Missing data estimation

The subject of missing data has arisen several times already (Sections 1, 3, 5) and the point made that the terminology and sheet entry Missing! refers only to variables (usually environmental -type variables) that are not recorded for some samples. It does not refer to designs which were intended to be balanced but for which some replicate samples were not analysed for some reason, over all variables. (Unbalanced replication is not generally a problem to handle in PRIMER, since balance is not required for most of the testing that PRIMER, and PERMANOVA+, are able to carry out.)

Some of the routines, including PCA (next section), require the user to enter a complete matrix, with no missing values. At a simple level, it is fairly clear why this should be so. For the trivial 2-variable case in which PCA was introduced in Chapter 4 of CiMC, imagine losing one variable value for one of the samples. What is now that sample's contribution to total variance? How can it be projected perpendicularly to the best-fitting line through the points? How can that first PC axis be determined at all without knowing the contribution of this sample, and so on? In fact, a solution to this was suggested in Section 5 when discussing computation of resemblance measures in the presence of missing entries – it is possible to adjust Euclidean distances, or any other distance/dissimilarity measure, for the crude bias that may come (and certainly will come for Euclidean distance) from some pairs of samples having more matching variables across the two samples than others do. The resulting (near-)Euclidean resemblance matrix is then complete and a choice can be made between MDS (possibly metric) or PCO in the PERMANOVA+ add-on software. The latter is a PCA when the matrix is Euclidean (though the missing data will make that identity not quite true). An alternative is to remove (listwise) as few variables and samples as possible, in a judicious balance, such that a complete matrix is left. The routines **Tools>Check**, **Select>Samples**>(•No missing values) or **Select>Variables**>(•No missing values) will help with this. When there are large blocks of missing data – a subset of the variables were simply not recorded at a large group of sites – then this is likely to be the most realistic option. In other situations, where there is very little missing data, it can seem very wasteful of valuable resources – a whole sample would have to be deleted because one variable is missing, or a whole variable deleted because it was not measured for one sample. In this case, there are then two realistic options – work always from a resemblance matrix and allow PRIMER to adjust automatically the pairwise distances for the crude bias, or use a completed data matrix obtained by estimating the missing values with the EM algorithm. If some restrictive distributional assumptions apply (with rather few missing values and good correlations between some of the variables), this can provide a less crude adjustment and should be attempted.

Revision #6
Created 24 September 2024 01:34:46 by Arden
Updated 11 February 2025 21:58:06 by Abby Miller