EM algorithm assumptions

Tools>Missing is designed to operate only on matrices for which: a) assumptions of multivariate normality can be made; b) there are many fewer variables than samples, so that there are enough data values to be able to estimate the parameters representing means, variances and correlations of all the variables, with reasonable stability; c) there are rather few missing data points (each of those is a new parameter that needs estimating also); d) the data points are thought of as 'missing at random', rather than missing because they were so extreme that they could not be recorded; e) the samples are treated as of unstructured design, rather than, for example, utilising information about their status as replicates from a set of *a priori* defined groups.

Many of these are the assumptions that the methods of PRIMER are trying to get away from, of course! But that is mainly because they are completely impossible to satisfy for assemblage data; they may be much more realistic for continuous, environmental-type data (including, for example, morphometric variables). The estimation technique that PRIMER uses is the standard statistical method under these conditions, namely the EM (expectation-maximisation) algorithm. It is rather tricky (and dangerous!) to give guidelines for when the method will prove acceptable, but you do have some help from the algorithm. Firstly, if you set it an impossible problem (far too many parameters to estimate for the number of data points you have) then it will fail a convergence threshold and display an error message (*max number of iterations exceeded*). Secondly, when it does converge, it is also able to provide an approximate standard deviation for its estimate of each missing value. If this is large then there has clearly been insufficient information to pin down a likely value for the missing cell. As a rough rule-of-thumb, you should not expect to be estimating more than about 5% of your data points if your analysis is to retain any credibility(!), and you should have enough samples *n* compared with (selected) variables *p* and missing cells *m*, so that there is a half-decent number of data points per estimated parameter $DpP = n/[(p+3)/2 + (m/p)]$ (around 7 is sometimes cited, in general contexts). When this criterion is far from being met using the whole matrix, you may be able to take a piecemeal approach, selecting just a small set of the most relevant variables to drastically reduce *p*. The method is clearly only going to provide you with something useful if there are variables that correlate fairly well with the one containing the missing data, so that it has some basis for the prediction. **Draftsman Plot** will work on datasheets with missing cells, so you can use this (and its correlation table) to select out good subsets of variables for estimating each missing data cell. Use of **Tools>Missing** should not be seen as an automatic process therefore – you must expect to have to work hard to justify any data points that you are making up! In the end, common sense is the best guide here, as always. Look at each estimated value – they are always displayed in the worksheet in red – and compare it with the range of values from the other samples for that variable. Does it look 'reasonable', or has something clearly gone wrong with the fitting routine? If all appears well, then it does have the objective credibility of being the maximum likelihood estimate of that cell, and not just some subjective value that you wish it was! Also, look at the standard deviation (σ) of the estimate in the results window and try sensitivity analysis. Add or subtract up to 2σ from each of the estimated cell values at random, and re-run your PCA (or MDS, ANOSIM etc.). Whatever you estimate for the missing values may make no difference to the outcome, if they are within a reasonable range of the other data – you then have a very credible analysis.

Revision #5
Created 24 September 2024 01:39:45 by Arden
Updated 11 February 2025 22:06:54 by Abby Miller