7.8 Species contributions to sample (dis)similarities – SIMPER

Dissimilarity breakdown between groups

The fundamental information on the multivariate structure of an abundance matrix is summarised in the Bray-Curtis similarities between *samples*, and it is by disaggregating these that one most precisely identifies the species responsible for particular aspects of the multivariate picture.[¶] So, first compute the *average dissimilarity* $\overline{\delta}$ between *all* pairs of inter-group samples (e.g. every sample in group 1 paired with every sample in group 2) and then break this average down into separate *contributions from each species* to $\overline{\delta}$.

For Bray-Curtis dissimilarity δ_{jk} between two samples *j* and *k*, the contribution from the *i* th species, $\delta_{jk} (i)$, could simply be defined as the *i*th term in the summation of equation (2.12), namely:

$$\delta_{jk} (i) = 100 \left| y_{ij} - y_{ik} \right| / \sum _ {i=1} ^ p \left(y_{ij} + y_{ik} \right) \tag{7.2} $$

$\delta_{jk}(i)$ is then averaged over all pairs (*j*,*k*), with *j* in the first and *k* in the second group, to give the *average contribution* $\overline{\delta}_i$ from the *i*th species to the overall dissimilarity $\overline{\delta}$ between groups 1 and 2.[†] Typically, there are many pairs of samples (*j, k*) making up the average $\overline{\delta}_i$, and a useful measure of how *consistently* a species contributes to $\overline{\delta} _ i$ across all such pairs is the *standard deviation* $SD(\delta_i)$ of the $\delta _ {jk}(i)$ values.[§] If $\overline{\delta}_i$ is large and $SD(\delta_i)$ small (and thus the ratio $\overline{\delta}_i / SD (\delta_i)$ is large), then the *i*th species not only contributes much to the dissimilarity between groups 1 and 2 but it also does so *consistently* in inter-comparisons of all samples in the two groups; it is a good *discriminating species*.

Table 7.1. Bristol Channel zooplankton {B}. Averages of transformed densities in site groups A and B of Fig. 7.8 (groups from unconstrained divisive tree method), then breakdown of average dissimilarity between groups A and B into contributions from each species (bold). Species ordered in decreasing contribution (until c.90% of average dissimilarity between A and B of 57.9 is attained, see last column). Ratio (also bold) identifies consistent discriminators by dividing average dissimilarity by its SD.

Species name	Av Ab Gp A	Av Ab Gp B	Av Diss	Diss /SD	Cum %
Centropages hamatus	*0.00*	*3.76*	*7.92*	*2.14*	*13.67*
Eurytemora affinis	*3.37*	*0.32*	*6.78*	*2.08*	*25.38*

Species name	Av Ab Gp A	Av Ab Gp B	Av Diss	Diss /SD	Cum %
Temora longicornis	*0.33*	*3.16*	*6.13*	*2.07*	*35.98*
Calanus helgolandicus	*1.09*	*3.64*	*6.03*	*1.62*	*46.40*
Acartia bifilosa	*3.05*	*5.56*	*5.51*	*1.39*	*55.92*
Pseudocalanus elongatus	*2.83*	*4.25*	*4.76*	*2.85*	*64.14*
Sagitta elegans juv	*0.17*	*1.71*	*3.35*	*1.97*	*69.93*
Pleurobrachia pileus juv	*1.23*	*0.58*	*2.71*	*1.04*	*74.61*
Paracalanus parvus	*0.17*	*1.20*	*2.63*	*0.85*	*79.16*
Sagitta elegans	*0.62*	*1.38*	*2.12*	*1.36*	*82.82*
Mesopodopsis slabberi	*0.47*	*0.99*	*1.72*	*1.34*	*85.80*
Pleuobrachia pileus	*0.81*	*0.46*	*1.62*	*1.14*	*88.60*
....	.....	.....	.....	.....	.....

For the Bristol Channel zooplankton data *{B}* of Fig. 7.8, Table 7.1 shows the results of breaking down the dissimilarities between sample groups A and B into species contributions. Species are ordered by the third column, by decreasing values of average dissimilarity contribution $\overline{\delta} _ i$ to total average dissimilarity $\overline{\delta} = \sum \overline{\delta} _ i = 57.9$. They could instead be ordered by the fourth (*Diss/SD*) column*, $\overline{\delta}_i / SD (\delta_i)$. The final column rescales the *Av Diss* values to a percentage of the total dissimilarity that is contributed by the *i*th species $(100 \overline{\delta} _ i/ \overline{\delta})$, and then cumulates this down the rows of the table. It can be seen that many species play some part in determining dissimilarity of groups A and B, and this is typical of such SIMPER analyses, particularly (as in this case) when a severe transformation has been used, since the intention is then to let many more species come into the reckoning. Here, c. 90% of the contribution to $\overline{\delta}$ is accounted for by the first 12 species, with 55% by the first five.

Naturally, the results agree well with the patterns of Fig. 7.8: *C. hamatus* and the *Temora* sp. are first and third in this list because they are scarcely found at all in group A but have good numbers in very many of the group B sites, the *Eurytemora* sp. between them having the opposite pattern. *Calanus* and *Pseudocalanus* spp. are found in group A, consistently so for the latter, but have much higher densities in group B, with a similar pattern (though much less consistency) for *Acartia,* with all 6 contributing 65% of the dissimilarity between those groups. This is also seen in the first two columns of Table 7.1, which are means of the abundances over all sites in each group. Note that this averaging is on *4th-root transformed* scales, so back-transforms of these averages represent major abundance differences (e.g. 1 back-transforms to a density of 1, 3.5 to 150, 5.6 to 1000 etc).

Alternatively, ordering the list by the ratio column (*Diss/SD*) highlights the consistent discriminators of the two groups and the contrast is well illustrated by *Acartia* and *Pseudocalanus* species. While *Acartia* has large numbers, particularly in group B, and higher mean density difference between the groups, ensuring it contributes to the between group dissimilarities, the shade plot shows this density to be variable within the groups and it moves down the consistent discriminator list. *Pseudocalanus* now heads the list even though its densities and mean difference are smaller, because of its greater consistency within groups.

Similarity breakdown within groups

In much the same way, one can examine the contribution each species makes to the average *similarity within* a group, $\overline{S}$. The mean contribution of the *i*th species,$\overline{S}_i$, could be defined by taking the average, over all pairs of samples (*j, k*) *within* a group, of the *i*th term in the Bray-Curtis similarity definition of equation (2.1), in its alternative form, namely:

$$ S _ {jk} (i) = 200 \times \min \left(y _ {ij}, y _ {ik} \right) / \sum _ {i=1} ^ p \left(y _ {ij} + y _ {ik} \right) \tag{7.3} $$

The more abundant a species is within a group, the more it will contribute to the *intra*-group similarities. It *typifies* that group if it is found at consistent abundance throughout, so that the standard deviation of its contribution $SD(S_i)$ is low, and the ratio $\overline{S} _ i /SD(S_i)$ high. Note that this says nothing about whether that species is a good *discriminator* of one group from another; it may be very *typical* of a number of groups.

Table 7.2 shows such a breakdown for group A of the Bristol Channel zooplankton data of Fig. 7.8. The average similarity within the group is $\overline{S} = 62.6$, with 70% of this contributed by the *Eurytemora*, *Acartia* and *Pseudocalanus* species; it is clear from the shade plot that these are the only major ‘players’ in group A. Here *Pseudocalanus*, though the least abundant of the three on average, heads the table, both in terms of contribution to average intra-group similarity and when consistency of that contribution is considered.

Table 7.2. Bristol Channel zooplankton {B}. Average of transformed density in A and breakdown of average similarity into contributions from each species (decreasing order until c.90% of similarity of 62.6 reached); also ratio of contribution to SD.

Species name	**Av Ab Gp A**	**Av Sim**	**Sim /SD**	**Cum %**
Pseudocalanus elongatus	*2.83*	*15.29*	*5.31*	*24.44*
Eurytemora affinis	*3.37*	*14.89*	*1.66*	*48.23*
Acartia bifilosa	*3.05*	*13.72*	*2.03*	*70.15*

Species name	**Av Ab Gp A**	**Av Sim**	**Sim /SD**	**Cum %**
Polychaete larvae	*1.09*	*4.45*	*1.41*	*77.27*
Schistomysis spiritus	*0.87*	*3.00*	*0.84*	*82.07*
Calanus helgolandicus	*1.09*	*2.38*	*0.53*	*85.86*
Pleurobrachia pileus	*0.81*	*2.34*	*0.67*	*89.61*
....	.....	.....	.....	.....

Interpretation

The dangers of taking the precise ordering in these tables too seriously, however, is well illustrated by noting that, if sites 9 and 24 had fallen into group B rather than A, which they did for the agglomerative clustering of this data (with *k-R* clustering giving a third – equally arbitrary – split; see Fig. 3.10), then the contribution and consistency of *Eurytemora* to the intra-group similarities of A would have been notably enhanced. This would have taken it to the head of the list both for contributions to similarity within group A and to dissimilarity between groups A and B.

Some of the confusion that can arise with interpreting SIMPER output stems from the failure to appreciate that SIMPER is not a hypothesis testing technique but an interpretation step that is only permissible once there has been a testing-based justification. So groups to be compared must either be defined *a priori* and then seen to be significantly different under pairwise testing by ANOSIM, or the groups have been determined in *a posteriori* testing by SIMPROF analyses. It is inevitable that two groups which are *not* significantly different will have *some* breakdown of their between-group dissimilarities (which will never be zero) into contributions from each species, but if the mean dissimilarity between two groups is no different (statistically) from that within the groups then it is not meaningful or sensible to look at that breakdown.

Another occasional source of confusion is that sometimes a species will have similar mean abundance in two groups but will still feature somewhere in the list of species contributing to the dissimilarities between them. One simple explanation[‡] is that if the densities (or biomass, area cover etc) are not negligible then samples from one group will inevitably have *some* dissimilarity to samples in the other group (except in the unlikely event that values are effectively identical in all replicates of both groups, in which case that species cannot feature in the list). The outcome will be that the standard deviation of those dissimilarities is relatively large, so that the *Diss/SD* ratio column is too small for that species to be taken seriously – on its own it would certainly not suggest that the groups differ (the implication of a low ratio). In other words, you need to keep an eye on both columns in bold in Table 7.1 (and 7.2) for any interpretation, whether you are primarily using the *Av Diss* column to better understand which species *have* contributed to the difference between those groups or *Diss/SD* to pick out a small number of key species you might monitor to characterise future changes, for example. This is the motivation for SIMPER’s reporting of these two criteria – they serve different practical requirements.

Extensions of SIMPER (Euclidean and 2-way)

The Bray-Curtis measure lends itself to this breakdown into species contributions, both in terms of the dissimilarities between groups and similarities within groups, because of its two equivalent definitions that are expressible as sums over species – of equations (7.2) and (7.3) respectively. Other coefficients can be used; for example, it is straightforward to break down (squared) Euclidean distances into contributions from each of a set of (usually normalised) environmental-type variables, since from equation (2.13):

$$ d_{jk}^2 (i) = (y_{ij} - y_{ik})^2 \tag{7.4} $$

needs simply to be summed over species *i = 1, ..., p*. This deals with identifying variables which primarily differentiate two groups of environmental samples (or other data for which Euclidean distance is relevant), but the reverse table of ‘nearness’ breakdowns within groups is less intuitively constructed.

[¶] *This is implemented in the SIMPER routine in PRIMER, both in respect of contribution to average similarity within a group and average dissimilarity between groups.*

[†] *Though this is a natural definition, it should be noted that, in the general unstandardised case, there is no unambiguous partition of δ_{jk} into contributions from each species, since the standardising term in the denominator of (7.2) is a function of all species values.*

[§] *The usual definition of standard deviation from elementary statistics is a convenient measure of variability here, but note that the δ_{jk}(i) values are not independent observations, and standard statistical inference cannot be used to define, for example, 95% confidence intervals for the mean contribution from the* i*th species.*

[‡] *A more subtle possibility is that SIMPER (in line with ANOSIM, which has the same property) is identifying a difference which is more a function of very strong dispersion differences between the groups rather than mean differences, where that arises from a consistent pattern of variance differences in the key species (but note that, quite often, community dispersion differences between groups arises from a totally different source – that of higher turnover or greater sparsity of species in one group than another).*

PRIMER does this by again tabulating a breakdown of squared (usually normalised) Euclidean distances, but for values within a group the table is therefore headed by variables which have zero or low contributions, taking the same or similar values within the group and thus accounting for little of its total squared Euclidean distance. For comparison between groups, the tables have a more familiar ‘feel’ in terms of the analogy with Bray-Curtis SIMPER output. That only squared Euclidean distance is partitioned, not Euclidean distance itself, is not generally of great concern in the context of PRIMER analyses, since they (ANOSIM, nMDS, BEST, RELATE etc) are usually only a function of ranks of the resemblances – identical whether Euclidean distance is squared or not.

Revision #23
Created 27 February 2022 13:21:11 by Arden
Updated 30 October 2024 03:54:58 by Abby Miller