Bootstrap averages in a reduced mMDS space

Though hopefully the above gives the motivation and an idea of the way the region estimates are constructed, the most important instruction of this Section is to read Chapter 18 of CiMC! The detailed reasoning and information it gives will not be repeated here but the upshot is that the best way of constructing the bootstrap averages which are then displayed (and smoothed, bias-corrected etc.) in 2-d *m*MDS, is to calculate them from m-dimensional metric MDS ordination co-ordinates created by running *m*MDS on the original dissimilarity matrix. This is carried out for a range of increasing values of m, starting from m=4 (up to m=10, though this limit is usually not reached) and stopping when the ordination configuration crosses a threshold for how well it matches the original dissimilarity matrix. In other words, m is chosen to be just high enough to give a 'near-perfect' representation of the dissimilarities. The criterion, as used in several guises in previous sections (the cophenetic correlation of cluster analysis, the matching coefficient for RELATE and BEST/BVStep, e.g. of a subset of species to the multivariate sample pattern for the full species set etc.) is just a matrix correlation ρ. Here this is between the original dissimilarities and the distances (which are Euclidean of course) among the sample points on the *m*MDS – in other words, ρ is the Pearson correlation of the points in the Shepard diagram. (In the context of *m*MDS and the need to retain the metric information in the original dissimilarities, as discussed above, it makes sense to use a standard Pearson correlation here and not the usual rank-based Spearman correlation). The default in **Analyse>Bootstrap Averages** for (•Auto m) choice is that the smallest m is chosen to make Pearson $\rho \ge$0.99, though this is under user control. The threshold criterion we adopted for successful reconstructions of the original dissimilarities, in the BVStep runs at the end of Section 14, was (Spearman) $\rho \ge$0.95, so the more severe $\rho \ge$0.99 could certainly be relaxed a little if necessary, without compromising the approach. As shown in Chapter 18, CiMC, the dimension m in which the bootstrapping operates must avoid being too large, otherwise an artefact of high-d bootstrapping becomes increasingly important, resulting in significant underestimation of true dispersion by the bootstrap averages, however many original replicates there are in a group (i.e. however well-behaved a univariate bootstrap might be). This explains the restriction to 4$\le$m$\le$10 in the (•Auto m) option, but the routine also permits manual choice of m, to allow the user to look at the outcome from a wider range of dimensionalities.

Starting from an active sheet which is the full sample resemblance matrix, **Analyse>Bootstrap Averages** therefore replaces this by m-dimensional *m*MDS co-ordinates (another approximation therefore in the series leading to our smoothed, nominal 95% region estimates! – but a very useful one, giving the technique some excellent properties). It is in this reduced space that *n* bootstrap samples are chosen, for a group with *n* replicates (*n* will differ for each group, in general) and their means calculated – so the *Bootstrap Averages* of this section are all simple averages for each of the m co-ordinates of an *m*MDS ordination. This is repeated *b* times – also under user choice, though the routine suggests a default which limits the overall number of bootstrap averages across all groups to 300. (However, most machines can run MDS for at least twice that number, hence the earlier encouragement to increase *b* to at least 100, if at all feasible).

Revision #4
Created 11 November 2024 02:18:44 by Arden
Updated 26 February 2025 04:28:16 by Abby Miller