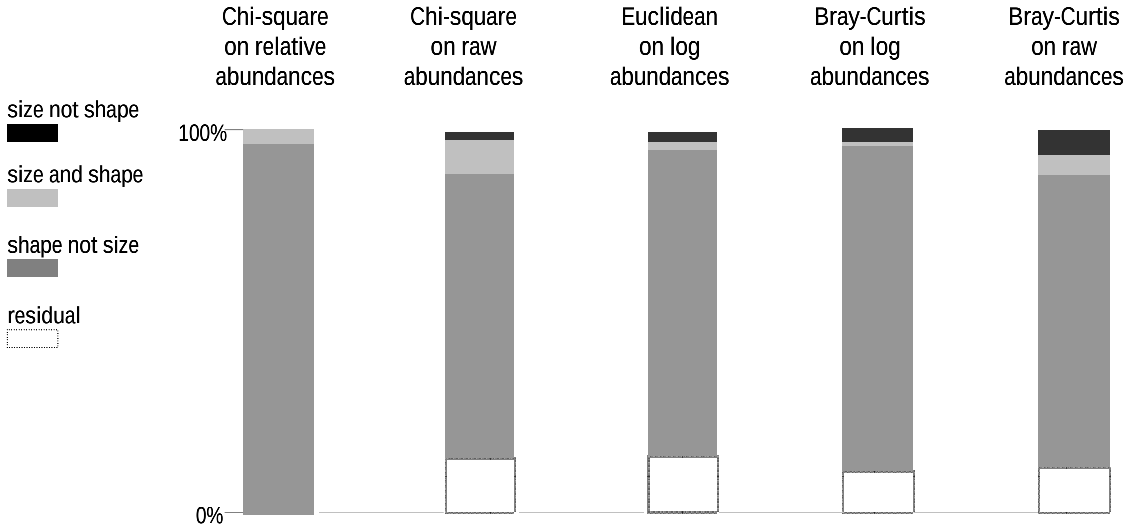

Figure 4: Decompositions of variance, as percentages, for five different proximity measures in terms of, from top to bottom, (1) component of size that is unrelated to shape; (2) shared component of size and shape; (3) component of shape that is unrelated to size; (4) residual variance.

does moderate the size component (3.55% uncorrelated with shape, 0.77% confounded with shape, totaling 4.32%). The Euclidean distances computed on log-transformed data that have not been relativized at all have a similar decomposition of variance, while the chi-square distances on raw data include a large part (8.7%) due to size that is confounded with the shape component. This way of decomposing the variance is enlightening to understand how much size and shape are being taken into account by each proximity measure.

Discussion and conclusions

As stated in the introduction, the choice of a proximity measure is a fundamental choice in any multivariate data analysis, and the properties of the measure need to be fully understood. Many ecologists choose the Bray-Curtis dissimilarity because it has a scale with endpoints that are easy to understand: 0 means exactly the same values in the two multivariate samples, and 1 (or 100%) means no co-presences of any species in the two samples. But, as shown in this report, the Bray-Curtis measure on the original data suffers from an exaggerated size component. This can be mitigated by making a logarithmic transformation of the data but then Bray-Curtis' simple definition becomes much less intuitive when applied to log-transforms of the data plus the obligatory 1 to cope with the zeros: can an ecologist, or anyone else for that matter, then really understand what 50% means on the scale, relative to the original data? The usual justification for either root- or log-transforming the data prior to computing Bray-Curtis is that the highly abundant species, with their larger variances, need to be reduced. Differences between abundant species are overly contributing to the measure, while smaller differences between rarer species are swamped. An alternative technical justification shown in this report is that overall size of the abundance values is excessively captured by the Bray-Curtis dissimilarity on the raw data, and its inherent relativization is clearly not compensating enough for the size differences between samples. For the usual situation in community ecology, where samples are obtained from equal areas or volumes, Greenacre (2010a) proposed the chi-square distance on the raw data, not the relative values as usually applied in canonical correspondence analysis, for example, in order to re-introduce the size component into the distance measure. This alternative chi-square distance, computed by dividing squared differences between the raw data by their expected value, appears to behave well in terms of its size and shape composition. In addition, it is advantageously a weighted Euclidean distance, with all the favorable properties of Euclidean metrics.

Finally, methods such as (canonical) correspondence analysis that use the classic chi-square distance on relative data are clearly analyzing pure shape, and indeed most data sets in community ecology have a major component of shape rather than size. But there will almost always be a part of the shape component that is related to size, as we have seen in the last example. This can be quantified and partialled out, if necessary. The components of size and shape as well as their common component can be quantified for any proximity measure applied to a particular data set, and this gives insight into the measure's properties.

Acknowledgments

Partial support of Spanish Ministry of Economy and Competitiveness grant MTM2012-37195 is hereby acknowledged.

References

Aitchison, J. , Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. 2000. Logratio analysis and compositional distance. Mathematical Geology 32: 271–275.

Aşan, Z. and Greenacre, M. 2011. Biplots of fuzzy coded data. Fuzzy Sets and Systems 183: 57–71.

Aschan, M., Fossheim, M., Greenacre, M. and Primicerio, R. 2013. Change in fish community structure in the Barents Sea. PLoS ONE: 84, e62748.

Bache, K. and Lichman, M. 2013. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. URL http://archive.ics.uci.edu/ml

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

Cramer, W. and Hytteborn, H. 1987. The separation of fluctuation and long-term change in vegetation dynamics of a rising seashore. Plant Ecology 69: 157-167.

Gower, J.C. and Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3: 5–48.

Greenacre, M. 2007. Correspondence analysis in practice. Second edition. Chapman & Hall / CRC Press, London. Published in Spanish translation by the BBVA Foundation, Madrid, 2008, and freely downloadable from www.multivariatestatistics.org

Greenacre, M. 2010a. Correspondence analysis of raw data. Ecology 91: 958–963.

Greenacre, M. 2010b. Biplots in practice. BBVA Foundation, Bilbao, Spain. Freely downloadable from www.multivariatestatistics.org.

Greenacre, M. 2013. Fuzzy coding in constrained ordinations. Ecology 94: 280–286.

Greenacre, M. and Primicerio, R. (2013). Multivariate analysis of ecological data. BBVA Foundation, Bilbao, Spain. Freely downloadable from www.multivariatestatistics.org

Kendall, D.G. 1977. The diffusion of shape. Advances in Applied Probability 9: 428–430.

Kendall, D.G. 1989. A survey of the statistical theory of shape. Statistical Science 4: 87–120.

Legendre, P. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia: 129, 271–280.

Oksanen, J., Guillaume Blanchet, F., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H. and Wagner, H. 2013. vegan: Community Ecology Package. R package version 2.0-9. URL: http://CRAN.R-project.org/package=vegan

Pawlowsky-Glahn, V. and Buccianti, A. (eds) 2011. Compositional data analysis: theory and applications. Wiley, Chichester, UK.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.