MIXMOD Statistical Documentation

May 15, 2008

Contents

1 Introduction

Finite mixture models is a powerful tool for density estimation, cluster analysis and discriminant analysis. MIXMOD is a software for MIXture MODelling which considered those three different aspects of mixtures and gives great place to the multivariate context. In its present version, MIXMOD is dealing with multivariate Gaussian mixture models for quantitative data and multivariate multinomial mixture models for qualitative data. Basing cluster or discriminant analysis on Gaussian mixture models is a classical and powerful approach since Gaussian models are useful both for understanding and suggesting powerful clustering criteria. One of the originality of MIXMOD is to consider a parameterization of the variance matrix of a cluster through its eigenvalue decomposition leading to many meaningful models for clustering and classification. In the same manner, different more or less parsimonious parameterizations are entering in the multinomial mixture models.

This documentation is organized as follows. In Section 2, the general setting of finite mixture modelling is sketched. In Section 3, the different available algorithms in MIXMOD for estimating mixture parameters are presented. In Section 4, the possible strategies for using MIXMOD algorithms and for initiating them are described. Moreover criteria to select a model are presented. Section 5 is devoted to the detailed presentation of the Gaussian mixture models considered in MIXMOD and to a mixture of factor analyser models useful to treat high dimensional supervised classification problems. Section 6 is devoted to the detailed presentation of multivariate multinomial mixture models.

2 Mixture model

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be n independent vectors in $\mathbb{R}^d$ such that each $\mathbf{x}_i$ arises from a probability distribution with density

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k) \tag{1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, ..., K$ and $p_1 + ... + p_K = 1$), $h(\cdot|\boldsymbol{\lambda}_k)$ denotes a d-dimensional distribution parameterized by $\boldsymbol{\lambda}_k$. As we will see in Section 5, h is for instance the density of a Gaussian distribution with mean $\boldsymbol{\mu}$ and variance matrix Σ and thus, $\boldsymbol{\lambda} = (\boldsymbol{\mu}, \Sigma)$.

It is worth noting that for a mixture distribution, a sample of indicator vectors or *labels* $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, $z_{ik} = 1$ or 0, according to the fact that $\mathbf{x}_i$ is arising from the kth mixture component or not, is associated to the observed data $\mathbf{x}$. The sample $\mathbf{z}$ can be *known* in which case we are in a discriminant analysis context where the problem is essentially to predict an indicator vector $\mathbf{z}_{n+1}$ from a new observed data vector $\mathbf{x}_{n+1}$. But the sample $\mathbf{z}$ can be *unknown* in which case we are in a density estimation context or cluster analysis context if the estimation of the $\mathbf{z}_i$'s are of primary interest. In each case, the vector parameter to be estimated is $\theta = (p_1, \ldots, p_K, \lambda_1, \ldots, \lambda_K)$.

2.1 Density estimation from a mixture model

Mixture modelling can be regarded as a flexible way to represent a probability density function, and thus providing a semi parametric tool for density estimation. When the labels $\mathbf{z}$ are unknown, maximum likelihood estimation of mixture models can be performed in MIXMOD via the EM algorithm of Dempster, Laird and Rubin (1977) or by a stochastic version of EM called SEM (see McLachlan and Peel, 2000). In each case, the parameter θ is chosen to maximize the observed log-likelihood

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p_k h(\mathbf{x}_i, \boldsymbol{\lambda}_k) \right). \tag{2}$$

2.2 Clustering with mixture model

Cluster analysis is concerned with discovering a group structure in a n by d data matrix $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i$ is an individual in $\mathbb{R}^d$. Consequently, the result provided by clustering is typically a partition of $\mathbf{x}$ into K groups defined with the labels $\tilde{\mathbf{z}} = \{\tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_n\}$, with $\tilde{\mathbf{z}}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{iK})$, $\tilde{z}_{ik} = 1$ or 0 according to $\mathbf{x}_i$ is assigned to the kth group or not.

Many authors have considered non hierarchical clustering methods in which a mixture of distributions is used as a statistical model. In this context, two commonly used maximum likelihood (m.l.) approaches have been proposed: the mixture approach and the classification approach. Loosely speaking, the mixture approach is aimed to maximize the likelihood over the mixture parameters, whereas the classification approach is aimed to maximize the likelihood over the mixture parameters and over the mixture component labels.

2.2.1 Mixture approach

In this approach, a partition of the data can directly be derived from the m.l. estimates $\hat{\theta}$ of the mixture parameters obtained, for instance, by the EM or the SEM algorithm described hereafter, by assigning each $\mathbf{x}_i$ to the component providing the largest conditional probability that $\mathbf{x}_i$ arises from it using a MAP (Maximum A Posteriori) principle. Denoting $\mathbf{z}_i$ the label of $\mathbf{x}_i$, the MAP principle is as follows

$$\tilde{z}_{ik} = \begin{cases} 1 \text{ if } k = \arg\max_{\ell=1\ldots,K} t_\ell(\mathbf{x}_i|\hat{\theta}) \\ 0 \text{ if not} \end{cases} \tag{3}$$

where

$$t_k(\mathbf{x}_i|\hat{\theta}) = \frac{\hat{p}_k h(\mathbf{x}_i|\hat{\boldsymbol{\lambda}}_k)}{\sum_{\ell=1}^K \hat{p}_\ell h(\mathbf{x}_i|\hat{\boldsymbol{\lambda}}_\ell)}.$$

2.2.2 The Classification Approach

The second approach available in MIXMOD is the classification approach. In this approach, the indicator vectors $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, identifying the mixture component origin, are treated as unknown parameters. The Classification Maximum Likelihood (c.m.l.) method is used to estimate both the parameters θ and $\mathbf{z}$. The classification likelihood criterion is defined by

$$CL(\theta, \mathbf{z}_1, \ldots, \mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k)]. \tag{4}$$

The CL criterion can be maximized by making use of a classification version of the EM algorithm, the so-called CEM algorithm (Celeux and Govaert 1992) which includes a classification step (C-step) between the E and M steps.

2.3 Discriminant Analysis

When the labels $\mathbf{z}$ are known, we are concerned with discriminant analysis: in discriminant analysis, data are composed by n observations $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n\}$

($\mathbf{x}_i \in \mathbb{R}^d$) and a partition of $\mathbf{x}$ into K groups defined with the labels $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$. The aim is to estimate the group $\mathbf{z}_{n+1}$ of any new individual $\mathbf{x}_{n+1}$ of $\mathbb{R}^d$ with unknown label.

In this context, the n couples $(\mathbf{x}_i, \mathbf{z}_i)$,..., $(\mathbf{x}_n, \mathbf{z}_n)$ are realizations of n identically and independently distributed random vectors $(\mathbf{X}_i, \mathbf{Z}_i)$,...,$(\mathbf{X}_n, \mathbf{Z}_n)$. The distribution of each $(\mathbf{X}_i, \mathbf{Z}_i)$ $(1 \leq i \leq n)$ is

$$f(\mathbf{x}_i, \mathbf{z}_i|\theta) = \prod_{k=1}^{K} p_k^{z_{ik}} [h(\mathbf{x}_i|\boldsymbol{\lambda}_k)]^{z_{ik}}, \tag{5}$$

where p_k is the prior probability of the kth group (the mixing proportion), $h(\mathbf{x}_i|\boldsymbol{\lambda}_k)$ is a probability density with parameters $\boldsymbol{\lambda}_k$ and the whole parameter is $\theta = (p_1, \dots, p_K, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K)$.

An estimate $\hat{\theta}$ of θ is obtained by the m.l. method

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{x}, \mathbf{z}) \tag{6}$$

where the log-likelihood function $L(\theta|\mathbf{x}, \mathbf{z})$ is defined by

$$L(\theta|\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k)). \tag{7}$$

This estimate $\hat{\theta}$ allows to assign any new point $\mathbf{x}_{n+1}$ with unknown membership in one of the K groups by the maximum a posteriori (MAP) procedure. Computing the conditional probability $t_k(\mathbf{x}_{n+1}|\hat{\theta})$ that $\mathbf{x}_{n+1}$ arises from the kth group

$$t_k(\mathbf{x}_{n+1}|\hat{\theta}) = \frac{\hat{p}_k h(\mathbf{x}_{n+1}|\hat{\boldsymbol{\lambda}}_k)}{\sum_{\ell=1}^{K} \hat{p}_\ell h(\mathbf{x}_{n+1}|\hat{\boldsymbol{\lambda}}_\ell)}, \tag{8}$$

the MAP procedure consists of assigning $\mathbf{x}_{n+1}$ to the group maximizing this conditional probability, i.e.

$$\hat{z}_{n+1\ k} = \begin{cases} 1 \text{ if } k = \arg\max_{\ell=1...,K} t_\ell(\mathbf{x}_{n+1}|\hat{\theta}) \\ 0 \text{ if not} \end{cases} . \tag{9}$$

3 Algorithms in MIXMOD

3.1 EM algorithm

Starting from an initial arbitrary parameter θ^0, the mth iteration of the EM algorithm consists of repeating the following E and M steps.

- **E step:** The current conditional probabilities that $z_{ik} = 1$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ are computed using the current value θ^{m-1} of the parameter:

$$t_{ik}^m = t_k^m(\mathbf{x}_i|\theta^{m-1}) = \frac{p_k^{m-1} h(\mathbf{x}_i|\boldsymbol{\lambda}_k^{m-1})}{\sum_{\ell=1}^K p_\ell^{m-1} h(\mathbf{x}_i|\boldsymbol{\lambda}_\ell^{m-1})}. \qquad (10)$$

- **M step:** The m.l. estimate θ^m of θ is updated using the conditional probabilities t_{ik}^m as conditional mixing weights. It leads to maximize

$$F(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}^m) = \sum_{i=1}^n \sum_{k=1}^K t_{ik}^m \ln \left[p_k \Phi(\mathbf{x}_i|\boldsymbol{\lambda}_k) \right], \qquad (11)$$

where $\mathbf{t}^m = (t_{ik}^m, i = 1, \ldots, n, k = 1, \ldots, K)$. Updated expression of mixture proportions are, for $k = 1, \ldots, K$,

$$p_k^m = \frac{\sum_{i=1}^n t_{ik}^m}{n}. \qquad (12)$$

Detailed formula for the updating of the $\boldsymbol{\lambda}_k$'s are depending of the component parameterization $\boldsymbol{\lambda}$ and cannot be detailed here.

3.2 SEM algorithm

The SEM algorithm is a stochastic version of EM incorporating between the E and M steps a restoration of the unknown component labels $\mathbf{z}_i$, $i = 1, \ldots, n$, by drawing them at random from their current conditional distribution. Starting from an initial parameter θ^0, an iteration of SEM consists of three steps.

- **E step:** The conditional probabilities t_{ik}^m $(1 \leq i \leq n, 1 \leq k \leq K)$ are computed for the current value of θ as done in the E step of EM.

- **S step:** A partition $P^m = (P_1^m, \ldots, P_K^m)$ of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is designed by assigning each point $\mathbf{x}_i$ at random to one of the mixture components according to the multinomial distribution with parameter $(t_{ik}^m, 1 \leq k \leq K)$.

- **M step:** The m.l. estimate of θ is updated using the cluster P_k^m as sub-sample $(1 \leq k \leq K)$ of the kth mixture component. This step leads generally to simple formula. For instance,

$$p_k^m = \frac{\text{card}(P_k^m)}{n}. \tag{13}$$

SEM does not converge pointwise. It generates a Markov chain whose stationary distribution is more or less concentrated around the m.l. parameter estimator. A natural parameter estimate from a SEM sequence $(\theta^r)_{r=1,\ldots,R}$ is the mean $\sum_{r=b+1}^{R} \theta^r/(R-b)$ of the iterates values where the first b burn-in iterates have been discarded when computing this mean. An alternative estimate is to consider the parameter value leading to the highest likelihood in a SEM sequence.

A remark is to be made. When several observations are associated to the same vector, they are assigned to the same mixture component in the S step. This choice can make a difference when concerned with qualitative data. It is expected to give a larger influence to the random assignments.

3.3 CEM algorithm

This algorithm incorporates a classification step between the E and M steps of EM. Starting from an initial parameter θ^0, an iteration of CEM consists of three steps.

- **E step:** The conditional probabilities t_{ik}^m $(1 \leq i \leq n, 1 \leq k \leq K)$ are computed for the current value of θ as done in the E step of EM.

- **C step:** A partition $P^m = (P_1^m, \ldots, P_K^m)$ of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is designed by assigning each point $\mathbf{x}_i$ to the component maximizing the conditional probability $(t_{ik}^m, 1 \leq k \leq K)$.

- **M step:** The m.l. estimates $(\hat{p}_k, \boldsymbol{\lambda}_k)$ are computed using the cluster P_k^m as sub-sample $(1 \leq k \leq K)$ of the kth mixture component as done in the M step of SEM.

CEM is a *K-means*-like algorithm and contrary to EM, it converges in a finite number of iterations. CEM is not maximizing the observed log-likelihood L (2) but is maximizing in θ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ the complete data log-likelihood CL (4) where the missing component indicator vector $\mathbf{z}_i$ of each sample point is included in the data set. As a consequence, CEM is not expected to converge to the m.l. estimate of θ and yields inconsistent estimates of the parameters especially when the mixture components are overlapping or are in disparate proportions (see McLachlan and Peel 2000, Section 2.21).

4 Using MIXMOD

4.1 Stopping rules

In MIXMOD they are three ways to stop an algorithm.

- An algorithm can be stopped after a pre-defined number of iterations (100 by default in MIXMOD). This possibility is available for EM, SEM and CEM.
- An algorithm can be stopped using a threshold for the relative change of the criterion at hand (the likelihood L or the classification likelihood CL). This possibility is available with EM and CEM. It is not recommended since EM can encounter slow convergence situations and CEM is converging in a finite number of iterations.
- An algorithm can be stopped at stationarity. Obviously, this possibility is only available for CEM.

4.2 Initialization strategies

The solution provided by EM can highly depend on its starting position especially in a multivariate context. Thus, it is important to have sensible ways for initiating EM to get a sensible optimum of the likelihood. Obviously, in some cases it is possible to start from a particular partition of the data or from a pre-defined θ^0 and those initializations of EM are possible in MIXMOD. But there is the need to have more general strategies. In MIXMOD, it is possible to easily link the algorithms EM, SEM and CEM in all imaginable ways. Thus, in Biernacki *et al.* (2003), we have experimented an efficient three step Search/Run/Select (S/R/S) strategy for maximizing the likelihood:

1. Build a search method for generating p initial positions. This could be based on random starts or the output from an algorithm like a Classification EM (CEM) algorithm, a Stochastic EM (SEM) algorithm or

short runs of the standard EM algorithm. The parameter p is depending on an allotment of iterations.

2. Run the EM algorithm a set number of times at each initial position with a fixed number of iterations.

3. Select the solution providing best likelihood among the p trials, say θ^*.

This three-step strategy can be compounded by repeating the three steps x times and using the $\theta_1^*, \ldots, \theta_x^*$ as the starting positions in step 1. By compounding, one increases starting position variation, but one must decrease the length of the EM runs possible within the steps in order to fix the total number of steps.

Possible variants of this strategy are now described.

Random initialization Usually this random initial position is obtained by drawing at random component means in the data set. Since this is probably the most employed way of initiating EM, it can be regarded as a reference strategy. An extension of this simple strategy consists of repeating it x times from different random positions and selecting the solution maximizing the likelihood among those x runs. This "xEM" strategy is the basic S/R/S algorithm.

Using the CEM algorithm Runs of CEM from random positions followed by EM from the position providing the highest *complete* data log-likelihood obtained with CEM. And, x repetitions of the previous strategy give rise to an additional strategy denoted "xCEM-EM".

Using short runs of EM By a short run of EM, we mean that we do not wait for convergence and that we stop the algorithm as soon as

$$\frac{L^m - L^{m-1}}{L^m - L^0} \leq 10^{-2}, \tag{14}$$

L^m denoting the observed log-likelihood at mth iteration. Here 10^{-2} represents a threshold value which has to be chosen on a pragmatic ground. It leads to the following strategies : several short runs of EM from random positions followed by a long run of EM from the solution maximizing the *observed* log-likelihood. And, x repetitions of the previous strategy lead to the so called "$x$em-EM" strategy.

Using Stochastic EM The stochastic EM algorithm generates an ergodic Markov chain. Thus a sequence of parameter estimates via SEM is expected to visit the whole parameter space with long sojourns in the neighborhood of sensible maxima of likelihood functions. This characteristic of SEM leads to the following strategies.

- "SEMmean-EM": A run of SEM, followed by a run of EM from the solution obtained by computing the mean values of the sequence of parameter estimates provided by SEM after a burn-in period. The idea underlying this strategy is that SEM is expected to spend most of the time near sensible likelihood maxima with a large attractive neighborhood.
- "SEMmax-EM": The *same* run of SEM followed by a run of EM from the position leading to the highest maximum likelihood value reached by SEM. Here, the idea is that a SEM sequence is expected to enter rapidly in the neighborhood of the global maximum of the likelihood function.

It is difficult to recommend a particular strategy among the ones presented above. However, the strategy "$x$em-EM" gives generally good performances and is the default strategy in MIXMOD.

4.3 Criteria to select a model

It is of high interest to automatically select a model and the number K of mixture components. However, choosing a sensible mixture model is highly dependent of the modelling purpose.

In MIXMOD, two criteria are proposed in a supervised setting: BIC and cross-validation. In an unsupervised setting, three criteria are available: BIC, ICL and NEC. In a density estimation perspective, BIC must be preferred. But in a cluster analysis perspective, ICL and NEC can provide more parsimonious answers. Nevertheless, NEC is essentially devoted to choose the number of mixture components K, rather than the model parameterization. Before describing those criteria, it can be noted that if no information on K is available, it is recommended to vary it between 1 and the smallest integer larger than $n^{0.3}$ (see Bozdogan 1993).

4.3.1 The Bayesian Information Criterion (BIC)

A finite mixture model is characterized by the number of components K and the vector parameter $\theta = (p_1, \ldots, p_K, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$. A classical way of choosing a model is to select this one maximizing the integrated likelihood,

$$(\hat{m}, \hat{K}) = \arg\max_{m,K} \mathbf{f}(\mathbf{x} \mid m, K) \tag{15}$$

where the integrated likelihood is

$$\mathbf{f}(\mathbf{x} \mid m, K) = \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x} \mid m, K, \theta)\pi(\theta \mid m, K)d\theta, \tag{16}$$

with the likelihood

$$\mathbf{f}(\mathbf{x} \mid m, K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid m, K, \theta), \tag{17}$$

and $\Theta_{m,K}$ being the parameter space of the model m with K components and $\pi(\theta \mid m, K)$ a non informative or a weakly informative prior distribution on θ for this model. An asymptotic approximation of the integrated likelihood, valid under regularity conditions, has been proposed by Schwarz (1978)

$$\log \mathbf{f}(\mathbf{x} \mid m, K) \approx \log \mathbf{f}(\mathbf{x} \mid m, K, \hat{\theta}) - \frac{\nu_{m,K}}{2} \log(n), \tag{18}$$

where $\hat{\theta}$ is the m.l. estimate of θ

$$\hat{\theta} = \arg\max_{\theta} \mathbf{f}(\mathbf{x} \mid m, K, \theta) \tag{19}$$

and $\nu_{m,K}$ is the number of free parameters in the model m with K components. It leads to minimize the so-called BIC criterion

$$\text{BIC}_{m,K} = -2L_{m,K} + \nu_{m,K} \ln n, \tag{20}$$

where $L_{m,K} = \log \mathbf{f}(\mathbf{x} \mid m, K, \hat{\theta})$ is the maximum log-likelihood for m and K. Despite the fact that those regularity conditions are not fulfilled for mixtures, it has been proved that the criterion BIC is consistent (Keribin 2000) and has been proved to be efficient on a practical ground (see for instance Fraley and Raftery 1998).

4.3.2 The Integrated Completed Likelihood (ICL)

The use of the integrated likelihood (16) does not take into account the ability of the mixture model to give evidence for a clustering structure of the data. An alternative is to consider the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ (or integrated completed likelihood) (Biernacki *et al.* 2000)

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) = \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta) \pi(\theta \mid m, K) d\theta, \tag{21}$$

where

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i \mid m, K, \theta) \tag{22}$$

with

$$f(\mathbf{x}_i, \mathbf{z}_i \mid m, K, \theta) = \prod_{k=1}^{K} p_k^{z_{ik}} \left[h(\mathbf{x}_i \mid \boldsymbol{\lambda}_k)\right]^{z_{ik}}. \tag{23}$$

This integrated completed likelihood can be approximated from a BIC-like approximation. That is

$$\log \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) \approx \log \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \hat{\theta}^*) - \frac{\nu_{m,K}}{2} \log n \tag{24}$$

where

$$\hat{\theta}^* = \arg\max_{\theta} \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta). \tag{25}$$

But $\mathbf{z}$ is unknown. It means that the objective functions to be maximized in (21) and (25) are not available and so is $\hat{\theta}^*$. However, for n large enough, $\hat{\theta}^*$ can be approximated by the m.l. estimator $\hat{\theta}$. Moreover, the missing data $\mathbf{z}$ can be replaced using the MAP principle: $\tilde{\mathbf{z}} = \text{MAP}(\hat{\theta})$. It leads finally to the ICL criterion to be minimized (Biernacki *et al.* 2000)

$$\text{ICL}_{m,K} = -2 \log \mathbf{f}(\mathbf{x}, \tilde{\mathbf{z}} \mid m, K, \hat{\theta}) - \frac{\nu_{m,K}}{2} \log n, \tag{26}$$

that we can also write as a BIC criterion penalized by an entropy term:

$$\text{ICL}_{m,K} = \text{BIC}_{m,K} - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{z}_{ik} \ln t_{ik}. \tag{27}$$

4.3.3 The Normalized Entropy Criterion (NEC)

This entropy criterion measures the ability of a mixture model to provide well-separated clusters and is derived from a relation highlighting the differences between the maximum likelihood (m.l.) approach and the classification maximum likelihood (c.m.l.) approach to the mixture problem. Recall that NEC is essentially devoted to choose the number of mixture components K, not the model m.

We note $\hat{\theta}$ the m.l. estimator of θ and

$$t_{ik} = t_k(\mathbf{x}_i|\hat{\theta}) = \frac{\hat{p}_k h(\mathbf{x}_i|\hat{\boldsymbol{\lambda}}_k)}{\sum_{k'=1}^{K} \hat{p}_{k'} h(\mathbf{x}_i|\hat{\boldsymbol{\lambda}}_{k'})} \tag{28}$$

the associated conditional probability that $\mathbf{x}_i$ arises from to the kth mixture component. Direct calculations show that

$$L_K = C_K + E_K, \tag{29}$$

with L_K the maximum log-likelihood,

$$C_K = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ \ln \ [\hat{p}_k h(\mathbf{x}_i|\hat{\boldsymbol{\lambda}}_k)], \tag{30}$$

and

$$E_K = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ \ln \ t_{ik} \geq 0. \tag{31}$$

This relation with the fact that the entropy term E_K measures the overlap of the mixture components (If the mixture components are well-separated $E_K \simeq 0$. But if the mixture components are poorly separated, E_K has a large value.) leads to the normalized entropy criterion (Celeux and Soromenho 1996)

$$\mathrm{NEC}_K = \frac{E_K}{L_K - L_1} \tag{32}$$

as a criterion to be minimized for assessing the number of clusters arising from a mixture.

Note that NEC_1 is not defined. Biernacki *et al.* (1999) proposed the following efficient rule to deal with this problem. Let $K^\star$ be the value minimizing NEC_K, $(2 \leq K \leq K_{\mathrm{sup}})$, K_{sup} being an upper bound for the number of mixture components. We choose $K^\star$ clusters if $\mathrm{NEC}_{K^\star} \leq 1$, otherwise we declare no clustering structure in the data.

4.3.4 The cross-validation criterion (CV)

This criterion is valid only in the discriminant analysis (supervised) context. In this situation, note that only the model m has to be selected. Cross validation is a resampling method which can be summarised as follows: Let S be the whole dataset. Consider random splits of S into V independent datasets $S_1, \ldots, S_V$ of approximatively equal sizes $n_1, \ldots, n_V$. (If n/V is an integer h, we have $n_1 = \ldots = n_V = h$.) The CV criterion is defined by

$$\mathrm{CV}_m = \frac{1}{n} \sum_{v=1}^{V} \sum_{i \in S_v} \delta(\hat{\mathbf{z}}_i^{(v)}, \mathbf{z}_i) \tag{33}$$

with δ the 0-1 cost and $\hat{\mathbf{z}}_i^{(v)}$ denotes the group to which $\mathbf{x}_i$ is assigned when designing the assignment rule from the entire data set $(\mathbf{x}, \mathbf{z})$ without S_v. When $V = 1$ the cross validation is known as the *leave one out* method, and, in this case, fast estimation of the n discriminant rules is implemented in the Gaussian situation (Biernacki and Govaert 1999). In MIXMOD, the default value for the cross validation criterion is $V = 10$.

4.3.5 The double cross-validation criterion (DCV)

The CV error rate described above gives an optimistic estimate of the actual error rate because the method includes the selection of one model among several ones. Thus, there is a need to assess the actual error rate from an independent sample. This is the purpose of the DCV criterion, implemented in MIXMOD version 1.7.

The double cross-validated error rate is computed in MIXMOD as follows: Repeat the three following steps for v in $1, \ldots, V$ with $S_v^- = S \setminus S_v$

- build the models using the S_v^- dataset
- select the best model regarding the CV criterion: $m_v^\star$
- estimate the error rate e_v of $m_v^\star$ using S_v

$$e_v = \frac{1}{n_v} \sum_{i \in S_v} \delta(\hat{\mathbf{z}}_i^{m_v^\star}, \mathbf{z}_i). \tag{34}$$

The DCV error rate ($\bar{e}$) is finally obtained by averaging the $e_1, \ldots, e_V$. The empirical standard error of the error rate is given by σ_e

$$\bar{e} = \frac{1}{V} \sum_{v=1}^{V} e_v \quad , \quad \sigma_e = \left(\frac{1}{V-1} \sum_{v=1}^{V} (e_v - \bar{e})^2 \right)^{1/2} . \tag{35}$$

Recall that in MIXMOD, the default value of V is 10.

4.4 Partial labeling of individuals

MIXMOD allows partial labeling. Recall that in density estimation or clustering context, observed data are $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, the corresponding labels $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ being unknown. On the contrary, in the discriminant analysis context, all the labels $\mathbf{z}$ are available to estimate the mixture parameter θ. In some cases, the following intermediate situation may occur: the set $\mathbf{x}$ of individuals is divided into two sets $\mathbf{x} = (\mathbf{x}^\ell, \mathbf{x}^u)$ where $\mathbf{x}^\ell = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ $(1 \leq m \leq n)$ are units with known labels $\mathbf{z}^\ell = \{\mathbf{z}_1, ..., \mathbf{z}_m\}$, and $\mathbf{x}^u = \{\mathbf{x}_{m+1}, ..., \mathbf{x}_n\}$ units with unknown labels $\mathbf{z}^u = \{\mathbf{z}_{m+1}, ..., \mathbf{z}_n\}$.

The m.l. mixture parameter estimate is derived by maximizing the following log-likelihood

$$L(\theta|\mathbf{x}, \mathbf{z}^\ell) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \ln[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k)] + \sum_{i=m+1}^{n} \ln \left(\sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k) \right) . \tag{36}$$

In a clustering context using the classification approach, the c.m.l. method, is maximizing the following completed log-likelihood

$$CL(\theta, \mathbf{z}^u|\mathbf{x}, \mathbf{z}^\ell) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \ln[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k)] + \sum_{i=m+1}^{n} \sum_{k=1}^{K} z_{ik} \ln[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k)]. \tag{37}$$

In practice, the modifications of the algorithms are straightforward. It is simply necessary to replace t_{ik} by z_{ik} for all k and $i = 1, \ldots, m$ in the M step of EM, and to fix z_{ik} to constant known values for all k and $i = 1, \ldots, m$ in the M step of SEM and CEM.

4.5 Weighting the units

In some cases, it arises that some units are duplicated. Typically, it happens when the number of possible values for the units is low in regard to the sample size.

To avoid entering unnecessarily large lists of units, MIXMOD allows to specify a weight w_i for each unit $\mathbf{y}_i$ $(i = 1,\ldots,r)$. The set $\mathbf{y}^w = \{(\mathbf{y}_1, w_1),\ldots,(\mathbf{y}_r, w_r)\}$ is strictly equivalent to the set with eventual replications $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, so we have the relation $n = w_1 + \ldots + w_r$.

All formula are easily adapted to take account of this weighting scheme. For instance, the log-likelihood L becomes

$$L(\theta|\mathbf{x}) = L(\theta|\mathbf{y}^w) = \sum_{i=1}^{r} w_i \ln\left(\sum_{k=1}^{K} p_k h(\mathbf{y}_i|\boldsymbol{\lambda}_k)\right), \tag{38}$$

and the proportion estimation equation at the mth iteration becomes

$$p_k^m = \frac{\sum_{i=1}^{r} w_i t_{ik}^m}{n}. \tag{39}$$

5 The Gaussian mixture model

5.1 Definition

In the Gaussian mixture model, each $\mathbf{x}_i$ is assumed to arise independently from a mixture with density

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) \tag{40}$$

where p_k is the mixing proportion ($0 < p_k < 1$ for all $k = 1, ..., K$ and $p_1 + ... + p_K = 1$) of the kth component and $h(\cdot|\boldsymbol{\mu}_k, \Sigma_k)$ denotes the d-dimensional Gaussian density with mean $\boldsymbol{\mu}_k$ and variance matrix Σ_k,

$$h(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) = (2\pi)^{-d/2}|\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}, \tag{41}$$

and $\theta = (p_1,\ldots,p_K, \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K, \Sigma_1,\ldots,\Sigma_K)$ is the vector of the mixture parameters. Thus, clusters associated to the mixture components are ellipsoidal, centered at the means $\boldsymbol{\mu}_k$ and variance matrices Σ_k determine their geometric characteristics.

5.2 Fourteen Gaussian models

5.2.1 Eigenvalue decomposition of variance matrices

Following Banfield and Raftery (1993) and Celeux and Govaert (1995), we consider a parameterization of the variance matrices of the mixture components consisting of expressing the variance matrix Σ_k in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{42}$$

where $\lambda_k = |\Sigma_k|^{1/d}$, D_k is the matrix of eigenvectors of Σ_k and A_k is a diagonal matrix, such that $|A_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order. The parameter λ_k determines the *volume* of the kth cluster, D_k its *orientation* and A_k its *shape*. By allowing some but not all of these quantities to vary between clusters, we obtain parsimonious and easily interpreted models which are appropriate to describe various clustering situations.

5.2.2 The general family

First, we can allow the volumes, the shapes and the orientations of clusters to vary or to be equal between clusters. Variations on assumptions on the parameters λ_k, D_k and A_k $(1 \leq k \leq K)$ lead to 8 general models of interest. For instance, we can assume different volumes and keep the shapes and orientations equal by requiring that $A_k = A$ (A unknown) and $D_k = D$ (D unknown) for $k = 1, \ldots, K$. We denote this model $[\lambda_k DAD']$. With this convention, writing $[\lambda D_k A D_k']$ means that we consider the mixture model with equal volumes, equal shapes and different orientations.

5.2.3 The diagonal family

Another family of interest consists of assuming that the variance matrices Σ_k are diagonal. In the parameterization (42), it means that the orientation matrices D_k are permutation matrices. We write $\Sigma_k = \lambda_k B_k$ where B_k is a diagonal matrix with $|B_k| = 1$. This particular parameterization gives rise to 4 models: $[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$ and $[\lambda_k B_k]$.

5.2.4 The spherical family

The last family of models consists of assuming spherical shapes, namely $A_k = I$, I denoting the identity matrix. In such a case, two parsimonious models are in competition: $[\lambda I]$ and $[\lambda_k I]$.

Finally, we get 14 different models (see Table 1). Those 14 Gaussian mixture models are implemented, specifying different clustering situations from the eigenvalue decomposition of the variance matrices of the mixture components. The main advantage of variance matrices eigenvalue decomposition is the simple geometric interpretation of the models. To stress this point, Figure 1 shows a contour plot for each model, for $K = 2$ groups with dimension $d = 2$, consisting of a single ellipse of isodensity per group.

model	number of parameters	M step	inertia criteria
$[\lambda DAD']$	$\alpha + \beta$	CF	$\|W\|$
$[\lambda_k DAD']$	$\alpha + \beta + K - 1$	IP	-
$[\lambda DA_kD']$	$\alpha + \beta + (K-1)(d-1)$	IP	-
$[\lambda_k DA_kD']$	$\alpha + \beta + (K-1)d$	IP	-
$[\lambda D_kAD_k']$	$\alpha + K\beta - (K-1)d$	CF	$\|\Sigma_k \Omega_k\|$
$[\lambda_k D_kAD_k']$	$\alpha + K\beta - (K-1)(d-1)$	IP	-
$[\lambda D_kA_kD_k']$	$\alpha + K\beta - (K-1)$	CF	$\Sigma_k \|W_k\|^{\frac{1}{d}}$
$[\lambda_k D_kA_kD_k']$	$\alpha + K\beta$	CF	$\Sigma_k n_k \ln(\frac{\|W_k\|}{n_k})$
$[\lambda B]$	$\alpha + d$	CF	$\|\text{diag}(W)\|$
$[\lambda_k B]$	$\alpha + d + K - 1$	IP	-
$[\lambda B_k]$	$\alpha + Kd - K + 1$	CF	$\Sigma_k \|\text{diag}(W_k)\|^{\frac{1}{d}}$
$[\lambda_k B_k]$	$\alpha + Kd$	CF	$\Sigma_k n_k \ln(\frac{\|\text{diag}(W_k)\|}{n_k})$
$[\lambda I]$	$\alpha + 1$	CF	$\text{tr}(W)$
$[\lambda_k I]$	$\alpha + d$	CF	$\Sigma_k n_k \ln \text{tr}(\frac{W_k}{n_k})$

Table 1: Some characteristics of the 14 models. We have $\alpha = Kd + K - 1$ in the case of free proportions and $\alpha = Kd$ in the case of equal proportions, and $\beta = \frac{d(d+1)}{2}$; CF means that the M step is closed form, IP means that the M step needs an iterative procedure. The last column gives the inertia type criterion to be minimized in the case of equal proportions for each model. Exact definition of W and W_k are given in (47) and (48).

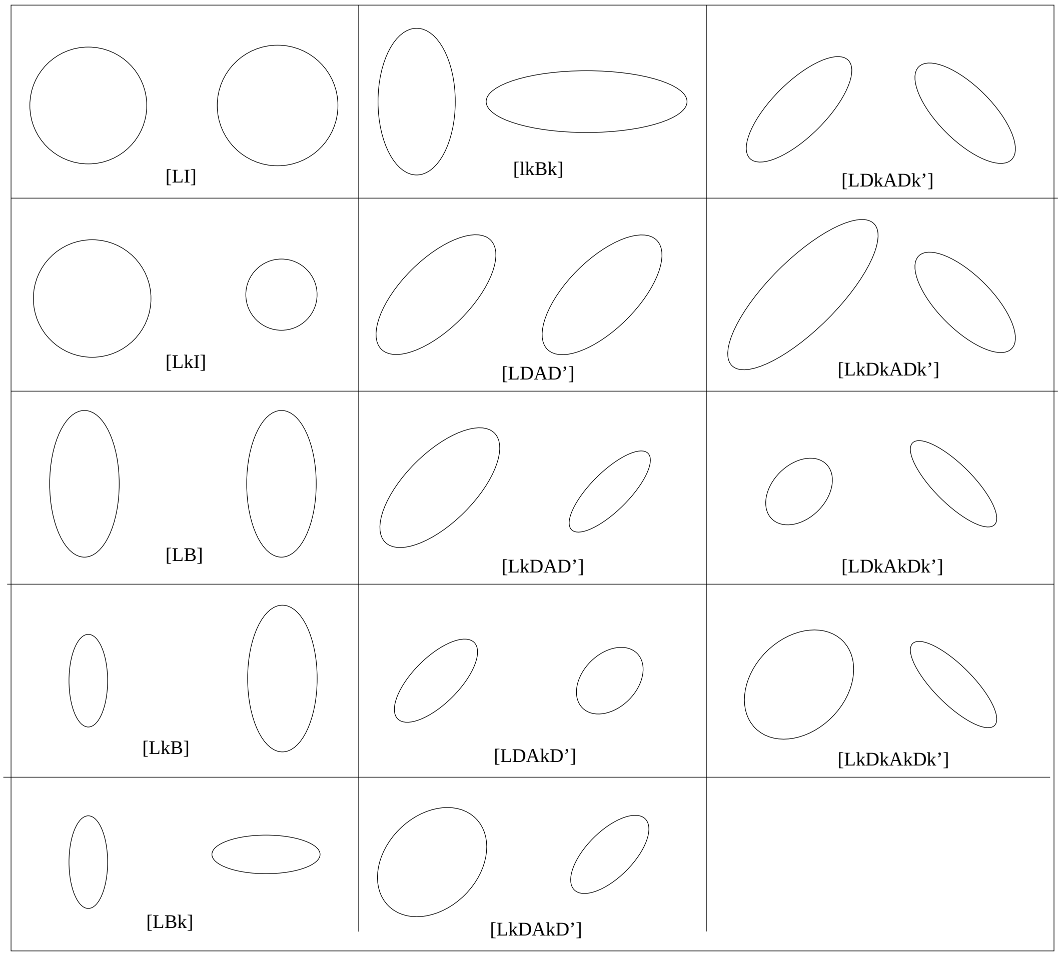

Figure 1: For two groups in two dimensions, this graphic displays the typical ellipse of isodensity per group for each of the 14 Gaussian models.

5.3 M step for each of the 14 models

The M step has to be detailed for each of the 14 models. It is obviously present in the EM algorithm and its variants (SEM, CEM), but it is also useful for the discrimination purpose since maximizing the likelihood with complete data $(\mathbf{x}, \mathbf{z})$ can be performed with a single iteration of the M step.

To unify the presentation, we make use of a classification matrix $\mathbf{c} = (c_{ik}, i = 1, \ldots, n; k = 1, \ldots, K)$ with $0 \leq c_{ik} \leq 1$ and $\sum_{k=1}^{K} c_{ik} = 1$, with the constraint $c_{ik} \in \{0, 1\}$ when $\mathbf{c}$ defines a partition as in the classification approach. With this convention, in both the mixture and the classification approaches, the M step consists of maximizing in θ the function:

$$F(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{c}) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \ln \left[p_k h(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)\right] \tag{43}$$

for fixed $\mathbf{c}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. When we are concerned with the EM algorithm, $\mathbf{c}$ defines a fuzzy classification and we have $c_{ik} = t_{ik}$ for $1 \leq i \leq n$ and $1 \leq k \leq K$. When we are concerned with the CEM algorithm, $\mathbf{c}$ defines a partition and we have $c_{ik} = 1$ if $\mathbf{x}_i$ belongs to the group k and 0 otherwise $(1 \leq i \leq n,\ 1 \leq k \leq K)$. Thus, for both approaches and for each of the considered models, the updating formulas for the proportions and the mean vectors of the mixture are, for $1 \leq k \leq K$,

$$\hat{p}_k = \frac{n_k}{n} \tag{44}$$

$$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k = \frac{\sum_{i=1}^{n} c_{ik} \mathbf{x}_i}{n_k} \tag{45}$$

where

$$n_k = \sum_{i=1}^{n} c_{ik}. \tag{46}$$

Remark that when $\mathbf{c}$ defines a partition $n_k = \text{card}(P_k)$. Moreover, we note W the within cluster scattering matrix

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \tag{47}$$

and W_k the scattering matrix of a cluster (or fuzzy cluster) , for $k = 1, \ldots, K$,

$$W_k = \sum_{i=1}^{n} c_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'. \tag{48}$$

The updating formulas for the variance matrices depend on the considered mixture model and are presented in the next subsections.

Table 1 summarizes some features of the 14 models. In this table, the first column specifies the model. The second column gives the number of parameters to be estimated. The third column indicates if the M step can be achieved with closed form formulas (CF) or if there is a need to make use of an iterative procedure (IP). The last column displays the inertia type criterion to be minimized for the case of equal proportions when the M step is closed form. These criteria can be derived from standard algebraic calculations. Some of them corresponds to standard criteria that was proposed without any reference to a statistical model. For instance, in clustering, $\mathrm{tr}(W)$ is the K-means criterion of Ward (1963), $|W|$ was suggested by Friedman and Rubin (1967) and $\Sigma_k n_k \ln \mathrm{tr}(\frac{W_k}{n_k})$ was proposed by Scott and Symons (1971). In discrimination, models $[\lambda C]$ and $[\lambda_k C_k]$ with equal proportions respectively correspond to classical linear and quadratic allocation rules (see for instance McLachlan 1982).

5.3.1 The general family

From Table 1, it can be seen that the inertia type criteria derived from the models $[\lambda DAD']$, $[\lambda D_k A_k D_k']$ and $[\lambda_k D_k A_k D_k']$ are classical clustering criteria (see Scott and Symons 1971, Maronna, Jacovkis 1974). On the contrary, the unusual models $[\lambda_k DAD']$, $[\lambda_k DA_k D']$ and $[\lambda_k D_k A D_k']$ which allow different volumes for the clusters do not lead to any inertia type criteria. Moreover, it is worth noting that the 8 models of the general family are invariant under any linear transformation of the data. We now detail the m.l. estimations of the variance matrices from a classification matrix **c** for the 8 situations.

Model $[\lambda DAD']$ In this well-known situation, the common variance matrix Σ is estimated by

$$\hat{\Sigma} = \frac{W}{n}. \tag{49}$$

Model $[\lambda_k DAD']$ In this situation, it is convenient to write $\Sigma_k = \lambda_k C$ with $C = DAD'$. M-step consists of two steps to minimize $\sum_{k=1}^{K} \mathrm{tr}(W_k C^{-1})/\lambda_k +$

$d\sum_{k=1}^{K} n_k \ln(\lambda_k)$

$$\text{Step 1 } (C \text{ fixed}): \qquad \lambda_k = \frac{\operatorname{tr}(W_k C^{-1})}{d n_k} \tag{50}$$

$$\text{Step 2 } (\lambda_k\text{'s fixed}): \qquad C = \frac{\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k}{|\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k|^{\frac{1}{d}}}. \tag{51}$$

Model $[\lambda D A_k D']$ In this situation and in the next one, there is no interest to assume that the terms of the diagonal matrices A_k are in decreasing order. Thus for the models $[\lambda D A_k D']$ and $[\lambda_k D A_k D']$ we do not assume that the diagonal terms of A_k are in decreasing order. First, direct calculation of λ is

$$\lambda = \frac{\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)}{nd}. \tag{52}$$

Then, M-step performs iteratively two steps to minimize $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$

$$\text{Step 1 } (D \text{ fixed}): \qquad A_k = \frac{\operatorname{diag}(D' W_k D)}{|\operatorname{diag}(D' W_k D)|^{\frac{1}{d}}} \tag{53}$$

$$\text{Step 2 } (A_k\text{'s fixed}): \qquad \text{see Flury and Gautschi (1986).} \tag{54}$$

Model $[\lambda_k D A_k D']$ In this situation, there is no need to isolate the volume and it is convenient to write $\Sigma_k = D A_k D'$ where $|A_k| = |\Sigma_k|$. M-step consists of two steps to minimize $\sum_{k=1}^{K} \left[\operatorname{tr}(D A_k^{-1} D' W_k) + n_k d \ln |A_k|\right]$

$$\text{Step 1 } (D \text{ fixed}): \qquad A_k = \operatorname{diag}(D' W_k D) \tag{55}$$

$$\text{Step 2 } (A_k\text{'s fixed}): \qquad \text{see Flury and Gautschi (1986).} \tag{56}$$

Model $[\lambda D_k A D_k']$ Considering for $k = 1, \ldots, K$ the eigenvalue decomposition $W_k = L_k \Omega_k L_k^{'}$ of the symmetric definite positive matrix W_k with the eigenvalues in the diagonal matrix Ω_k in decreasing order, we have

$$D_k = L_k, \quad A = \frac{\sum_{k=1}^{K} \Omega_k}{|\sum_{k=1}^{K} \Omega_k|^{\frac{1}{d}}} \quad , \quad \lambda = \frac{|\sum_{k=1}^{K} \Omega_k|^{\frac{1}{d}}}{n}. \tag{57}$$

Model $[\lambda_k D_k A D'_k]$ Using again the eigenvalue decomposition $W_k = L_k \Omega_k L'_k$, M-step consists of three steps to minimize $\sum_{k=1}^K \operatorname{tr}(W_k D_k A^{-1} D'_k)/\lambda_k + d \sum_{k=1}^K n_k \ln(\lambda_k)$

$$\begin{aligned} &\text{Step 1 } (D_k\text{'s, } A \text{ fixed}): & \lambda_k &= \frac{\operatorname{tr}(W_k D_k A^{-1} D_k^{'})}{d n_k} & (58)\\ &\text{Step 2 } (\lambda_k\text{'s, } A \text{ fixed}): & D_k &= L_k & (59)\\ &\text{Step 3 } (\lambda_k\text{'s, } D_k\text{'s fixed}): & A &= \frac{\sum_{k=1}^K \frac{1}{\lambda_k}\Omega_k}{|\sum_{k=1}^K \frac{1}{\lambda_k}\Omega_k|^{\frac{1}{d}}}. & (60)\end{aligned}$$

Model $[\lambda D_k A_k D'_k]$ In this situation, it is convenient to write $\Sigma_k = \lambda C_k$ where $C_k = D_k A_k D'_k$. Direct calculation shows that

$$C_k = \frac{W_k}{|W_k|^{\frac{1}{d}}}, \quad \lambda = \frac{\sum_{k=1}^K |W_k|^{\frac{1}{d}}}{n}. \tag{61}$$

Model $[\lambda_k D_k A_k D'_k]$ This is the most general situation and we have

$$\hat{\Sigma}_k = \frac{1}{n_k} W_k. \tag{62}$$

5.3.2 The diagonal family

For this more parsimonious family of models, the eigenvectors of Σ_k $(1 \leq k \leq K)$ are the vectors generating the basis associated to the d variables $(D_k = J_k)$. If the J_k are equal, the variables are independent. If the J_k are different, the variables are independent conditionally to the $\mathbf{z}_i$ $(1 \leq i \leq n)$. In this situation, Gaussian mixture with diagonal variance matrices can be viewed as an elegant model for weighting variables in a cluster analysis context. It leads to adaptive weighting algorithms assuming same weights for each cluster if the J_k's are assumed equal and different weights for each cluster if the J_k's are assumed different. We considered four models of interest. The main features of these four models are summarized in Table 1. The three inertia type criteria for the models $[\lambda B]$, $[\lambda B_k]$ and $[\lambda_k B_k]$ are simple adaptations of the corresponding criteria of the general family. The interesting model $[\lambda_k B]$ does not lead to an inertia type criterion. Moreover, it is worth noting that the 4 models of the diagonal family are invariant under any scaling of the variables but not under any linear transformation. We now derive the m.l.

estimation of the variance matrices from a classification matrix **c** for each of the four situations.

Model $[\lambda B]$ We have

$$B = \frac{\text{diag}(W)}{|\text{diag}(W)|^{\frac{1}{d}}}, \quad \lambda = \frac{|\text{diag}(W)|^{\frac{1}{d}}}{n}. \tag{63}$$

Model $[\lambda_k B]$ M-step consists of two steps to minimize $\sum_{k=1}^K \text{tr}(W_k B^{-1}) + d \sum_{k=1}^K n_k \ln(\lambda_k)$

$$\text{Step 1 } (B \text{ fixed}): \qquad \lambda_k = \frac{\text{tr}(W_k B^{-1})}{d n_k} \tag{64}$$

$$\text{Step 2 } (\lambda_k\text{'s fixed}): \qquad B = \frac{\text{diag}\left(\sum_{k=1}^K \frac{1}{\lambda_k} W_k\right)}{|\text{diag}\left(\sum_{k=1}^K \frac{1}{\lambda_k} W_k\right)|^{\frac{1}{d}}}. \tag{65}$$

Model $[\lambda B_k]$ We have

$$B_k = \frac{\text{diag}(W_k)}{|\text{diag}(W_k)|^{\frac{1}{d}}}\ , \quad \lambda = \frac{\sum_{k=1}^K |\text{diag}(W_k)|^{\frac{1}{d}}}{n}. \tag{66}$$

Model $[\lambda_k B_k]$ We get

$$B_k = \frac{\text{diag}(W_k)}{|\text{diag}(W_k)|^{\frac{1}{d}}}, \quad \lambda_k = \frac{|\text{diag}(W_k)|^{\frac{1}{d}}}{n_k}. \tag{67}$$

5.3.3 The spherical family

We consider here very parsimonious models for which the variance matrices are spherical. Two situations have to be considered: $\Sigma_k = \lambda I$ and $\Sigma_k = \lambda_k I$, I denoting the $(d \times d)$ identity matrix. The inertia type criterion $\text{tr}(W)$ of the model $[\lambda I]$ is certainly the oldest and the most employed clustering criterion. On the contrary, as far as we know, the criterion

$$\sum_{k=1}^K n_k \ln \frac{\text{tr}(W_k)}{n_k} \tag{68}$$

has been proposed for the first time by Banfield and Raftery (1993). Note that the 2 models of the spherical family are invariant under any isometric transformation. We derive the m.l. estimations of the volumes of the clusters for these models.

Model $[\lambda I]$ We get

$$\lambda = \frac{\mathrm{tr}(W)}{dn}. \tag{69}$$

Model $[\lambda_k I]$ We get

$$\lambda_k = \frac{\mathrm{tr}(W_k)}{dn_k}. \tag{70}$$

Formally models $[\lambda I]$ and $[\lambda_k I]$ do not seem to be very different and the increase of the number of parameters when considering model $[\lambda_k I]$ instead of model $[\lambda I]$ is small (see Table 1). In fact, these two models can lead to very different clustering structures.

5.4 Mixture of Factor Analyzers

In order to deal with high dimensional data, mixture of factor analyzers have been considered by several authors including Bouveyron *et al.* (2007), McNicholas and Murphy (2008), McLachlan and Peel (2000), Chapter 8, Tipping and Bishop (1999). In MIXMOD, a family of eight Gaussian mixture models introduced by Bouveyron *et al.* (2007) have been implemented for discriminant analysis in high dimensional spaces. They are denoted as HD (for High Dimensional) models in the following.

5.4.1 The general high dimensional model

The same eigenvalue decomposition of the mixture component variance matrices Σ_k, $\forall k = 1, ..., K$, is considered:

$$\Sigma_k = D_k \Delta_k D_k^t,$$

where D_k is the orthogonal matrix of the eigenvectors of Σ_k and Δ_k is a diagonal matrix containing the eigenvalues of Σ_k. It is further assume that

Δ_k has the following form (Note that in this section, there is no need to isolate the volume of the variance matrices.):

$$\Delta_k = \left(\begin{array}{cc} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{k\delta_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} \end{array} \right) \begin{array}{l} \left.\vphantom{\begin{matrix} a \\ \ddots \\ a \end{matrix}}\right\} \quad \delta_k \\ \left.\vphantom{\begin{matrix} b \\ \ddots \\ b \end{matrix}}\right\} \quad (d - \delta_k) \end{array}$$

where $a_{kj} \geq b_k$, for $j = 1, ..., \delta_k$ and $\delta_k < d$. The class-specific subspace generated by the δ_k first eigenvectors corresponding to the eigenvalues a_{kj} and containing the mean μ_k is denoted $\mathbb{E}_k$. In the orthogonal of $\mathbb{E}_k$, the component variance is characterized with a single parameter b_k. The projectors on $\mathbb{E}_k$ and $\mathbb{E}_k^{\perp}$ are denoted P_k and $P_k^{\perp}$. Figure 2 summarizes the model which is referred as $[a_{kj} b_k D_k \delta_k]$ in the following.

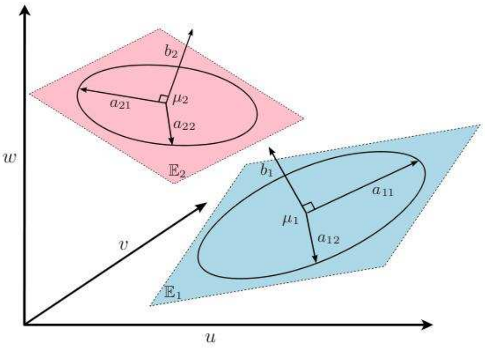

Figure 2: The parameters of the model $[a_{kj} b_k D_k \delta_k]$ in the case of two classes.

5.4.2 Sub-models of model $[a_{kj} b_k D_k \delta_k]$

Starting from the general model $[a_{kj} b_k D_k \delta_k]$ and allowing the elements of the model to vary or to be equal between classes, leads to 28 different models related to different types of regularization. In MIXMOD eight useful models have been selected:

- Two models with free dimension δ_k:
  - the model $[a_{kj}b_kD_k\delta_k]$
  - the model $[a_kb_kD_k\delta_k]$
- Six models with fixed dimension δ:
  - the model $[a_{kj}b_kD_k\delta]$
  - the model $[a_jb_kD_k\delta]$
  - the model $[a_{kj}bD_k\delta]$
  - the model $[a_jbD_k\delta]$
  - the model $[a_kb_kD_k\delta]$
  - the model $[a_kbD_k\delta]$

Their main features are summarized in Table 2. The second column of this table gives the number of parameters to be estimated. The third column provides the asymptotic order of the number of parameters to be estimated (with the assumption $K \ll \delta_k \ll d$). The last column gives this number in the particular case $K = 4$, $d = 100$ and $\forall k$, $\delta_k = 10$. These values are also given for the standard classification methods QDA and LDA. It is worthwhile to note that, in this cases, all HD models are more parsimonious than both QDA and LDA. Some particular situations lead to standard discriminant methods. For example, if $\delta_k = (d-1)$, for $k = 1, ..., K$, the model reduces to QDA. Moreover, if $a_{kj} = a_j$, $b_k = b$ and $D_k = D$, for $i = 1, ..., k$, it reduces to LDA.

5.4.3 The MAP step

The MAP decision rule for model $[a_{kj}b_kD_k\delta_k]$ yields to classify $\mathbf{x}$ in class C_{k^*} if $k^* = \text{argmin}_{k=1,\dots,K}\{\Gamma_k(\mathbf{x})\}$ with

$$\begin{aligned}\Gamma_k(\mathbf{x}) &= \|\mu_k - P_k(\mathbf{x})\|^2_{\mathcal{A}_k} + \frac{1}{b_k}\|\mathbf{x} - P_k(\mathbf{x})\|^2 \\ &+ \sum_{j=1}^{\delta_k} \log(a_{kj}) + (d-\delta_k)\log(b_k) - 2\log(\pi_k) + p\log(2\pi),\end{aligned}$$

Model	Number of	Asymptotic	Values of n for $K=4$,
	parameters n	order	$p=100$ and $d=10$
$[a_{kj}b_kD_k\delta_k]$	$\rho+\bar{\tau}+2K+D$	$Kd\delta$	4231
$[a_kb_kD_k\delta_k]$	$\rho+\bar{\tau}+3K$	$Kd\delta$	4195
$[a_{kj}b_kD_kd]$	$\rho+K(\tau+\delta+1)+1$	$Kd\delta$	4228
$[a_jb_kD_kd]$	$\rho+K(\tau+1)+\delta+1$	$Kd\delta$	4198
$[a_{kj}bD_kd]$	$\rho+K(\tau+\delta)+2$	$Kd\delta$	4225
$[a_jbD_kd]$	$\rho+K\tau+\delta+2$	$Kd\delta$	4195
$[a_kb_kD_kd]$	$\rho+K(\tau+2)+1$	$Kd\delta$	4192
$[a_kbD_kd]$	$\rho+K(\tau+1)+2$	$Kd\delta$	4189
QDA	$\rho+Kd(\delta+1)/2$	$Kp^2/2$	20603
LDA	$\rho+\delta(\delta+1)/2$	$p^2/2$	5453

Table 2: Features of the HD models: $\rho = Kd+K-1$ is the number of parameters required for the estimation of means and proportions, $\bar{\tau} = \sum_{k=1}^K \delta_k[p-(\delta_k+1)/2]$ and $\tau = \delta[d-(\delta+1)/2]$ are the number of parameters required for the estimation of $\tilde{D}_k$ and $\tilde{D}$, and $D = \sum_{k=1}^K \delta_k$. For asymptotic order, it is assumed that $K \ll \delta \ll d$.

$\|.\|_{\mathcal{A}_k}$ being a norm on $\mathbb{E}_k$ such that $\|\mathbf{x}\|^2_{\mathcal{A}_k} = \mathbf{x}^t\mathcal{A}_k\mathbf{x}$ with $\mathcal{A}_k = \tilde{D}_k\Delta_k^{-1}\tilde{D}_k^{\,t}$.

This decision rule is based on two distances: the distance between the observation and the subspace $\mathbb{E}_k$, and the distance between the projection of $\mathbf{x}$ on $\mathbb{E}_k$ and the mean of the class. It also depends on the variances a_{kj} and b_k and on prior probabilities π_k. Figure 3 depicts the decision rule. It illustrates the fact the projection on $\mathbb{E}_k^{\perp}$ is not required, reducing dramatically the number of parameters to be estimated and avoiding numerical difficulties.

5.4.4 Estimation of the model parameters

The parameters of the mixture of factor analyzers models are estimated through the maximum likelihood approach. Estimation of parameters π_k and μ_k of class C_k are

$$\hat{\pi}_k = \frac{n_k}{n},\ \hat{\mu}_k = \frac{1}{n_k}\sum_{\mathbf{x}_k\in C_k}\mathbf{x}_j.$$

In what follows, we make use of $W_k = \sum_{\mathbf{x}_j\in C_k}(\mathbf{x}_j-\hat{\mu}_k)^t(\mathbf{x}_j-\hat{\mu}_k)$, $n_k = \text{card}(C_k)$, $W = \sum_{i=k}^K \hat{\pi}_k W_k$, λ_{kj} which denotes the jth largest eigenvalue of

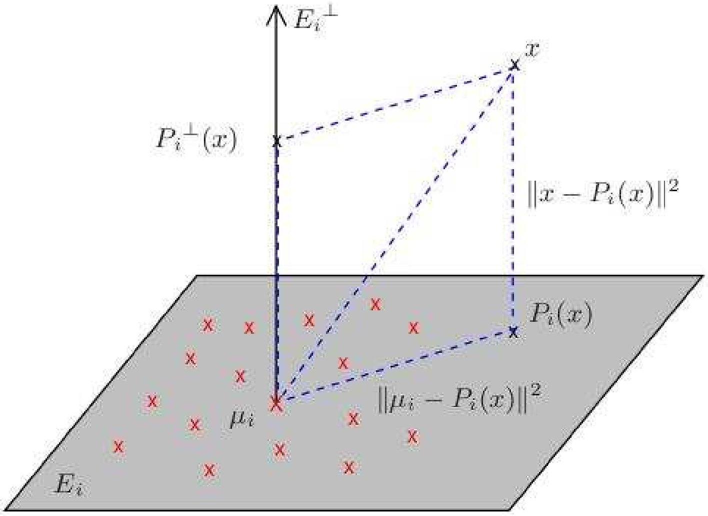

Figure 3: The subspaces $\mathbb{E}_k$ and $\mathbb{E}_k^\perp$ of the class C_k.

W_k and λ_j the jth largest eigenvalue of W, to define the m.l. estimate of the mixture component variance that are now presented. Details can be found in Bouveyron *et al.* (2007).

Models with free δ_k Assuming that dimensions δ_k are known for $k = 1, ..., K$, the following closed form estimators of model parameters are derived.

Subspace $\mathbb{E}_k$ The δ_k first columns of D_k are estimated by the eigenvectors associated with the δ_k largest eigenvalues λ_{kj} of W_k.

Model $[a_{kj} b_k D_k \delta_k]$ The estimators of a_{kj} are the δ_k largest eigenvalues λ_{kj} of W_k divided by n_k and

$$\hat{b}_k = \frac{1}{n_k(d-\delta_k)} \left(\text{trace}(W_k) - \sum_{j=1}^{\delta_k} \lambda_{kj} \right). \tag{71}$$

Model $[a_k b_k D_k \delta_k]$ The estimator of b_k is given by (71) and

$$\hat{a}_k = \frac{1}{n_k \delta_k} \sum_{j=1}^{\delta_k} \lambda_{kj}, \tag{72}$$

Models with common δ_k Assuming that parameter δ is known, we obtain the following closed form estimators for the parameters of the models with common δ_k, equal to δ.

Subspace $\mathbb{E}_k$ The δ first columns of D_k are estimated by the eigenvectors associated with the δ largest eigenvalues λ_{kj} of W_k.

Model $[a_{kj} b_k D_k d]$ The estimators of a_{kj} are the δ largest eigenvalues λ_{kj} of W_k divided by n_k and

$$\hat{b}_k = \frac{1}{n_k(d-\delta)} \left(\text{trace}(W_k) - \sum_{j=1}^{d} \lambda_{kj} \right). \tag{73}$$

Model $[a_j b_k D_k d]$ The estimator of b_k is given by (73) and

$$\hat{a}_j = \frac{1}{n} \sum_{k=1}^{K} \hat{\pi}_k \lambda_{kj}. \tag{74}$$

Model $[a_{kj} b D_k d]$ The estimators of a_{kj} are the δ largest eigenvalues λ_{kj} of W_k divided by n_k and

$$\hat{b} = \frac{1}{n(d-\delta)} \left(\text{trace}(W) - \sum_{k=1}^{K} \hat{\pi}_k \sum_{j=1}^{d} \lambda_{kj} \right). \tag{75}$$

Model $[a_j b D_k d]$ The estimators of a_j are given by (74) and the estimator of b is given by (75).

Model $[a_k b_k D_k d]$ The estimator of b_k is given by (73) and

$$\hat{a}_k = \frac{1}{nd} \sum_{j=1}^{d} \lambda_{kj}, \tag{76}$$

Model $[a_k b D_k d]$ The estimator of a_k is given by (76) and the estimator of b is given by (75).

Estimation of intrinsic dimensions The last parameters to be estimated are the intrinsic dimensions δ_k of the K classes. It is not possible to estimate the dimensions δ_k using the maximum likelihood approach and minimizing the cross validated error rate is considered. However, this minimization technique is not implemented in the present version of MIXMOD and the user has to provide all the intrinsic dimensions δ_k.

6 The multinomial mixture model

6.1 Definition

We consider now that data are n objects described by d categorical variables, with respective number of categories $m_1, \ldots, m_d$. The data can be represented by n binary vectors $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

$(i = 1, \ldots, n)$ where $x_i^{jh} = 1$ if the object i belongs to the category h of the variable j and 0 otherwise. Denoting $m = \sum_{j=1}^{d} m_j$ the total number of categories, the data are defined by the matrix $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with n rows and m columns. Binary data can be seen as a particular case of categorical data with d dichotomous variables, i.e. $m_j = 2$ for any $j = 1, \ldots, d$.

The latent class model assumes that the d ordinal variables are independent given the latent variable. Formulated in mixture terms (Everitt 1984), each $\mathbf{x}_i$ arises independently from a mixture of multivariate multinomial distributions defined by

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\alpha}_k) \tag{77}$$

where p_k is the mixing proportion $(0 < p_k < 1$ for all $k = 1, ..., K$ and $p_1 + ... + p_K = 1)$ of the kth component and where, for $k = 1, \ldots, K$,

$$h(\mathbf{x}_i|\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}} \tag{78}$$

with $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$. In (78), we recognize the product of d conditionally independent multinomial distributions of parameters $\boldsymbol{\alpha}_k^j$. The mixture parameters is denoted by $\theta = (p_1, \ldots, p_{K-1}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$.

This model may present problems of identifiability (see for instance Goodman 1974) but most situations of interest are identified.

6.2 Five multinomial models

In order to propose more parsimonious models than the previous one, we present the following extension of the parameterization of Bernoulli distributions used by Celeux and Govaert (1991) for clustering and also by Aitchison and Aitken (1976) for kernel discriminant analysis.

The basic idea is to impose the vector $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ to take the form $(\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j)$ with $\gamma_k^j > \beta_k^j$. Since $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$, we have $(m_j - 1)\beta_k^j + \gamma_k^j = 1$ and, consequently, $\beta_k^j = (1 - \gamma_k^j)/(m_j - 1)$. The constraint $\gamma_k^j > \beta_k^j$ becomes finally $\gamma_k^j > 1/m_j$. Then, the vector $\boldsymbol{\alpha}_k^j$ can be broken up into the two following parameters:

- $\mathbf{a}_k^j = (a_k^{j1}, \ldots, a_k^{jm_j})$ where $a_k^{jh} = 1$ if h corresponds to the rank of γ_k^j (in the following, this rank will be noted $h(k, j)$), 0 otherwise;

- $\varepsilon_k^j = 1 - \gamma_k^j$ which corresponds to the probability that the data $\mathbf{x}_i$ arising from the kth component are such that $x_i^{jh(k,j)} \neq 1$.

In other words, the multinomial distribution associated to the jth variable of the kth component is reparameterized by a center $\mathbf{a}_k^j$ and the dispersion ε_k^j around this center. Thus, it allows us to give an interpretation similar to the center and the variance matrix used for continuous data in the Gaussian mixture context.

Since, the relationship between the initial parameterization and the new one is given by:

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j & \text{if } h = h(k,j) \\ \varepsilon_k^j/(m_j - 1) & \text{otherwise,} \end{cases} \tag{79}$$

Equation (78) can be rewritten with $\mathbf{a}_k = (\mathbf{a}_k^j; j = 1, \ldots, d)$ and $\boldsymbol{\varepsilon}_k = (\varepsilon_k^j; j = 1, \ldots, d)$

$$h(\mathbf{x}_i|\boldsymbol{\alpha}_k) = \tilde{h}(\mathbf{x}_i|\mathbf{a}_k, \varepsilon_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} \left((1 - \varepsilon_k^j)^{a_k^{jh}} (\varepsilon_k^j/(m_j - 1))^{1 - a_k^{jh}} \right)^{x_i^{jh}}. \tag{80}$$

In the following, this model will be denoted by $[\varepsilon_k^j]$. In this context, three other models can be easily deduced. We note $[\varepsilon_k]$ the model where ε_k^j is independent of the variable j, $[\varepsilon^j]$ the model where ε_k^j is independent of the component k and, finally, $[\varepsilon]$ the model where ε_k^j is independent of both the variable j and the component k. In order to maintain some unity in the notation, we will denote also $[\varepsilon_k^{jh}]$ the most general model introduced at the previous section. The number of free parameters associated to each models is given in Table 3.

6.3 M step for each of the five models

The M step has to be detailed for each of the five models presented above. Using notation already defined in the Gaussian mixture context, the M step consists of maximizing in θ the function:

$$F(\theta|\mathbf{x}, \mathbf{c}) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \ln \left[p_k h(\mathbf{x}_i|\boldsymbol{\alpha}_k) \right] \tag{81}$$

model	number of parameters
$[\varepsilon]$	$\delta + 1$
$[\varepsilon^j]$	$\delta + d$
$[\varepsilon_k]$	$\delta + K$
$[\varepsilon_k^j]$	$\delta + Kd$
$[\varepsilon_k^{jh}]$	$\delta + K\sum_{j=1}^{d}(m_j - 1)$

Table 3: Number of free parameters of the five multinomial models. We have $\delta = K - 1$ in the case of free proportions and $\delta = 0$ in the case of equal proportions.

for fixed *classification* matrix $\mathbf{c}$ (obtained at the previous E or S or C steps) and with data matrix $\mathbf{x}$. We now detail the m.l. estimations of the parameters $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ of the multinomial distributions. In the following, we adopt the notation $e_k^{jh} = n_k - \sum_i c_{ik} x_i^{jh}$ and also $h(k,j)$ for the value of h which minimizes e_k^{jh}. In other terms, $h(k,j)$ still denotes the rank of the modality which occurs the most frequently for a given variable j and a given component k. For convenience, we use also $e_k^j = e_k^{jh(k,j)}$.

Model $[\varepsilon_k^{jh}]$

$$\alpha_k^{jh} = 1 - e_k^{jh}/n_k. \tag{82}$$

Model $[\varepsilon_k^j]$

$$\alpha_k^{jh} = \begin{cases} 1 - e_k^j/n_k & \text{if } h = h(k,j) \\ e_k^j/(n_k(m_j - 1)) & \text{otherwise.} \end{cases} \tag{83}$$

Model $[\varepsilon_k]$

$$\alpha_k^{jh} = \begin{cases} 1 - (\sum_j e_k^j)/(n_k d) & \text{if } h = h(k,j) \\ (\sum_j e_k^j)/(n_k d(m_j - 1)) & \text{otherwise.} \end{cases} \tag{84}$$

Model $[\varepsilon^j]$

$$\alpha_k^{jh} = \begin{cases} 1 - (\sum_k e_k^j)/n & \text{if } h = h(k,j) \\ (\sum_k e_k^j)/(n(m_j - 1)) & \text{otherwise.} \end{cases} \tag{85}$$

Model $[\varepsilon]$

$$\alpha_k^{jh} = \left\{ \begin{array}{ll} 1 - (\sum_{j,k} e_k^j)/(nd) & \text{if } h = h(k,j) \\ (\sum_{j,k} e_k^j)/(nd(m_j - 1)) & \text{otherwise.} \end{array} \right. \tag{86}$$

Using the new parameterization In fact, we could prefer to express the M step with the new parameterization $\mathbf{a}_k$ and $\boldsymbol{\varepsilon}_k$ (for models $[\varepsilon_k^j]$, $[\varepsilon_k]$, $[\varepsilon^j]$ and $[\varepsilon]$) instead of $\boldsymbol{\alpha}_k$, in particular for the meaningful interpretation of the terms $\mathbf{a}_k$ and $\boldsymbol{\varepsilon}_k$. In this case, it is easy to deduce expressions of $\mathbf{a}_k$ and $\boldsymbol{\varepsilon}_k$ from the expressions given above for $\boldsymbol{\alpha}_k$ with the following relationships:

$$a_k^{jh} = \left\{ \begin{array}{ll} 1 & \text{if } h = h(k,j) \\ 0 & \text{otherwise,} \end{array} \right. \tag{87}$$

and

$$\varepsilon_k^j = 1 - \alpha_k^{jh(k,j)}. \tag{88}$$

References

Aitchison, J. and Aitken, C. G. G. (1976), "Multivariate Binary Discrimination by the Kernel Method," *Biometrika*, *63*, 413-420.

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and non Gaussian Clustering," *Biometrics, 49*, 803-821.

Biernacki, C. Celeux, G. and Govaert, G. (1999), "An improvement of the NEC criterion for assessing the number of components arising from a mixture," *Pattern Recognition letters*, No 20, 267-272.

Biernacki, C. and Govaert, G. (1999). Choosing Models in Model-based Clustering and Discriminant Analysis. *Journal of Statistical Computation and Simulation, 64, 49-71.*

Biernacki, C. Celeux, G. and Govaert, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol *22*, No 7, 719-725.

Biernacki, C. Celeux, G. and Govaert, G. (2003) "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models". *Computational Statistics and Data Analysis*, *41*, 561-575.

Bouveyron C., Girard S. and Schmid C., High Dimensional Discriminant Analysis, *Communications in Statistics: Theory and Methods, 36*, 2607-2623.

Bozdogan, H. (1993), "Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix," in *Information and Classification*, O. Opitz, B. Lausen, and R. Klar (eds.), Heidelberg: Springer-Verlag, pp. 40-54.

Celeux, G. and Govaert, G. (1991), "Clustering Criteria for Discrete Data and Latent Class Models," *Journal of Classification*, *8*, 157-176.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, **14**, 315-332.

Celeux, G. and Govaert, G. (1995) "Parsimonious Gaussian models in cluster analysis". *Pattern Recognition, 28, 781-793.*

Celeux, G. and Soromenho, G. (1996) "An entropy criterion for assessing the number of clusters in a mixture model". *Journal of Classification, 13, 195-212.*

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statis. Soc. B*, **39**, 1-38.

Everitt, B. (1984). *An Introduction to Latent Variable Models.* London, Chapman and Hall.

Fraley, C. and Raftery, A. E. (1998): How Many Clusters ? Answers via Model-based Cluster Analysis. *The Computer Journal, 41*, 578-588. 215-231.

Flury, B. W., Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM J. Scientific Statist. Comput.*, **7**, 169-184.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *JASA,* **62**, 1159-1178.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, *61*,

Keribin, C. (2000). Consistent estimation of the order of mixture. *Sankhya*, *62*, 49-66.

Maronna, R. and Jacovkis, P. M. (1974). Multivariate procedures with variable metrics. *Biometrics*, **30**, 499-505.

McLachlan, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. in *Handbook of Statistics* (Vol. 2), P. R. Krishnaiah and L. N. Kanal (Eds.). Amsterdam: North-Holland, pp. 199-208.

McLachlan, G. J. and Peel D. (2000). *Finite Mixture Models.* New York, Wiley.

McNicholas, P.D. and Murphy, T.B. (2008) Parsimonious Gaussian Mixture Models. *Statistics and Computing*, to appear.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, *6*, 461-464.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387-397.

Tipping, M. E. and C. M. Bishop (1999). Mixtures of probabilistic principal component analysers. N*eural Computation 11*, 443-482.

Ward, J.H. (1963) Hierarchical groupings to optimize an objective function. *JASA*, **58**, 236-244.