"Combining Mixture Components for Clustering" published in the Journal of Computational and Graphical Statistics. Instructions for the provided Matlab functions.

Jean-Patrick Baudry
Université Paris-Sud XI

Adrian E. Raftery
University of Washington

Gilles Celeux
INRIA

Kenneth Lo
University of Washington

Raphaël Gottardo
University of British Columbia

March 29, 2018

1 Description of the files

The following files are provided. They have been developped and tested with Matlab R2007b, with the Statistics toolbox. An example on how they should be used is given bellow.

Let us denote, as in the paper

- d for the data dimension ;
- n for the sample size ;

MixCombi.m is the main file. The user provides the BIC (and optionally the ICL solution) solution and the program computes and returns the whole hierarchy obtained by combining from the BIC solution, as described in the paper. The user may get the BIC solution for example from the Mixmod (`http://www.mixmod.org/`, see *Run_Mixmod2* bellow) or the Mclust (`http://www.stat.washington.edu/mclust/`) softwares. The following variables must be created by the user:

- *exp.data.obs* ($n \times d$ matrix which contains the sample) ;
- *exp.data.BIC.K* (number of components selected by BIC, denoted by K_{BIC} from now on) ;
- *exp.data.BIC.mu* ($K_{BIC} \times d$ matrix with the K_{BIC} Gaussian components mean parameters of the BIC solution) ;
- *exp.data.BIC.S* ($d \times d \times K_{BIC}$ matrix with the K_{BIC} Gaussian components Covariance matrices of the BIC solution) ;

- *exp.data.BIC.p* ($1 \times K_{BIC}$ vector with the K_{BIC} Gaussian components mixing proportions of the BIC solution).

The user may also provide the ICL solution (optional):

- *exp.data.ICL.K* (number of components selected by ICL, denoted by K_{ICL} from now on) ;
- *exp.data.ICL.mu* ($K_{ICL} \times d$ matrix with the K_{ICL} Gaussian components mean parameters of the ICL solution) ;
- *exp.data.ICL.S* ($d \times d \times K_{ICL}$ matrix with the K_{ICL} Gaussian components Covariance matrices of the ICL solution) ;
- *exp.data.ICL.p* ($1 \times K_{ICL}$ vector with the K_{ICL} Gaussian components mixing proportions of the ICL solution).

The program returns a K_{BIC} structure exp.res.combi with the labels, posterior probabilities (denoted by *tau*), entropies for each combined solution, as well as a matrix denoted by M for each of those solutions, which describes the $K+1 \to K$ combining step: for example, $exp.res.combined(3).M = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ means that the three-component combined solution is obtained from the four-component solution by merging the first and the third of its components.

PlotResults plots the results obtained from *MixCombi.m* in case $d = 2$. Analogous functions are provided for the cases $d = 3$ (3D plots) and $d = 4$ (plot of the two first coordinates, which may be easily changed in the code). The labels and titles of the plots should be explicit enough.

PlotEntropy plots all the entropy plots (simple, rescaled, etc.) considered in the paper, from the results obtained from *MixCombi.m*. The labels of the plots should be explicit enough.

Run_Mixmod2 Runs Mixmod (version 2.1.1) and returns the results (BIC solution) in a structure which can be directly used in *MixCombi.m*. Mixmod has to be installed and the Mixmod paths lines 16,17 in the file have to be updated according to your installation. The following variables must be created by the user:

- *exp.data.obs* ($n \times d$ matrix which contains the sample) ;
- *exp.cond.Kmin* (minimum number of Gaussian components) ;
- *exp.cond.Kmax* (maximum number of Gaussian components) ;
- *exp.cond.models* (model to be considered, according to the Mixmod syntax).

Data The files *4.1.mat*, *4.2.mat*, *4.3.mat*, *4.4.1.mat*, *4.4.2.mat*, *GvHD-.mat*, *GvHD+.mat* contain the data and BIC solutions corresponding to the examples reported in the paper.

func repertory: contains functions called by the preceding files. Some are more or less documented: for example,
>> *help MAP_combi*
will return a brief description of the MAP_combi function. But it should not be necessary to directly use those functions at first.

2 Example

Here is an example on how the first example reported in the paper (section 4.4.1) may be reproduced by readers.

>> clear % Clear all

>> load 4.1.mat % Load the data, BIC and ICL solutions

>> exp % display the loaded structure

>> MixCombi % computes and display the combined solutions

'Criterion'	'K'	'ENT'	'Lcc'
'ICL'	'4'	'3'	'-2044'
'Combined K=2'	'2'	'0'	'-1951'
'Combined K=3'	'3'	'1'	'-1952'
'Combined K=4'	'4'	'5'	'-1956'
'Combined K=5'	'5'	'41'	'-1992'
'BIC'	'6'	'122'	'-2074'

>> PlotResults % Plot the data, BIC and ICL solutions and the combined solutions.

>> PlotEntropy % Plot the entropy plots...

All illustrations in the paper may be reproduced the same way. Think of using *PlotResults_3D* (resp. *PlotResults_4D*) for the "3D uniform cross" example (resp. the GvHD datasets).