Subject: Re: Cluster analysis
Posted by George N. White III on Sun, 05 Dec 2004 14:43:58 GMT
View Forum Message <> Reply to Message

On Fri, 3 Dec 2004, Chris Jacobsen wrote:

> In our work we started out using the "stock" IDL
> cluster routine but we have added to it a bit.
> Still, we have not changed the basic algorithm.
> We've found that preparation of the data can
> make a big difference. If the variation in
> variable X is 100 times bigger than the variation
> in variable Y, then the clustering (which looks
> at simple Euclidian distance) will not see the
> variation in Y very well. One approach is
> to subtract the mean of each variable, and
> apply a scale factor to the data in variable
> Y so that it is spread out over the same distance
> as in variable X.

Sound advice. You should also consider whether a non-linear transform (e.g. alog()) should be applied to some variables. Many people overlook the rank transform, which gives you a distance measure that is essentially the number of observations with values between the two points. This is a way to make sense of comparisons between different physical quantities.

--
George N. White III <aa056@chebucto.ns.ca>