Subject: Re: clustering
Posted by James Kuyper on Mon, 23 Jul 2007 18:16:25 GMT
View Forum Message <> Reply to Message

little davey wrote:

> Is it the case that you MUST use standardize() before you call
> CLUST_WTS()?

No, you don't need to standardize. The scaling of the data affects the results you're going to get. The algorithm is built to treat a difference of 1.0 in a variable between two data points as being equally significant, no matter which of the variables that difference occurs in. This is a fine assumption for variables that have equivalent meanings; such as the x, y, and z coordinates when you're clustering stars.

However, in most contexts for most variables that's simply not true. You can use the scaling to tell CLUST_WTS() treat differences in one variable as more important than differences in another variable. That's fine if you have a clear idea as to which variables are more important than others, and by what factor.

However, the most common case is where the analyst doesn't have clear advance knowledge of the relative importance of the different variables; the analysis is being done to get some idea as to which are the important variables. Scaling the variables according their standard deviations, as STANDARDIZE() does,provides a comforting illusion of objectivity to the choice of scale factors. Unfortunately, if there are only very small variations in an unimportant variable, standardizing it would give that variable undue importance in the clustering. Nothing can substitute for good judgment on the part of the analyst. However, STANDARDIZE() does produce fairly good results in many cases, which implies that the objectivity it provides not quite as illusionary as I've suggested.