Appendix 1

Notes on Program Multimix

This program fits a mixture of multivariate distributions using the EM algorithm. The models that can be fitted are multivariate normal, latent class and location models having one categorical attribute. The data file can contain both categorical and continuous data, or either of these data types.

Note: The desired form of the data matrix is to have attributes in a partition cell being contiguous. To achieve this, the data is read in by specifying the column of the data array into which the J^{th} attribute of the data file is stored in an order variable, JP(J). All further references to the attribute J, refer to the rearranged order of the attributes.

The program currently has a maximum of

1500 observations	(IOB = 1500)
6 groups	(IK6 = 6)
15 attributes and partition cells	(IP15 = 15)
10 levels of categories	(IM10 = 10)
200 iterations to convergence	(ITER = 200)

NB. If these parameters are altered, remember to alter parameters (IK6 and IP15) in the subroutine, DETINV.

The parameter file contains:-

NG - the number of groups (distributions) in the finite mixture to be fitted.
NOBS - the number of observations.
NVAR - the number of attributes.
NPAR - the number of partition cells (sets of attributes associated within each distribution).
JP(J) - The column of the data array into which the J^{th} attribute of the data file will be stored, J = 1, NVAR. For example, suppose that we want the 3rd attribute in the first column, attribute 4 in the second column, attribute 7 in the 3rd column, and then attributes 1, 2, 5 and 6. Then JP(J) = 4 5 1 2 6 7 3, for $\text{J} = 1, \ldots, 7$.
IP(L) - the number of attributes in the L^{th} partition cell, $\text{L} = 1, \ldots, \text{NPAR}$.
IPC(L) - number of continuous attributes in the L^{th} partition cell.
ISV(L) - partition cell L starts at attribute J.

e.g. if attributes 6, 7, and 8, are in the same partition cell, then ISV(L) = 6, and IEV(L) = 8.

IEV(L) - partition cell L ends at attribute J.

IPARTYPE(L) - indicator giving the type of model for each partition cell.

$$\text{IPARTYPE(L)} = \begin{cases} 1 & \text{for a categorical model;} \\ 2 & \text{for a multivariate normal model;} \\ 3 & \text{for a location model.} \end{cases}$$

IVARTYPE(J) - an indicator for the type of each attribute

$$\text{IVARTYPE(J)} = \begin{cases} 1 & \text{for a categorical attribute;} \\ 2 & \text{for a multivariate normal attribute;} \\ 3 & \text{for a categorical attribute in a location model;} \\ 4 & \text{for a multivariate normal attribute in a location model,} \end{cases}$$

NCAT(J) - the number of categories for the J^{th} categorical attribute. For continuous attributes, NCAT(J) is entered as 0.

ISPEC - indicator variable determining whether the observations are specified into groups.

$$\text{ISPEC} = \begin{cases} 1 & \text{observations are not specified into groups;} \\ 2 & \text{observations are specified into groups.} \end{cases}$$

(1) ISPEC = 1 — read in the estimates of the parameters.

PI(K)- estimated mixing proportions for each group.

THETA(K,J,M) - estimated probability that the J^{th} categorical attribute is at level M, given that in group K

EMU(K,L,J) - estimated mean vector for group K, partition cell L and attribute J.

EMUL(K,L,J,M) - estimated mean vector for group K, partition cell L, attribute J, at the M th level of the categorical attribute in the location model.

VARIX(K,L,I,J) - estimated covariance between attributes I and J for group K, partition cell L where $\text{I} = 1, \ldots, \text{IPC(L)}$ and $\text{J} = 1, \ldots, \text{IPC(L)}$.

Note: The parameters that are required, are read in for each partition cell, $\text{L} = 1, \ldots, \text{NPAR}$. For example, if the attributes within the partition cell are all categorical, that is, ITYPE(L) = 1, then $\text{THETA}(\text{K}, \text{J}, \text{M})$, for $\text{M} = 1, \ldots, \text{NCAT(J)}$ is required for the attribute in that partition cell.

If the attributes within the partition cell are continuous, multivariate normal attributes, that is, ITYPE(L) = 2, then estimates of EMU(K,L,J) are required for each attribute.

If the attributes within the partition cell follow the location model, that is, ITYPE(L) = 3, then $\text{THETA}(\text{K}, \text{J}, \text{M}), \text{M} = 1, \ldots, \text{NCAT(J)}$ is required for the categorical attribute, and $\mathrm{EMUL(K, L, J, M), M = 1, \ldots, IM(L)}$ is required for each continuous multivariate normal attribute. (Note that IM(L) is the number of categories of the categorical attribute associated with the location model.)

The estimates are read in for group 1, and then for group 2 etc.

(2) ISPEC = 2, read in

IGRP(I) - variable specifying which group each observation is in. After reading in this variable, the program proceeds to the M step to calculate estimates of the parameters.

Creating an input file

A FORTRAN program (READ3.FOR) has been written to help set up the parameter input file for program MULTIMIX. If ISPEC = 2, that is, the grouping of the data is specified, the program requires the grouping to be in a separate file. This file must be in existence before the program for creating the parameter input file is run. This program has a very basic error subroutine to check whether the number of attributes in each partition cell matches with the total number of attributes, and the type of each attribute matches with the partition cell type. This program is being extended to facilitate ease of input.

Format of the parameter file

Format is free field.
Input:-
NG NOBS NVAR NPAR ISPEC
(JP(J), J = 1, NVAR)
(IP(L), L = 1, NPAR)
(IPC(L), L = 1, NPAR)
(ISV(L), L = 1, NPAR
(IEV(L), L = 1, NPAR)
(IPARTYPE(L), L = 1, NPAR)
(IVARTYPE(J), J = 1, NVAR)
(NCAT(J), J = 1, NVAR)

If ISPEC = 1, read in estimates of the parameters. See section below.
If ISPEC = 2, read in specified grouping of observations
(IGRP(I), I = 1, NOBS)

Estimates of the parameters

(PI(K), K = 1, NG)
****
For each partition cell, L = 1, NPAR read in the required parameters.
(THETA(K,J,M), M = 1, NCAT(J)) - for categorical attributes
- repeat for each attribute, J = ISV(L), IEV(L)
(EMU(K,L,J), J = 1, IPC(L)) - for the multivariate normal model
(THETA(K,J,M), M = 1, NCAT(J)) - for the categorical attribute in location model
(EMUL(K,L,J,M), M = 1, IM(L)) for each continuous attribute J = 1, IPC(L) in the location model

Repeat from **** for each group.
Read in the estimates of the variance for each partition cell. (For continuous attributes only)
((VARIX(K,L,I,J), J = 1, IPC(L)), I = 1, IPC(L)) - repeat for each group.