Section: Numerical classification

TWINSPAN (hierarchical divisive classification)

Theory R functions Examples

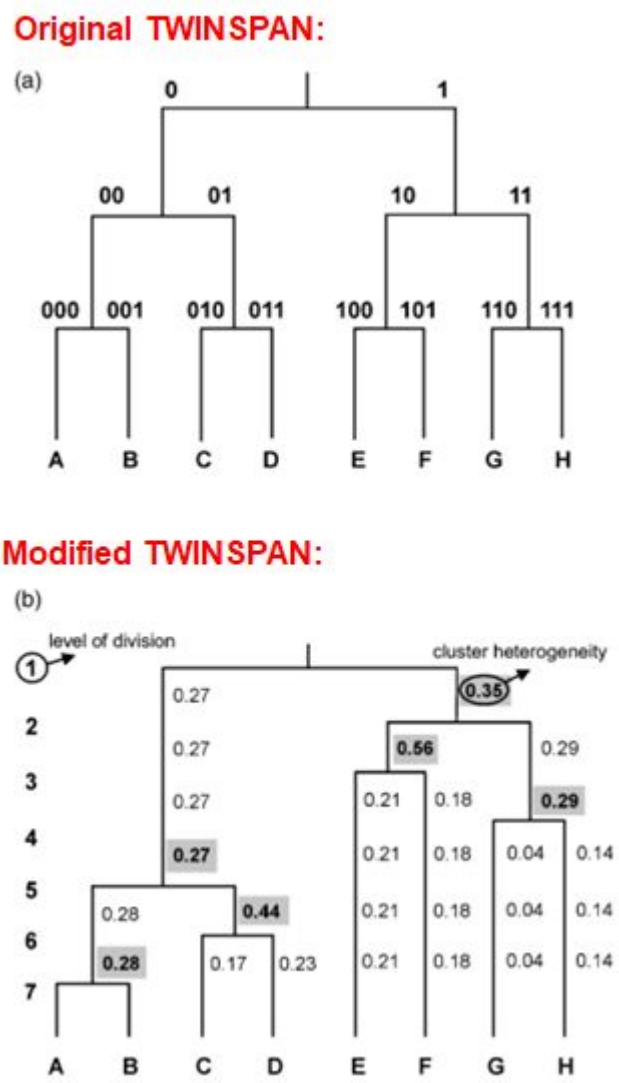

Figure 1: Dendrogram of the original (a) and modified (b) TWINSPAN algorithm. While in the original TWINSPAN, at each level of the division each cluster is divided into two clusters (unless the cluster contains too few samples), in the modified TWINSPAN only the most compositionally heterogeneous cluster is divided into two clusters.

TWINSPAN (abbreviation standing for Two-way indicator species analysis) is hierarchical and divisive method of numerical classification, which uses the results of ordination (namely CA) to divide the whole dataset into subdivisions. The method has been introduced by Mark O. Hill in 1979. It is not the only divisive algorithm in hand (others like DIANA or COINSPAN exist), but it is with no doubt far the most commonly used one.

The algorithm itself is rather complex, and consist of the following steps:

1. ordination of samples along the first axis of correspondence analysis (CA1) and splitting the axis near the middle;
2. identify the indicator species which have high fidelity to each side (negative and positive) of the axis, and use them to further refine the classification of samples which are near the middle to avoid their misclassification;
3. take samples in each subdivision and apply steps 1 and 2 on them.

Two stopping rules are applied to stop the division: minimum size of the subdivision (for example 5 – groups with five and fewer samples are not further divided) and the number of levels to which subdivision advances (for example 3 – only three levels of division are used). In each level of division, all groups of samples are divided (unless they are too small), which means that the number of resulting clusters is 2, 4, 8, 16, 32, ... 2^2 for one, two, three, four, five ... *n* levels of divisions. A simple modification of the original algorithm allows the user to choose the desired number of clusters: in step 3, instead of dividing all subdivisions, divide only the one that is the most compositionally heterogeneous (has the highest cluster heterogeneity, measured by one of chosen beta diversity metrics). This **modified TWINSPAN** (Roleček et al. 2009) allows choosing any number of clusters (Fig. 1).

Because the concept of indicator species is working with species presences and absences, the whole TWINSPAN algorithm is using only presence-absence data. To include also quantitative information about species abundances, species with higher abundances are multiplicated in the matrix by being converted into pseudospecies (i.e. the more abundant is the species in the original data, the more pseudospecies represent it in the pre-processed data); the conversion is done using pre-defined cut-levels. Result of TWINSPAN is a hierarchical dendrogram showing the relationship between individual subdivisions, a list of (one or several) indicator species for each split, and also the optional two-way ordered table, where both sites and species are ordered; sites are ordered according to the splits of the hierarchical classification, while species are ordered to form the blocks within the groups (Fig. 2).

TWINSPAN is often criticized for being a not-elegant sequence of arbitrary and not fully documented steps. On the other side, ecologists often love it, since it does return results which are ecologically quite intuitive. Vegetation ecologists especially have a long tradition in using TWINSPAN, mainly because the author (M. O. Hill) is himself a vegetation ecologist and he designed the method to closely resemble the traditional Braun-Blanquet approach of classifying vegetation (e.g. because it produces the two-way ordered table and emphasized the use of indicator species with high fidelity to individual groups, Fig. 2).

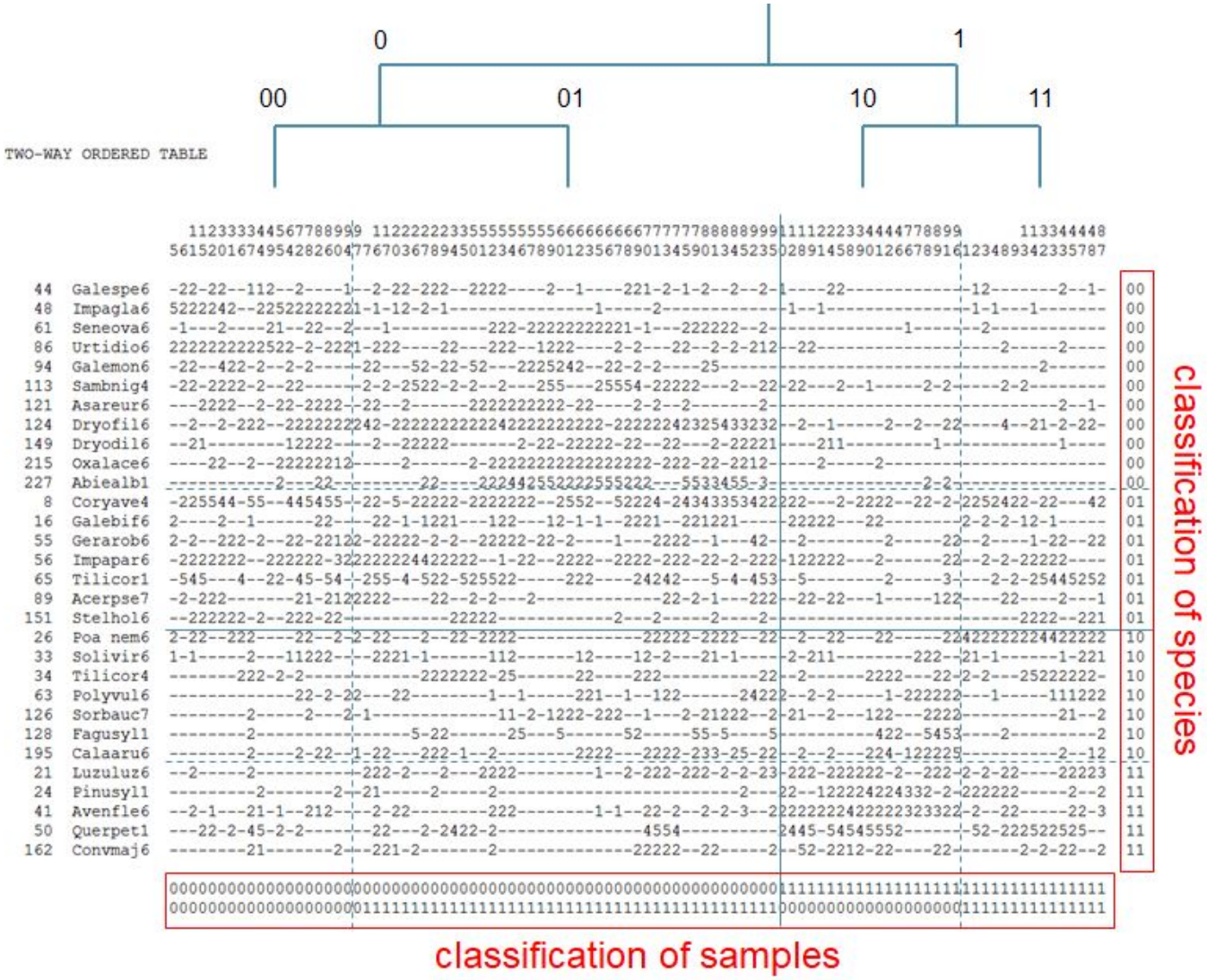

Figure 2: Two-way ordered table resulting from TWINSPAN algorithm (this example is based on Vltava dataset, which has been reduced to contain only species occurring in at least 30 samples (out of 97) to contain fewer species; the calculation was done in twinspanR library.

The true algorithm is actually much more complex, and even the original description by Hill (1979) does not contain all details (some changes have been introduced later by other authors directly in the FORTRAN code of the TWINSPAN program). Perhaps the most detailed description of the algorithm with attention to some of the details is given in Kent (2012). Some software offers TWINSPAN (note that the implementation in each of them actually slightly differs, since some are using a different version of the FORTRAN code): TWINSPAN for Windows, PC-ORD, CAP and JUICE. In R, I created a simple experimental package `twinspanR`, which is an R-wrapper around the twinspan.exe program and works only on Windows platform (this implementation includes both original and modified TWINSPAN algorithm).