Multivariate Analysis Methods: Background and Example

Fionn Murtagh[1]
Space Telescope — European Coordinating Facility
European Southern Observatory
Garching bei München, F.R. Germany

Abstract

Multivariate statistical methods deal with the inherently very difficult problem of detecting patterns in data. These patterns can take many forms — natural groups, inherent dimensionality, correlations, dependencies, and so on. Often, therefore, different methods bring different features of the data to light.

Following a brief overview of some prominent multivariate methods, we illustrate their use on IRAS data. We indicate how different multivariate methods can be "chained together" to yield powerful tools for uncovering structure in data.

1 Multivariate data analysis methods

When faced with large quantities of multiple-parameter data, multivariate data analysis algorithms can offer considerable time-savings, together with ensuring consistency and "objectivity" of treatment. Being multivariate (multidimensional), they allow the simultaneous treatment of many variables.

There are many types of multivariate algorithms, but among the most commonly used are algorithms for *cluster analysis*, *discriminant analysis* and *principal component analysis*.

Given a set of objects, each characterised by the same set of variables, clustering methods will produce groups of the objects. The objects in the resulting groups will either be more similar in feature space to one another than to non-group members, or satisfy some other homogeneity or compactness criterion. "Similarity" is most often defined by the Euclidean distance, but other metrics may well merit consideration. The question of "standardization" or "normalization" (centring the objects in the multidimensional space and rescaling them to have unit variance) may also have to be addressed before carrying out the clustering. Of course the groups of objects found by a clustering algorithm are in the parameter/variable space, and this will not necessarily have a direct relationship with positional, 2- or 3-dimensional space.

A method of clustering (closely related to widely used hierarchical clustering methods and percolation methods; for the latter, see, *e.g.*, Schulman and Seiden 1986), is the

[1] Affiliated to Astrophysics Div., Space Science Dept., European Space Agency.

minimal spanning tree. It is a graph theoretic representation of the set of points and has been used in studies of galaxy clustering (*e.g.* Barrow *et al.* 1985).

One of the aims of discriminant analysis methods is to assess a known assignment of objects to groups. Thus, such methods can be used to study the results of a cluster analysis. Discriminant methods also can be used for assignment of objects to already existing groups. When used for this second objective (*i.e.* assignment), discriminant analysis has been referred to as "supervised classification" (because of the need to define the training set, — perhaps by a visual study of a relatively small number of objects), while cluster analysis has been termed "unsupervised classification".

Principal components analysis is used for dimensionality reduction The best linear combinations of the axes in the initial parameter space are sought. Thereby new and often fewer coordinate axes (the underlying "principal components" of the data) are determined. These may be used for interpreting the data or for providing the best possible planar projection(s) of the data.

Comprehensive background material on multivariate methods is available in Murtagh and Heck (1987b). Other general references on this area include the MIDAS Users' Guide (1985) and Murtagh and Heck (1987a). MIDAS software includes all methods discussed here. In particular a storage-economic hierarchic clustering method is available: in-core storage of an $n \times m$ input matrix (where n and m are number of rows and columns, respectively) is required rather than the usual $O(n^2)$ storage. Also a

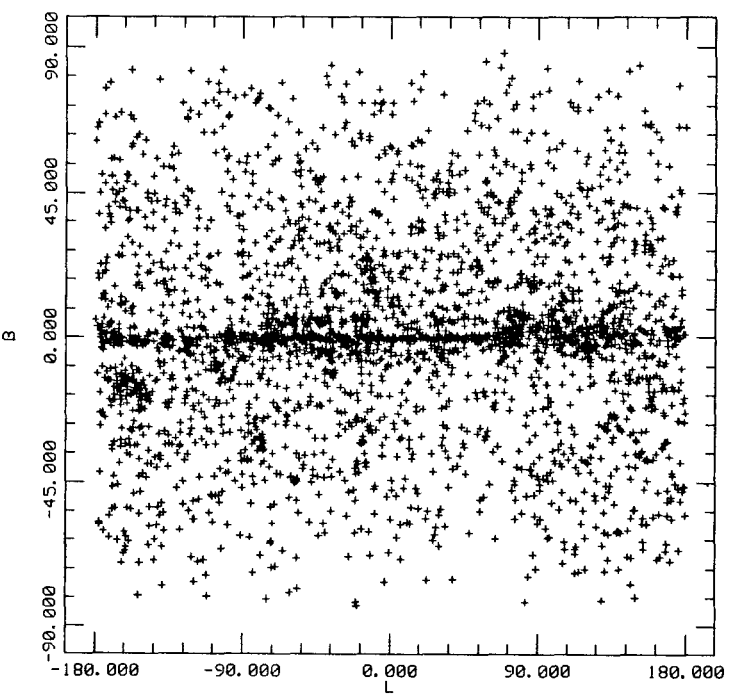

Fig. 1. Galactic longitude and latitude locations of all 3178 objects.

very efficient routine for the minimal spanning tree is available, based on an initial preprocessing of parameter space (Rohlf 1978).

2 Application to IRAS Point Source Catalog data

The data used consisted of a sample of 3181 IRAS PSC objects. The article by Adorf and Meurs (1988) should, in particular, be referred to for more background on the data used (further references may be obtained in Meurs *et al.* 1988). The 3181 objects were taken from approximately 30 000 non-stellar objects for computational/storage convenience; the selection was random (every tenth object from the collection of 30 000-odd objects was taken). The aim was to investigate the main classes of these objects with a view towards finding relatively well defined groups of objects for further study.

Three colours and a log *flux* value (see Meurs *et al.* 1988) were used to characterize these objects. In the notation of the last-mentioned reference, these are: c_{12}, c_{23}, c_{34} and $\log f_{100}$. Three objects having missing values were deleted, leaving a set of 3178 objects characterisable as points in a four-dimensional parameter space. Again, in the last-mentioned reference, it is described how clusters corresponding to "thin (galactic) plane", "cirrus" and "galaxies" are of interest: these can be represented in the plot of galactic longitude and latitude positions associated with the 3178 objects (*Fig. 1*).

A principal components analysis (PCA) was carried on the 3178×4 matrix. The PCA was carried out on a correlation matrix, *i.e.* the 3178 objects were *centred* and *reduced* in the parameter space. A plot of the objects in the principal plane (*i.e.* the plane defined by principal components 1 and 2) did not seem particularly interesting. For instance, no grouping of the objects was visible in this optimal planar representation.

However, the three principal components accounted for more than 96.5 % of the variance. These three new parameters were used as input for a cluster analysis (*i.e.* the input data matrix was of dimensions 3178×3). If, as an alternative, clustering had been employed on the initial data, then some form of standardization would have been necessary.

The minimum variance hierarchical method was used. This was because this method is recommendable for determining cohesive groups and also because an efficient and storage-economic algorithm was available (Murtagh 1985). This method took the longest time of all methods used: to cluster the 3178 objects, about 10–15 minutes elapsed time was required on a VAX 8600. A complete hierarchy of partitions is provided by such a method. Knowing that three classes were primarily of interest, the three-cluster partition alone was examined.

The galactic longitude and latitude coordinates associated with the 3178 objects were plotted for the three different groups found. These are shown in *Fig. 2*. As can be seen, from a visual point of view they are quite satisfactory.

To assess these results, a multiple discriminant analysis (MDA; or canonical discriminant analysis) was carried out. MDA may be informally described as a PCA on the

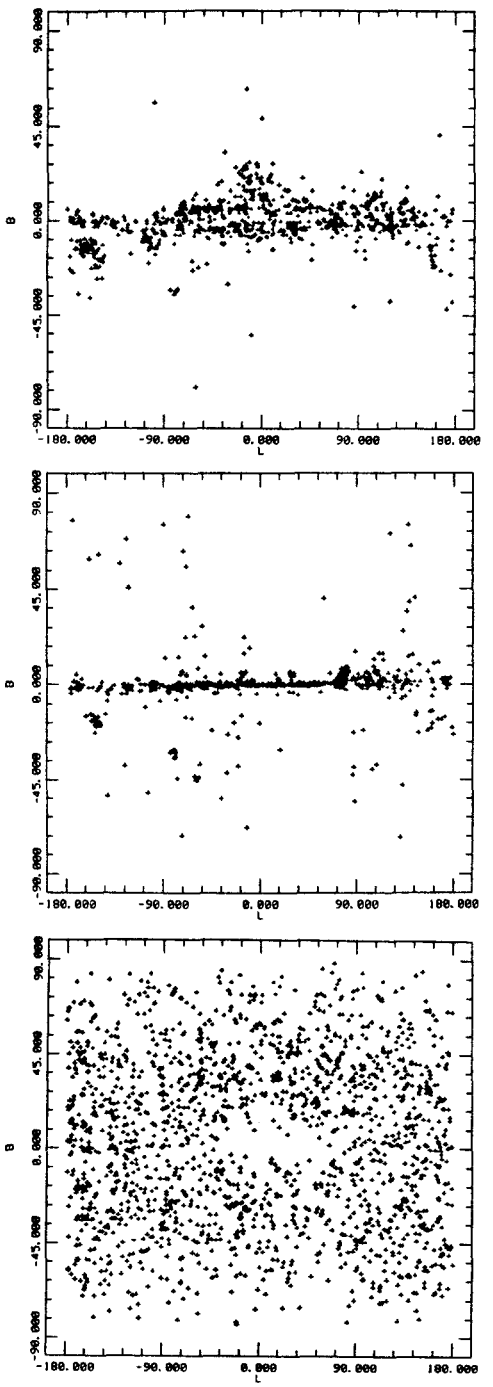

Fig. 2. Galactic longitude and latitude locations of (from top down) "galaxy","thin plane" and "cirrus" groups.

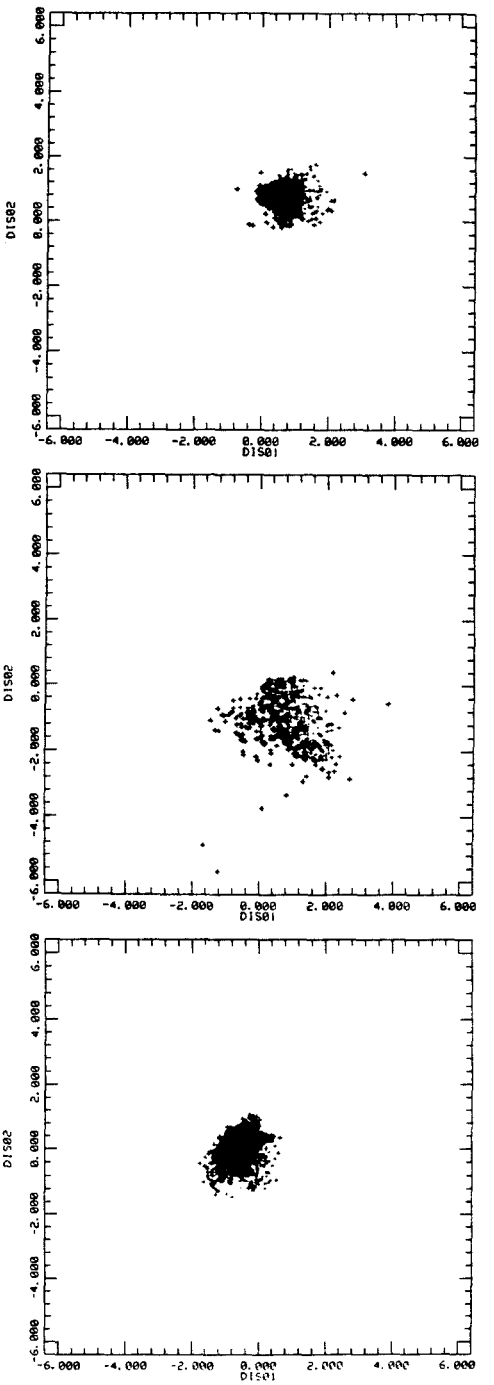

Fig. 3. Discriminant factor plane showing (from top down) "galaxy","thin plane" and "cirrus" groups.

groups. It attempts to separate the groups optimally using hyperplanes. The MDA implemented took the known assignments of the 3178 objects to the three groups (found by the cluster analysis above) in the original four–dimensional space (c_{12}, c_{23}, c_{34} and $\log f_{100}$). This was done because conclusions derived would be more meaningful for these parameters rather than the derived principal components.

Two discriminant factors (*i.e.* axes) of nearly equal "discriminating power" (measured by the eigenvalues) were obtained. Discriminant factors 1 and 2 were found to be defined as

$$f_1 = -0.5c_{12} + 0.1c_{23} - 1.8c_{34} + 0.1 \log f_{100}$$
$$f_2 = 0.9c_{12} - 2.0c_{23} - 0.5c_{34} - 0.0 \log f_{100}$$

The above equations indicate the relative importance of these parameters for these discriminating factors. The projections of the three groups in the plane defined by discriminant factors 1 and 2 are shown in *Fig. 3* (f_1 and f_2, above, are denoted by *DIS*01 and *DIS*02). Some overlap is evident, but also some clear "regions" of unequivocal group membership. Such unequivocally classed objects could be used as "pure" samples of "galaxy", "thin plane" or "cirrus".

It may be noted in the above equations that the parameter $\log f_{100}$ is almost irrelevant from the point of view of the discriminating factors. This points to the redundancy of this parameter if one uses the above equations for the assignment of new objects to one or other of the three classes.

The most time consuming algorithm used was the cluster analysis one. Its computational requirements are $O(n^2)$ when n objects are being classified. Although a sample of 3000-odd objects was used in the foregoing, a clustering was also carried out on the 30 000–odd set of objects. Clustering (hierarchic clustering using Ward's minimum variance criterion) of 31 760 objects took, on a VAX 8600 machine, 11.11 hours of CPU time. Note that (i) most widely available unsupervised clustering methods require in-core storage of dissimilarities and therefore would not work for such a large number of objects; and (ii) no special speed-up techniques were availed of (Murtagh 1985).

Acknowledgements

H.–M. Adorf and E.J.A. Meurs made their data available for the application described here, and helpfully discussed the issues involved.

References

Adorf, H.–M., Meurs, E.J.A., 1988. *These proceedings*, p. 315.

Barrow, J.D., Bhavsar, S.P., Sonoda, D.H., 1985. *Mon. Not. R. astr. Soc.*, **216**, 17.

Meurs, E.J.A., Adorf, H.–M., Harmon, R.T., 1988. In *Astronomy from Large Databases*, eds. Murtagh, F., Heck, A., European Southern Observatory, Garching, p. 49.

The MIDAS Users Guide, 1985 (and subsequent versions). Chapter 13, *Multivariate Statistical Methods*, European Southern Observatory, Garching.

Murtagh, F., 1985. *Multidimensional Clustering Algorithms*, Physica–Verlag, Heidelberg and New York.

Murtagh, F., Heck, A., 1987a. *Astr. Astrophys. Suppl.*, **68**, 113. (*copy of bibliography available from author*).

Murtagh, F., Heck, A. 1987b. *Multivariate Data Analysis*, Kluwer Academic, Dordrecht.

Rohlf, F.J., 1978. *Information Processing Letters*, **7**, 44.

Schulman, L.S., Seiden, P.E., 1986. *Science*, **233**, 425.