multimix, multimix-prep – automatically discover classes in data

multimix fits a mixture of multivariate distributions to a set of observations using the EM algorithm. The data file may contain both categorical and continuous variables. **multimix** prompts for the names of the data and parameter files. The assignment of the observations to groups and the posterior probabilities are written to *GROUPS.OUT*. Parameter estimates, convergence information, and group assignment probabilities are written to *GENERAL.OUT*. If **multimix** does not converge after *ITER*=200 iterations, the estimates of the parameters will be written to *EMPARAMEST.OUT*. This file can then be used as the parameter input file for **multimix** if desired. **multimix** is limited to 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)

Recompilation is required to change these parameters. The data file has one line for each observation. Each line has one entry for each variable. Only the first *NVAR* entries on each line are read. The parameter file contains free field values which describe the data and the fitting models. **multimix-prep** will ask the user a series of questions and write a suitable parameter file. (The experienced user finds it faster to edit old parameter files into new ones.)

multimix requires variables in a partition to be stored contiguously. Hence the data is read in with the variable order being specified by *JP*(J). *INTYPE*(J) and *NCAT*(J) both refer to the rearranged data. The first five values are The number of groups (distributions) in the finite mixture to be fitted. The number of observations. The number of attributes. The number of partition cells (sets of attributes associated within each distribution). Flag indicating how the starting point is specified for the fit:

1 Initial parameter estimates are specified.

2 Observations are specified into groups. Next come eight arrays of data: is the column of the data array into which the *J*th attribute of the data file will be stored, where *J* varies from 1 to *NVAR*. For example, suppose we want the third 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 J=1,...,7. is the number of attributes in the *L*th partition cell, *L*=1,...,*NPAR*. is the number of continuous attributes in the *L*th partition cell. gives the index *J* of the start of partition cell *L*. E.g. if attributes 6, 7, and 8 are in the same partition cell *L*, then ISV(L)=6 and IEV(L)=8. gives the index *J* of the end of partition cell *L*. is an indicator giving the type of model for partition *L*:

1 for a categorical model.
2 for a multivariate normal model.
3 for a location model. is an indicator for the type of attribute *J*:
1 for a categorical attribute.
2 for a multivariate normal attribute;
3 for a categorical attribute in a location model;
4 for a multivariate normal attribute in a location model. is the number of categories for the *J*th categorical attribute. For continuous attributes, *NCAT*(*J*) should be 0. If observations are specified into groups (*ISPEC*=2), then the groups are next: is the index of the group that observation *I* is in. If observations are not specified into groups (*ISPEC*=1), then estimates of the parameters are next: is the estimated mixing proportion for group *K* The parameters for each group depend on the type of attribute: is the estimated probability that the *J*th categorical attribute is at level *M*, given that in group *K*. Repeat for each attribute, **categorical attributes only** is the estimated mean vector for group *K*, partition cell *L* and attribute *J*. **multivariate normal model only** is the estimated probability that the *J*th categorical attribute in the location model is at level *M*, given that in group *K*. **categorical attributes only** is the estimated mean vector for group *K*, partition cell *L* and attribute *J*, at the *M*th level of the categorical attribute in the location model. **multivariate normal model only** An entry in *VARIX* is the estimated covariance between attributes *I* and *J* for group *K*, partition cell *L*, where The required parameters are read in for each partition cell, For example, if the attributes within the partition cell are all categorical, that is, then for is required for the attribute in that partition cell. If the attributes within the partition cell are continuous, multivariate normal attributes, that is then estimates of are required for each attribute. If the attributes within the partition cell follow the location model, that is, then is required for the categorical attribute, and is required for each continuous multivariate normal attribute. (Note that 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. See */usr/share/doc/multimix/examples*. *GROUPS.OUT* **multimix** output on success: the assignment of the observations to groups and the posterior probabilities. *EMPARAMEST.OUT*. **multimix** output on failure to converge: current parameter estimates. This file can then be used as the parameter input file for **multimix** if desired. Lynette A. Hunt <lah@waikato.ac.nz> and Murray Jorgensen <maj@waikato.ac.nz>.