CLUSTER ANALYSIS BY LEAST MEANS
Warren S. Sarle, SAS Institute Inc.

Introduction
Most existing clustering methods have only heuristic justification. A more satisfying approach statistically is to define clustering models and to fit the models to data using criteria such as least squares (LS) or maximum likelihood (ML). The data analyst can then assess the appropriateness of the model and the fitting criterion for the data and the purpose of the analysis.

ML estimation of clusters has been discussed by Wolfe (1970), Hartigan (1975), Oehlert (1979), and Symons (1981), with the multivariate normal mixture model receiving the most attention. For most clustering problems, however, it is difficult to devise a plausible probability model. LS methods often provide useful results when a satisfactory probability model is not available and are especially valuable for summarizing large data sets.

Clustering Models for Rectangular Data
For rectangular multivariate data, the usual regression equation \( Y = XB + E \) becomes a clustering model if we treat the \( X \) values, as well as the \( R \) values, as parameter estimates instead of independent observations, and if we place appropriate constraints on the estimated \( X \) and/or \( B \) values.

A model equation for rectangular data is thus:

\[
Y = XB + E
\]

where

\[
Y = n \times p \text{ variables},
\]

\[
X, B \text{ are to be determined by LS},
\]

\[
E = \text{residuals},
\]

and trace\((E'E)\) is to be minimized.

The well-known k-means method for disjoint clusters of observations, implemented in SAS 82 by the FASTCLUS procedure, is obtained from the constraints:

\[
x_{ij} = 0, 1
\]

\[
\theta(x_{ij}) = 1
\]

where "#" is read "the number of." For example:

\[
\begin{array}{cccccc}
2 & 3 & 1 & 0 & 0 & 0 \\
3 & 2 & 1 & 0 & 0 & 0 \\
2 & -3 & 0 & 1 & 0 & 2.5 \ 2.5 \\
3 & -2 & 0 & 1 & 0 & 2.5 \ -2.5 \\
& -1 & 2 & 0 & 0 & 1 \ -1.5 \ 1.5 \\
& -2 & 1 & 0 & 0 & 1 \ -1.5 \ -1.5 \\
& -1 & -2 & 0 & 0 & 1 \\
& -2 & -1 & 0 & 0 & 1 \\
\end{array}
\]

\(X\) indicates the clusters and \(B\) gives the cluster means. \(X\) is simply a design matrix for a one-way design, and we can view the cluster analysis as an attempt to discover what design is most consistent with the data.

Ward’s method, available in the CLUSTER procedure in SAS 82, is a hierarchical algorithm for fitting the same model as k-means.

The LS criterion for the k-means model is equivalent to ML if it is assumed that the observations are sampled with equal probability from population clusters that are spherical multivariate normal with equal variances.

Oblique component clustering, provided by the SAS procedure VARCLUS, obtains disjoint clusters of variables. Assuming without loss of generality that the variables have mean zero, the constraints are:

\[
\#(b_{jk}) = 0 = 1
\]

For example:

\[
\begin{array}{cccccc}
2 & 1 & 0 & 1 & 0.37 & 1.12 \\
1 & 2 & 1 & 0 & 0.37 & 1.12 \\
0 & 0 & 2 & 2 & 1.49 & 0 \ 0.447 \ .447 \ 0 \ 0 \\
0 & 0 & -2 & -2 & -1.49 & 0 \ 0 \ 0.447 \ .447 \\
-1 & -2 & -1 & 0 & -0.37 & -1.12 \\
-2 & -1 & 0 & -1 & -0.37 & -1.12 \\
\end{array}
\]

Non-zero elements of \(B\) indicate the clusters of variables. \(X\) contains the first principal component of each cluster.

Factorial clustering, not yet available in SAS, produces overlapping clusters of observations. The constraints are:

\[
x_{ij} = 0, 1
\]

651
For example:

\[
\begin{array}{cccccc}
2 & 3 & 4 & 1 & 2 & 1 \\
3 & 2 & 1 & 1 & 1 & 1 \\
2 & -3 & 1 & 1 & 0 & -1.5 \\
3 & -2 & 1 & 1 & 0 & 4.0 \\
-2 & 0 & 1 & 1 & 1 & 1 \\
-1 & 2 & 1 & 1 & 0 & 4.0 \\
-2 & -1 & 1 & 1 & 0 & 4.0 \\
\end{array}
\]

X indicates clusters and is a design matrix for a 2\(^9\) factorial design. B gives the "effects" of the clusters in the ANOVA sense. A cluster containing all objects will provide an intercept. As with k-means, factorial clustering can be viewed as an attempt to determine what design would be most consistent with the observed data. Factorial clustering can also be considered a discrete analog of factor analysis with binary latent variables, the reverse of latent structure analysis.

Both observations and variables can be clustered simultaneously using the equation:

\[Y = XBZ + E\]

where constraints on X and Z determine the types of clusters. Hartigan (1975) has proposed some related methods.

**Clustering Models for Square Data**

Sometimes square data matrices are encountered, in which both rows and columns correspond to a single set of objects, rather than to objects and variables. The data may be distances, similarities, or other symmetric or asymmetric relations between the objects.

A model equation for square data is:

\[Y = XB' + E\]

\[\begin{array}{cccccc}
\text{n} & \text{x} & \text{p} & \text{x} & \text{q} & \text{x} & \text{p} \\
\end{array}\]

where

- X = measures of some relation among m objects.
- X, B are to be determined by LS.
- E = residuals.
- and trace(E'E) is to be minimized.

If we apply the constraints:

\[x_{ij} = 0,1\]

\[\#(x_{ij}=1) = 1\]

\[j\]

we obtain a model for disjoint clusters that does not yet have a name, although Forrest Young has proposed NEMOCLUS as the name of an as yet unwritten SAS procedure to fit the model. For example:

\[
\begin{array}{cccccc}
1 & 2 & 7 & 9 & 10 & 26 \\
3 & 2 & 8 & 8 & 10 & 0 \\
4 & 5 & 3 & 0 & 1 & 1 \\
6 & 5 & 5 & 3 & 0 & 1 \\
\end{array}
\]

X indicates the clusters. Each element of B is the mean of a submatrix of X determined by the clusters. If the data are asymmetric, then B will also be asymmetric. Missing data can be accommodated easily.

This model is appropriate for data that are neither similarities nor distances but express the strength of some relation between the row and column objects, such as export/import data between countries. Countries with similar patterns of exports and imports would be clustered together, rather than countries with high or low levels of mutual trade. A version of this model for binary data (directed graphs) was proposed by Oehlert (1979).

The ADCLUS model of Shepard and Arabie (1979) requires a square symmetric similarity matrix, the diagonal elements of which are ignored. ADCLUS produces overlapping clusters of objects. The constraints are:

\[x_{ij} = 0,1\]

\[b_{jk} = 0 \text{ if } j \neq k\]

\[\geq 0 \text{ if } j = k\]

For example:

\[
\begin{array}{cccccc}
. & 8 & 3 & 0 & 0 & 1110 \\
8 & 4 & 1 & 1 & 1111 & 1 \\
3 & 4 & 1 & 1 & 1111 & 1 \\
0 & 1 & 1 & 1 & 1010 & 1 \\
0 & 1 & 1 & 1 & 1010 & 1 \\
\end{array}
\]

X indicates the clusters. B gives the increment in the similarity of two objects accounted for by each cluster. The similarity of objects i and j is approximated by the sum of b_{kk} for all clusters k containing both i and j. A cluster containing all objects will provide an intercept.

Many generalizations of ADCLUS are possible. One approach, tentatively called generalized additive clustering or GADCLUS, allows the similarity matrix to be asymmetric and is given by the equation:

\[Y = XB'X' - (J-X)B'X'\]

\[+ Xb_{kk}(J-X)'\]

\[+ (J-X)B_{kk}(J-X)'\]

\[+ E\]
where \( J \) is a matrix of ones and \( B_1 \) through \( B_4 \) are all diagonal matrices. For the equation as given, the estimates of the matrices are not unique, but one or more of the four terms may be dropped to yield several different well-determined models.

Clustering procedures in SAS

All current clustering procedures in SAS and several proposed procedures are summarized in the accompanying table. The proposed procedures may not be implemented exactly as described here, and no commitment is made that such procedures will be provided by SAS Institute.

ADCLUS*, GADCLUS, FASTCLUS (k-means), CLUSTER (Ward's method), MADCLUS (factorial clustering), and VARCLUS (oblique component clustering) all involve LS methods as described above. DISCLUS would be oriented toward non-metric data. IPFPHC uses a variant of single-linkage for asymmetric similarity data (Hubert, 1973). HICLUS would implement the standard hierarchical methods. MULTICLUS would do ML estimation for multivariate normal mixtures. FACTOR uses many of the usual criteria for factor rotation.

References


Hubert, L. (1973), "Min and Max Hierarchical Clustering using Asymmetric Similarity Measures." Psychometrika, 38, 63-72.


**CLUSTERING PROCEDURES IN SAS**

Proposed procedures are in parentheses

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Type of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disjoint</td>
</tr>
<tr>
<td>Distance or Similarity Data</td>
<td>DISCLUS</td>
</tr>
<tr>
<td>M</td>
<td>Clustering Observations</td>
</tr>
<tr>
<td>L</td>
<td>t</td>
</tr>
</tbody>
</table>

*Name to be changed in SAS 82.