

## Review of scalable binary dissimilarities

Tóthmérész, Béla

Ecological Institute, Debrecen University  
Debrecen, P. O. Box 71, H-4010 Hungary  
tothmerb@delfin.klte.hu

**Abstract** Classical and weighted binary dissimilarities are reviewed from the point of view of scaling. New parametric binary dissimilarity index families are also proposed. Detailed review of weight functions is presented along with the introduction of a new formalism, useful presenting a unified discussion of weighted dissimilarities. The usefulness of weight functions and their relationships to the classical binary dissimilarities are discussed. The effect of weight functions is demonstrated by influence curve introduced in the paper. The nonlinear, scalable Rényi weight was one of the most useful during the assessment.

**Keywords** Parametric dissimilarity index families; Weighted binary dissimilarities; Weight functions; Scalable weights; Influence function.

## Introduction

Instead of that the importance of scaling is part of the folklore in ecology, there are just a rather few quantitative techniques available to quantify such research problems. Frequently it may guess that two communities are similar to each other for the frequent species and, at the same time, they are quite different for the rare species. We should have techniques to quantify such research problems. To achieve this goal, scalable similarity and dissimilarity functions have to be developed. In the followings I present a list of the *main types of scalable dissimilarities* (sc1–sc7), which is based on biological, as well as statistical considerations:

- sc1. Weighting of the elements of  $2 \times 2$  contingency tables. The elements of  $2 \times 2$  contingency tables are weighted according to the biological importance of the cells,  $a$ ,  $b$ ,  $c$ , and  $d$ .
- sc2. Weighting of the species by weight functions in the case of binary distances. A weight is attached to each species by a weight function, which provides the weights based on the data structure.

- sc3. Weighted quantitative dissimilarities. There is a weight attached to each species, as it is in the binary case. Instead of this, the importance of the weight is less than in the binary case, because the attributes of the species is not limited to 0 or 1 as it is in the binary case.
- sc4. Scaling based on the power function. A typical example of this kind of dissimilarities is the Minkowski metric. If we take rather strictly, almost all of the scalable dissimilarities related to the power function in some way.
- sc5. Scaling by diversity distances. These dissimilarities are related to the scalable diversities and to the informational divergences.
- sc6. Rarefaction scaling or sample-size scaling. The basic philosophy is quite different than in the previous cases. The scaling is based on rarefaction: the similarity of samples, just like the number of species, depends on the size of the samples.
- sc7. Scaling based on Hausdorff-type distances. These distances are based on set functions. The distance of two groups of samples are calculated.

In the paper the methods mentioned in the paragraphs sc1 and sc2 are reviewed, mainly related to my recent research. These are largely related to binary (presence/absence) data. The techniques of sc3–sc6 are mainly quantitative. Therefore, they are different from technical point of view from the methods mentioned in the paragraphs sc1 and sc2. Distances mentioned in the paragraph sc7 are based on a very different philosophy, which makes possible to use both binary or quantitative data. The methods of the paragraph sc3 is formally related to the techniques discussed in the paper; therefore, they are mentioned briefly to emphasize the continuity of the techniques presented in the list. I would like to emphasize, however, that instead of the evident formal similarity of the methods sc2 and sc3, the differences are much larger than it is evident for the first sight. I also would like to stress that each group of these methods need a separate review paper, because they are vital in quantifying ecological processes which depends on the scale of the study. Majority of these techniques also should be tested from statistical point of view, and should be assessed from the point of view of usefulness of ecological data management.

The next section of the paper is a kind of re-interpretation of the classical binary dissimilarities in a scaling context. Majority of these dissimilarities can be found in the standard books of multivariate analysis (Orlóci 1978,

Legendre and Legendre 2000); although, in the paper the emphasis is on the scalable interpretation. New, generalized, scalable dissimilarities of this kind, are also proposed here. The weighted binary dissimilarities are discussed in the following section. The first of these was published by Podani (1978). Tóthmérész (1994) developed a general interpretation based on the weight functions, and also published the uniform and the hump weights. These weights are reviewed and other, new weight functions are also introduced in the paper. Scalable weight functions are newly developed, published and assessed in this paper. In the final section there are comments on the scalable interpretation, and further perspectives are also discussed including the relationships to the quantitative generalizations of the weighted dissimilarities.

Tab. 1: Notation for a two-species presence/absence cross-classification.

are compared.		plot $r$		
		present	absent	row total
plot $t$	present	$a$	$b$	$a + b$
	absent	$c$	$d$	$c + d$
	column total	$a + c$	$b + d$	$a + b + c + d$

### Weighting of the elements of $2 \times 2$ contingency tables

One of the simplest measure of the similarity of the species composition of two samples,  $r$  and  $t$ , is the number of common species:

$$s_a(r, t) = a, \quad (1)$$

where  $a$  and the later used  $b$ ,  $c$ , and  $d$  are the usual elements of the  $2 \times 2$  contingency table (see Table 1).  $r$  and  $t$  are the identification numbers of the sample plots to be compared;  $r, t = 1, \dots, nP$ , where  $nP$  is the number of plots. Instead of its simplicity, the usefulness of (1) is limited, because 20 common species may be lot, when the species pool is 25. However, it means just a few common species, when there are 500 species altogether. Therefore, usually it is standardized by the number of species in the two compared samples,  $a + b + c$ , or by the total number of species,  $a + b + c + d$ . The first one is the Jaccard-similarity

$$s_{Ja}(r, t) = \frac{a}{a + b + c}, \quad (2)$$

while the other is usually mentioned as the Russel-Rao coefficient of similarity:

$$s_{RR}(r, t) = \frac{a}{a + b + c + d}. \quad (3)$$

In some respect it is also increases the similarity when the same species are missing from the compared plots; therefore,  $d$  should also be represented in the counter of the similarity measure. One of the most frequently used measure of this kind is the simple matching coefficient of similarity:

$$s_{sm}(r, t) = \frac{a + d}{a + b + c + d}. \quad (4)$$

Biologically the pieces of information contained by the cells of the  $2 \times 2$  contingency table are not really equal. The presence of a species may be more relevant information than the lack of an another species. Especially in the case of a study where there are a lot of missing species; e.g. studying the vegetation along a long gradient. Or one may arguing that the alternative occurrence of certain species are really important. All these basic philosophies are represented by the weighting of the cells of the contingency table. The differential weighting may be regarded as *the most elementary types of scaling*.

Just a few of these are mentioned briefly. In the case of Sørensen similarity the joint occurrence of species has double weight. It also may be interpreted as the number of joint species standardized by the average number of species of the compared plots:

$$s_{So}(r, t) = \frac{a}{\frac{(a + b) + (a + c)}{2}} = \frac{2a}{2a + b + c}. \quad (5)$$

In the case of the Rogers-Tanimoto coefficient of similarity the alternative occurrence,  $b + c$ , has double weight:

$$s_{RT}(r, t) = \frac{a + d}{a + 2(b + c) + d}. \quad (6)$$

There are a lot of other kind of weights. They can be regarded, technically, as the members of a *family of dissimilarities, which has a scale parameter*. Gower and Legendre (1986) proposed two measures of this kind ( $\zeta, \eta > 0$ ):

$$s_{\theta a}(r, t) = \frac{a}{a + \zeta(b + c)}, \quad (7)$$

and

$$s_{\theta ad}(r, t) = \frac{a + d}{a + d + \eta(b + c)}. \quad (8)$$

There may be defined even a more general family of that kind ( $v, w, z \geq 0$ ) as:

$$s_{sc}(r, t) = \frac{va + wd}{va + z(b + c) + wd}. \quad (9)$$

Just a few particular cases:  $v = z = 1$  and  $w = 0$  produces (2), while  $v = 1$ ,  $z = 2$  and  $w = 1$  gives (6). Evidently (7) and (8) are also a special cases of (9).

The situation is more sophisticated when a simple, non-constant function is also used for weighting, like

$$s_{scf}(r, t) = \frac{va + f(d)}{va + z(b + c) + f(d)}. \quad (10)$$

It is evident, that (9) is a special case of (10). They are identical, when  $f$  is a constant function. When we choose  $f(d) = \sqrt{a \cdot d}$ , then the Baroni-Urbarni-Busher similarity is received:

$$s_{BB}(r, t) = \frac{a + \sqrt{ad}}{a + b + c + \sqrt{ad}}. \quad (11)$$

Actually,  $\sqrt{a \cdot d}$  is the geometric mean of the number of joint and missing species.

Asymmetric generalizations are also possible ( $v, z_b, z_c \geq 0$ ):

$$s_{as}(r, t) = \frac{va + f(d)}{va + z_b b + z_c c + f(d)}. \quad (12)$$

Asymmetric weighting of  $b$  and  $c$  may be important, because there are multivariate methods, which can display the asymmetric dissimilarities matrices, making interpretable asymmetric relationships. In niche theory, asymmetric overlap measures had always been used to characterize the structure of a community. Kulczynski coefficient may be mentioned as a special case of (12), when  $f(d) = 0$ , and also  $z_c = 0$ . This is an asymmetric dissimilarity, which can be interpreted as the number of common species divided by the number of species of the first compared plot:

$$s_K(r, t) = a/S_r \quad \text{and} \quad s_K(t, r) = a/S_t.$$

### Weight functions for the species

In the case of weighted binary dissimilarities a weight is attached to each species. This weight is based on the occurrence of the species in the sample plots. To cope with the situation we need to introduce a few technical notations. Let  $x_{ri}$  denote any quantitative feature (cover, frequency, phytomass, etc.) of species  $i$  in the plot  $r$  ( $r = 1, \dots, nP$ ,  $i = 1, \dots, ST$ ).  $nP$  is the number of plots, and  $ST$  is the total number of species in the plots. An indicator function

$$I : \mathbb{R} \rightarrow \{0, 1\}, x_{ri} \mapsto I_{ri} \quad (13)$$

is defined in the following way:

$$I_{ri} = \begin{cases} 1, & \text{if species } r \text{ is present in plot } i, \\ 0, & \text{if species } r \text{ is not present in plot } i. \end{cases} \quad (14)$$

This is nothing else just a bookkeeping device; still it is important, because very useful from practical point of view. The number of species,  $S_r$ , of the plot  $r$  can be written as

$$S_r = \sum_{i=1}^{ST} I_{ri}. \quad (15)$$

Using this notation the elements of the classical  $2 \times 2$  contingency tables are defined as follows for the plots  $r$  and  $t$  and for the species pool  $i = 1, \dots, ST$ :

$$\begin{aligned} a &= \#\{(I_{ri} = 1) \text{ and } (I_{ti} = 1)\}, & b &= \#\{(I_{ri} = 1) \text{ and } (I_{ti} = 0)\}, \\ c &= \#\{(I_{ri} = 0) \text{ and } (I_{ti} = 1)\}, & d &= \#\{(I_{ri} = 0) \text{ and } (I_{ti} = 0)\}, \end{aligned}$$

where " $\#$ " means "number of".

Now, a weighted dissimilarity,  $d_w(r, t)$ , is defined as

$$d_w(r, t) = \sum_{i=1}^{ST} w(i) |I_{ri} - I_{ti}|, \quad (16)$$

where  $w(i)$  is the weight of the species  $i$ . It is clear from (16) that a weight function,  $w : i \mapsto w(i)$ , is attached to each species and the difference in presence/absence,  $|I_{ri} - I_{ti}|$ , is weighted by this quantity. There are three types of weights: *absolute weights*, *relative weights*, and *standardized*

*weights*. Formally, a weight function defines a relative weight when the weights of the species sum up to 1:

$$\sum_{i=1}^{ST} w(i) = 1. \quad (17)$$

In the case of a standardized weights, the maximum of the weights is 1. Therefore, each weight function can be used as an absolute weight, a relative weight or a standardized weight. The relative weights  $w_r(i)$  can always be derived from the absolute weights,  $W(i)$ , as

$$w_r(i) = \frac{W(i)}{\sum_{i=1}^{ST} W(i)}.$$

The standardized weights,  $w_{st}(i)$ , can also be derived from the absolute weights as follows

$$w_{st}(i) = \frac{W(i)}{\max\{W(i); i = 1, \dots, ST\}}.$$

The weight function of the binary dissimilarity proposed by Podani (1978) was the following

$$w_L(i) = \frac{\sum_{r=1}^{nP} I_{ri}}{\sum_{i=1}^{ST} \sum_{r=1}^{nP} I_{ri}} = \frac{\sum_{i=1}^{ST} I_{ri}}{\sum_{r=1}^{nP} S_r}. \quad (18)$$

It is a weight function which is increasing linearly with the frequency of the species occurrence in the plots, and it takes its maximum when the species are present in each plot. Therefore, it may be mentioned as a *linear weight*. This weight function is suggesting that the more frequent the species the more important is its contribution to the dissimilarity of plots. Naturally, this weight function can be defined simply as

$$W_L(i) = \sum_{r=1}^{nP} I_{ri}, \quad (19)$$

and its relative form is (18), as pointed out by Tóthmérész (1996). Anyway, the *relative frequency of the occurrence of the species  $i$  in the plots*, which is defined as

$$\pi_i = \frac{\sum_{r=1}^{nP} I_{ri}}{nP} = \frac{\#\{(x_i > 0); r = 1, \dots, nP\}}{nP} \quad (20)$$

should be regarded as the *natural linear weight of a species*. It is slightly different from (18), and it may be mentioned as the *relative occurrence weight*, and denoted by  $W_{\pi}(i)$ . Its importance is also emphasized by the fact that  $\pi_i$  has central role in defining all the weight functions coming below.

In some respect it is natural to use a symmetric, hump-like weight function, which downweights both the extremely rare and the extremely frequent species. Tóthmérész (1996) proposed a nonlinear weight function, which has these properties:

$$W_H(i) = -\pi_i \log \pi_i - (1 - \pi_i) \log(1 - \pi_i). \quad (21)$$

It is based on the Shannon entropy.

It is also possible to create a symmetric variant of the relative occurrence weight as a wedge-like weight function:

$$W_w(i) = 1 - 2 \cdot |\pi_i - 0.5|. \quad (22)$$

Shortly, it may be mentioned as the *wedge weight*.

When each of the species have the same weight, the weight function is a constant function. When this constant weight is the *unit weight*

$$W_u(i) = 1.0, \quad (23)$$

then it may be mentioned as the *unweighted* case. In this case (16) is identical with the Hamming distance. Biologically, it is the number of differential species,  $b + c$ . Using the relative unit weights, (16) is equivalent by the complement of the simple matching coefficient (Tóthmérész 1996).

### Scalable weight functions

There is a simple and straightforward way to create a scalable weight function. It is based on the properties of the power function. In this case the weight function is defined as ( $\alpha \geq 0$ )

$$W_P(i; \alpha) = \pi_i^{\alpha}, \quad (24)$$

where  $\alpha$  may be interpreted as a scale parameter. When  $\alpha = 0$  it is identical with the unit weight. (24) can be regarded as the scalable variant of (20) and it may be mentioned as the *power weight*. Based on this analogy it is

possible to produce a scalable version of (22) using the same idea which led to (24). It may be mentioned as the *scalable wedge weight* ( $\alpha > 0$ )

$$W_{wsc}(i; \alpha) = 1 - |2 \cdot (\pi_i - 0.5)|^\alpha. \quad (25)$$

In this case (24) tends to underweight even the moderately frequent species. It is a remedy to use the following *anti-power weight* ( $\alpha > 0$ ):

$$W_{aP}(i; \alpha) = 1 - (1 - \pi_i)^\alpha. \quad (26)$$

A more sophisticated way to create a scalable weight function is to use Rényi entropy as a scalable generalization of Shannon entropy ( $\alpha \geq 0$ ,  $\alpha \neq 1$ ):

$$W_R(i; \alpha) = \frac{1}{1 - \alpha} \log \{ \pi_i^\alpha + (1 - \pi_i)^\alpha \}. \quad (27)$$

Here,  $\alpha$  serves as a scale parameter. One extreme of (27) is a constant weight, when  $\alpha = 0$ .

It is also natural to choose Daróczy diversity as a scalable weight function. It is defined as ( $\alpha \geq 0$ ,  $\alpha \neq 1$ ):

$$W_D(i; \alpha) = (\pi_i^\alpha + (1 - \pi_i)^\alpha - 1) / (2^{1-\alpha} - 1). \quad (28)$$

Again, (21) is received when  $\alpha \rightarrow 1$ .  $W_R(i; \alpha)$  and  $W_D(i; \alpha)$  are closely related as it was demonstrated by Daróczy (1970).

There is an another approach to define a scalable weight function which is at least as simple as the basic idea of (24). From biological point of view, it may be regarded even more natural, because in a slightly different form it has been used for a long time. It always been a tendency to eliminate the rare species from the samples, as regarded them to be "noise", i.e. to demolish the basic tendencies of the vegetation with their accidental occurrence. Therefore, it may be proposed to eliminate the rare species from the samples before a multivariate or other types of analyses. Orłóci (1973, 1975, 1976, 1978) proposed such techniques attached to the performed multivariate analyses. These techniques also may be used in other context. Tóthmérész (1993) also proposed a special rare species elimination procedure, which is attached to the analysis of gradients. The elimination of extremely rare or rare species always been a practice during data management. From the point of view of binary comparisons, however, the extremely frequent species have the same influence on the discrimination of the samples as the rare species. When a species is present only in one plot, then it has the same influence

when a species is just missing from 1 plot. Biologically these two situations may not be equivalent. Also, this interpretation is not true for the case of quantitative variables. Still, it is true for a binary data set, which is a special situation. Therefore, statistically it is equally justified to downweight the rare and the extremely frequent species. One of the simplest way to do it is to create a *threshold weight* which eliminates symmetrically the  $T$  less frequent and the  $T$  most frequent species ( $1 \leq T < ST/2$ ):

$$W_T(i; T) = \begin{cases} 1, & \text{if } T < [i] \leq ST - T, \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

Naturally, an *asymmetric threshold weight*,  $W_{aT}(i; T_1, T_2)$ , also may be defined if somebody is addicted to the non-binary word, regarding the frequent species as containing more reliable information on the patterns of organisms than the rare ones ( $T_1, T_2 < ST/2$ ):

$$W_{aT}(i; T_1, T_2) = \begin{cases} 0, & \text{if } [i] < T_1 \text{ or } [i] \geq ST - T_2, \\ 1, & \text{otherwise.} \end{cases} \quad (30)$$

A useful way to characterize a weight function is to display the graph of the weight function. This characterization is especially relevant, when the weight function is used for weighting quantitative variables. It may be slightly misleading in the case of binary variables and weighted dissimilarities like (16), because the distance is the sum of the weights multiplied by the  $\delta(r, t) = |I_{ir} - I_{it}|$  differences. Therefore, it is also useful to plot

$$w(i) \cdot P(|I_{ir} - I_{it}| = 1) = w(i) \cdot P(\delta(t, r) = 1) \quad (31)$$

against  $\pi_i$ , the relative frequencies of the occurrence of the species in the plots;  $P(|I_{ir} - I_{it}| = 1)$  is the probability of the alternative occurrence of the species  $i$  in the plots  $r$  and  $t$ . (31) may be interpreted as the *influence curve* of a weight function. In the case of studying a large sample the probability  $P(|I_{ir} - I_{it}| = 1)$  is  $\pi_i \cdot (1 - \pi_i)$ . Therefore,  $w(i) \cdot \pi_i \cdot (1 - \pi_i)$  is plotted against  $\pi_i$ . Moreover, to make easier the graphical comparisons of the weight functions I propose the standardized weights to plot against  $\pi_i$ . In this case the maximum of each weight is 1, and the maximum of (31) is 0.25. This representation is clearly showing the influence of the weights for different  $\pi_i$ 's. There is an another representation displaying the effect of weights described in the paper of Tóthmérés (1996), which is not discussed here. It is based on the PCoA or NMDS ordination of a linear gradient by the

studied weighted dissimilarities. Instead of these graphics displays of the weight functions, the testing of their usefulness in the case of real field data sets is also inevitable.

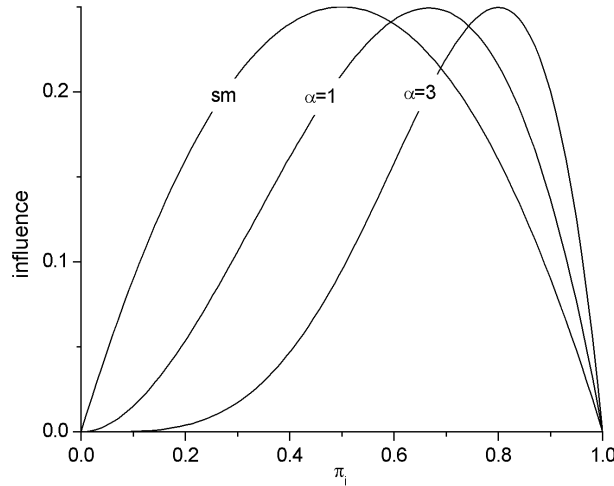


Fig. 1: Influence curves of the standardized power weight functions for the  $\alpha = 1$  and  $\alpha = 2$  scale values; sm means the weight value producing the simple-matching coefficient.

It is evident, that the threshold weights represent a different attitude than the other weight functions. Because of this difference each of the weights can be combined by a (symmetric or asymmetric) threshold weight. Simply it means, that the left and/or the right tail (or both tails) of an influence curve can be truncated by defining a threshold. Below and over that threshold the influence of the species is regarded as zero. Formally, the *generalized threshold weight* is defined as follows ( $T_1, T_2 < ST/2$ ):

$$W_{gT}(i; T_1, T_2) = \begin{cases} 0, & \text{if } [i] < T_1 \text{ or } [i] \geq ST - T_2, \\ w(i), & \text{otherwise,} \end{cases} \quad (32)$$

where  $w(i)$  is an arbitrary weight function.

## Discussion and Future Perspectives

**Abundance and occurrence** What is the relevance of scaling in the binary case? We have decided to ignore the quantitative relations when we accepted to collect just binary data. This situation, however, is totally different than the quantitative case. The scaling is not related directly to the abundance of the species, but to the occurrence of the species in the sample plots). The information related to the abundances is ignored, when presence/absence data is used. However, the scaling types mentioned in the paragraphs sc1 and sc2 are using the information related to the occurrence of the species in the sample plots. It is demonstrated by the following example. There are 10 sample plots and the number of individuals are displayed for the species  $i$  and  $j$  by the vectors below:

$$\begin{aligned} \mathbf{x}_i &= (1, 2, 1, 0, 2, 1, 3, 1, 1, 2)^t, \\ \mathbf{x}_j &= (2517, 0, 0, 957, 0, 0, 4593, 0, 0, 0)^t, \end{aligned}$$

where  $t$  means the transpose of a vector. Species  $j$  is more abundant in the samples than species  $i$  if we quantify it by the total number of individuals; there were 8067 individuals of species  $j$  while just 14 individuals for the species  $i$ . However, species  $j$  was present only in 3 plots, while species  $i$  was present 9 out of 10 plots. From that point of view, the species  $i$  is more frequent. The high total number of individuals were produced by the rather high number of individuals in the plots where the species occurred. It may be resulted in by a very strong aggregation of the species.

Depending on the choice of the weight function the methods mentioned in the paragraph sc3 may be able to use both the information on the abundances and the occurrences in the plots. Distances based on se funtions mentioned on the paragraph sc7 are based on a very different concept that makes possible to use both kind of information types.

**Parametric index families** Both in the case of diversities and in the case of dissimilarities there are two basic kinds of scaling. One of these is like the Rényi diversity, when there is a *one-parametric diversity index family*, and there are classical diversity indices along this scale parameter. The dissimilarities (7)–(12) may be mentioned as the examples of this kind of scaling. In the case of the other kind of scaling, the scale parameter is related in some way to a real biological parameter, like in the case of  $ES(m)$ -diversity, where  $m$  may be interpreted as the size of a sub-sample or as the size of a sub-sample plot. This kind of scaling may be mentioned as *spatial*

*series analysis* (Tóthmérész 1994) or *space series analysis* (Podani 1992). The techniques of sc6, and sc7 may be mentioned as typical examples of this kind of scaling. Generally sc1–sc3 are slightly different than sc4–sc7. The main characteristic of these methods that a weight or a weight function is attached to the species. This is why I write *weighting* instead of scaling, although I regard it as the simplest case of scaling.

***a/d*-invariance and/or asymmetry** What is the basic motivation of introducing a weighted dissimilarity like (12)? In ecology there is a long tradition of using *a/d*-invariant binary dissimilarities and as well as asymmetric binary dissimilarities. The difference is demonstrated by the Table 2. It would be important to make clear difference between these two kinds of phenomena. When the binary variables are nominal variables, like gender, it is evident that the dissimilarity should not depend on the coding of the variable. In the case of ecology the presence or the absence of a species is a kind of quantitative information (presence is always more than the absence), therefore these are ordinal variables. In this case it is highly justified to use *a/d*-invariant dissimilarities, as the botanists and zoologists always did. The asymmetry is totally different kind of problem; evidently those dissimilarities are asymmetric, where  $d(r, t) \neq d(t, r)$ . This kind of dissimilarities are also useful during ecological studies, although these are less frequently used, which may be motivated by the fact that the majority of multivariate techniques prefer symmetric matrices. Niche theory may be mentioned, where asymmetric overlap measures always been regarded as natural.

Tab. 2: Types of binary dissimilarities according to their symmetry and/or coding invariance.

	symmetric	asymmetric
<i>a/d</i> -invariant or invariant to coding	$d(r, t) = \frac{b + c}{ST}$	$d(r, t) = b$ $d(t, r) = c$
not <i>a/d</i> -invariant	$d(r, t) = \frac{b + c}{a + b + c}$	$d(r, t) = b/(a + b)$ $d(t, r) = c/(a + c)$

**Influence functions** The statistics (31) shows clearly the influence of a weight function depending on the occurrence probability of the species. It is

evident that each of the introduced weight function has unimodal influence curve. Therefore, it implies for symmetric weights, that the highest is the influence of a species which occurs in the half of the plots. (In the case of the power weight it also depends on the value of the scale parameter.) This means, that those species are the most influential which occurred at half of the plots. Therefore, implicitly we facilitate to divide the studied plots to two equal halves. When the plots are from three different kinds of habitats, it may not be the best strategy. It implies the question, whether there would be room for influence curves more than one maximum. It is not yet explored, however, what is the relevance to introduce such weight functions. From practical point of view, it looks evident that there is no reason to use an influence function which has more than 2 peaks. 3 peaks would mean that we are preferring to divide the sample plots to 4 parts. In this case, however, the influence functions with one peak are at least as good as the more sophisticated ones. The reason is simple, that the number of plots and the number of species are finite during the field studies and the very subtle differences produced by an overly sophisticated weight function could not be realized. Anyway, a weight function, which has bimodal influence curve can be defined as ( $\alpha \geq 0$ ):

$$W_{bi}(i; \alpha) = \pi_i^\alpha + (1 - \pi_i)^\alpha. \quad (33)$$

It is the unit weight for  $\alpha = 0$ , and  $\alpha \leq 2$  is producing unimodal weight function, and it has two peaks in the case of  $\alpha > 2$ . I prefer to mention it as the *bipower wedge weight*. The graph of the weight function is shown by Fig. 2

**Weighted quantitative dissimilarities** There is a direct way to generalize the weighted binary dissimilarities to quantitative variables in the following way:

$$d_{wQ}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{ST} W(i) \cdot \delta(x_i, y_i)}{\sum_{i=1}^{ST} W(i)} = \sum_{i=1}^{ST} w(i) \cdot \delta(x_i, y_i), \quad (34)$$

where the data vectors (samples)  $\mathbf{x}$  and  $\mathbf{y}$  are compared, and  $\delta(x_i, y_i)$  is an arbitrary difference of the coordinates  $x_i$  and  $y_i$ . This is a limited generalization to quantitative dissimilarities. In some respect (34) is similar to the metric of correspondence analysis, because the rows and the columns are also used during the calculation of two samples. This generalization and its consequences to the data management is not discussed here in details.

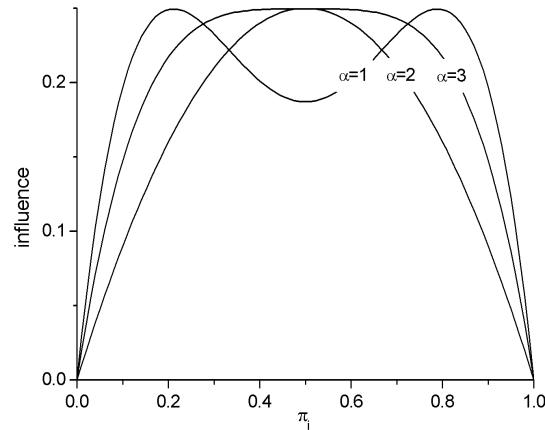


Fig. 2: Graph of the bipower wedge weight for different scale parameter values.

The main reason of it is, that instead of the formal similarity of (16) and (34) there are massive differences between them. In the binary case the choice of the weight function has crucial influence on the value of the  $d_w(r, t)$  distance. This is not always the case for quantitative and mixed variables.

In the case of variables that have the same magnitude, like percentage cover, it is a natural choice to use  $\delta(x_i, y_i) = |x_i - y_i|$  or  $|x_i - y_i|^\alpha$  where  $\alpha > 0$ . In the case of a heterogeneous data set, it is better to use a self-standardized difference which may be scaled by the power function

$$\delta(x_i, y_i) = \left( \frac{|x_i - y_i|}{x_i + y_i} \right)^\alpha .$$

Based on the power function, the following general weighted dissimilarity can be defined ( $\alpha, \beta > 0$ )

$$d_{gWQ}(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^{ST} w(i) \cdot |\delta(x_i, y_i)|^\alpha \right)^\beta , \tag{35}$$

$\delta(x_i, y_i)$  may be defined as  $|x_i - y_i|$  or in a self-standardized form like  $|x_i - y_i|/(x_i + y_i)$ . What is really crucial here, how to choose the  $w(i)$  weight functions? They may be borrowed from the binary case. Therefore, to use the weights discussed in this paper. The combination of the information of

the quantitative variables and the weights based on binary information may be especially fruitful. It is worth for a detailed study. The other possibility to introduce weights of their own right, fitted to the new task.

The dissimilarities used for mixed type of variables formally are also weighted dissimilarities looks like (34). Although, the role of the weight is evidently different than in the binary case. In the case of weighted binary dissimilarities the weights has dominant influence on the  $d_w(r, t)$  distance. In the case of mixed variables these are just to standardize the differences caused by the different types of variables. Each of the generalized similarity coefficients, which are able to use mixed variables, are special cases of (34), because these are weighting and/or standardizing the differences separately for each variable (species). The general coefficient of similarity proposed by Gower (1971) and the generalized Euclidean metric of Podani (1980) may be mentioned as this kind of dissimilarities. These are the representatives of those kind of techniques mentioned in the paragraph sc3. This is a very broad category. The Euclidean distance with oblique coordinates (Orlóci 1978) may also be regarded as this kind of scaling. In this case the differences of the abundances of the species  $\delta(x_i, y_i)$  is weighted in a proper way by the elements of the variance-covariance matrix of the species occurrences. For large data set this is especially computer-intensive, because it needs a lot of calculation. On the other hand it is especially useful because it naturally amalgamate the properties of a difference-type dissimilarity (Euclidean distance) and a ratio-type dissimilarity (covariance).

## References

- Daróczy, Z. 1970. Generalized information functions. *Information and Control* 16: 36–51.
- Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857–871.
- Gower, J. C. and Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* 3: 5–48.
- Legendre, L. and P. Legendre, 1983: Numerical Ecology. Elsevier, Amsterdam.
- Orlóci, L. 1973. Ranking characters by a dispersion criterion. *Nature* 244: 371–373.
- Orlóci, L. 1975. Measurement of redundancy in species collection. *Vegetatio* 31: 65–67.
- Orlóci, L. 1976. Ranking species by information criterion. *J. Ecol.* 64: 417–419.
- Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research*. 2nd ed. The Hague: Junk.

- Podani, J. 1978. A method for clustering of binary (floristical) data in vegetation research. *Acta Botanica Sci. Hung.* 24, 121–137.
- Podani, J. 1980. SYN-TAX: Számítógépes programcsomag ökológiai, cönológiai és taxonómiai osztályozások végrehajtására. *Abstracta Botanica* 6: 1–158.
- Podani, J. 1992. Space series analysis of vegetation: processes reconsidered. *Abstracta Botanica* 16: 25–29.
- Tóthmérész, B. 1993. Noise elimination in gradient analysis. *Abstracta Botanica* 17: 155–158.
- Tóthmérész, B. 1994. Statistical analysis of spatial pattern in plant communities. *Coenoses* 9: 33–41.
- Tóthmérész, B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* 6: 283–290.
- Tóthmérész, B. 1996. Weighted dissimilarity measures for binary data. *Abstracta Botanica* 20: 105–108.