



## EM for mixtures - Initialization requires special care

Jean-Patrick Baudry, Gilles Celeux

► **To cite this version:**

Jean-Patrick Baudry, Gilles Celeux. EM for mixtures - Initialization requires special care. 2015.  
<hal-01113242>

**HAL Id: hal-01113242**

**<https://hal.inria.fr/hal-01113242>**

Submitted on 4 Feb 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# EM for mixtures

## Initialization requires special care

Jean-Patrick Baudry · Gilles Celeux

Received: date / Accepted: date

**Abstract:** Maximum likelihood through the EM algorithm is widely used to estimate the parameters in hidden structure models such as Gaussian mixture models. But the EM algorithm has well-documented drawbacks: its solution could be highly dependent from its initial position and it may fail as a result of degeneracies. We stress the practical dangers of these limitations and how carefully they should be dealt with. Our main conclusion is that no method enables to address them satisfactorily in all situations. But improvements are introduced by, first, using a penalized loglikelihood of Gaussian mixture models in a Bayesian regularization perspective and, second, choosing the best among several relevant initialization strategies. In this perspective, we also propose new recursive initialization strategies which prove helpful. They are compared with standard initialization procedures through numerical experiments and their effects on model selection criteria are analyzed.

**Keywords:** Gaussian mixture models, EM algorithm, initialization strategies, recursive initialization, regularized likelihood, model selection criteria

### 1 Introduction

The EM algorithm (Dempster et al., 1977) is one of the most used algorithms in statistics (40 240 citations in Google Scholar, early February 2015). It is concerned with models with missing data structure and mixture models are

---

J.-P. Baudry  
LSTA, Université Pierre et Marie Curie  
Tel.: +33-144273353  
E-mail: Jean-Patrick.Baudry@upmc.fr

G. Celeux  
Inria Saclay-Île-de-France  
Tel.: +33-169155777  
E-mail: Gilles.Celeux@inria.fr

certainly the favorite domain of application of the EM algorithm (see McLachlan and Peel, 2000). Often the presence of missing data makes maximum likelihood (ml) inference difficult. The idea of the EM algorithm is to maximize at each iteration the conditional expectation of the complete data loglikelihood of the model at hand to derive the ml estimator of its parameter. The context is as follows: the observed data  $\mathbf{y} = (y_1, \dots, y_n)$  are assumed to arise from the probability distribution  $f(\cdot; \theta)$ ,  $\theta$  being the model parameter. The complete data are  $\mathbf{x} = (x_1, \dots, x_n)$  with  $(x_i = (y_i, z_i))$ , for  $i = 1, \dots, n$ , the  $z_i$ s being missing. The density of a complete observation  $x = (y, z)$  is  $g(x; \theta)$  and the two densities  $f$  and  $g$  are linked by the relation

$$f(y; \theta) = \int g(y, z; \theta) dz.$$

Denoting  $L_c(\theta) = \prod_i g(x_i; \theta)$  the complete likelihood, the two steps of the EM algorithm when  $\theta^r$  is the current parameter are

**E step** Compute  $Q(\theta, \theta^r) = E_{\theta^r}[\log L_c(\theta) | \mathbf{y}]$ .

**M step**  $\theta^{r+1} = \arg \max_{\theta} Q(\theta, \theta^r)$ .

The advantages and drawbacks of EM are well-documented (see for instance McLachlan and Krishnan, 2008). Its advantages are that (i) the likelihood  $L(\theta) = \prod_i f(y_i; \theta)$  is increasing at each iteration, (ii) the E and M steps lead generally to closed form and simple formulas easy to program. Its main drawbacks are that (i) EM may converge painfully slowly, (ii) its solutions may be highly dependent on its initial position  $\theta^0$ . Its advantages made it the most popular algorithm to estimate missing structure models and many extensions have been proposed to answer its drawbacks (see for instance Chapters 5 and 6 in McLachlan and Krishnan, 2008).

This article is focused on the possible influence of the EM algorithm initialization for the estimation and the selection of a mixture model. To be specific, we restrict our presentation to Gaussian mixture models. But all our considerations are expected to be relevant for all latent structure models for which the EM algorithm encounters many local maxima, slow convergence situations, or degeneracies.

In a Gaussian mixture model, the observed data  $(\mathbf{y} = y_1, \dots, y_n)$ ,  $y_i \in \mathbb{R}^d$  are assumed to arise from the density

$$f(y; \theta) = \sum_{k=1}^K p_k \phi(y; \eta_k),$$

where  $p_k \geq 0$  for  $k = 1, \dots, K$  with  $\sum_{k=1}^K p_k = 1$  are the mixing proportions, and where  $\Phi(\cdot; \mu, \Sigma)$  is the density of a Gaussian distribution with mean  $\mu$  and covariance matrix  $\Sigma$ . The vector parameter of the mixture model is denoted  $\theta = (p_1, \dots, p_{K-1}, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$ . The missing data are the labels  $z_i$  from which  $y_i$  arises for  $i = 1, \dots, n$ . These labels are binary indicator vectors in  $\{0, 1\}^K$ :  $z_{ik} = 1$  if  $y_i$  arises from the  $k$ th component, otherwise  $z_{ik} = 0$ . The

EM algorithm for this model consists of

**E step** For  $i = 1, \dots, n$ ,  $k = 1, \dots, K$  compute the current conditional probabilities that  $y_i$  arises from the Gaussian distribution  $\phi(y_i; \mu_k^r, \Sigma_k^r)$

$$\tau_{ik}^r = \frac{p_k^r \phi(y_i; \mu_k^r, \Sigma_k^r)}{\sum_{\ell=1}^K p_\ell^r \phi(y_i; \mu_\ell^r, \Sigma_\ell^r)}.$$

**M step**

$$\begin{aligned} p_k^{r+1} &= \frac{\sum_{i=1}^n \tau_{ik}^r}{n} \\ \mu_k^{r+1} &= \frac{\sum_{i=1}^n \tau_{ik}^r y_i}{\sum_{i=1}^n \tau_{ik}^r} \\ \Sigma_k^{r+1} &= \frac{\sum_{i=1}^n \tau_{ik}^r (y_i - \mu_k^{r+1})(y_i - \mu_k^{r+1})'}{\sum_{i=1}^n \tau_{ik}^r}. \end{aligned}$$

*Remark* We gave the EM formulas for the most general multivariate Gaussian mixture model with no restrictive assumptions on the mixing proportions or the component covariance matrices. More parsimonious models could be assumed, based on the eigenvalue decomposition of the component covariance matrices (see Banfield and Raftery, 1993). Detailed formulas for EM with fourteen different models are in Celeux and Govaert (1995). Note that for some covariance matrix decompositions the M step is no more closed form and requires a numerical optimization.

When using the EM algorithm, there is no guarantee that it provides the ml estimate of the model at hand since its solution depends on its initial position. Moreover, mixture ml estimation can be jeopardized with degeneracies. It can blur statistical inference for estimation and also for model selection. As a matter of fact, selecting a proper mixture model is an important and difficult problem. In particular, the choice of the number of components in a mixture model is quite influential. This choice depends on the purpose of the modeling, but it also depends on the quality of the assessed ml solutions. Actually, most model selection criteria and procedures are based on the likelihood of the consistent ml estimators. Using poor ml estimates at some place can lead to a misleading choice of mixture model.

All the traps of the EM algorithms (high dependence on starting values, slow convergence, convergence towards insensible or spurious maxima) arise more often for complex models. In particular, EM is jeopardized with mixture models with a great number of components or in high dimension settings or when many local maxima of the likelihood function are present. In such situations, this algorithm should be used with a great care to provide relevant and stable parameter estimates.

The aim of this article is to propose relevant procedures to get honest ml estimates derived with the EM algorithm. This is by the way a necessary condition to get a relevant assessing of the number of components in mixture models.

The article is organized as follows. Popular mixture model selection criteria are reviewed in Section 2. Section 3 is devoted to initialization strategies for EM to estimate a mixture model. First, standard strategies are reviewed. Then a recursive initialization procedure is proposed. The issue of degeneracies is considered in Section 5 and it is proposed to avoid them by replacing the likelihood with a regularized likelihood in a Bayesian perspective (Fraley and Raftery, 2007). Section 6 is devoted to numerical experiments comparing different initialization procedures and their effect on mixture model selection. The numerical experiments on both simulated and real data sets, highlight the performances of the compared initialization strategies to get sensible local maximizers of the penalized likelihood. A discussion section ends this article.

## 2 Mixture model selection criteria

This section is focused on the selection of the number of components in a mixture model. Thus, it is convenient to index the parameter  $\theta$  with the number of components of the mixture at hand and denote it  $\theta_K$ . We only sketch some popular model selection criteria and refer the reader to the Chapter 6 in McLachlan and Peel (2000) for an extensive presentation on assessing the order of a mixture model.

*BIC criterion* If the purpose of the mixture modeling is to get a semi-parametric density estimation, the BIC criterion (Schwarz, 1978) has been proved to be practically adequate for choosing  $K$  (Roeder and Wasserman, 1997; Fraley and Raftery, 2002). Actually, under some regularity conditions, it is proved to asymptotically select the model minimizing the Kullback-Leibler distance to the true distribution, so the true model when it is available (Keribin, 2000). This criterion has been derived within a Bayesian framework. It is an asymptotic approximation, as the number of observations goes to infinity, of the logarithm of the integrated likelihood

$$f_K(\mathbf{y}) = \int f(\mathbf{y}; \theta_K) \pi(\theta_K) d\theta_K,$$

$\pi$  being a prior distribution for  $\theta_K$ . It is

$$\text{BIC}(K) = \log L(\hat{\theta}_K) - \frac{D_K}{2} \log n, \quad (1)$$

$D_K$  being the dimension of  $\theta_K$ . Note that BIC is a pseudo-Bayesian criterion since it does not depend on the prior distribution  $\pi(\theta_K)$ .

*ICL criterion* If the purpose of the mixture modeling is model-based clustering the ICL criterion could be preferred to BIC (see Biernacki et al., 2000). ICL is an asymptotic BIC-like approximation of the logarithm of the complete integrated likelihood

$$g_K(\mathbf{y}, \mathbf{z}) = \int g(\mathbf{y}, \mathbf{z}; \theta_K) \pi(\theta_K) d\theta_K,$$

where the missing labels  $z_i$  are replaced with their most probable values  $\hat{z}_i$  for the ml estimate  $\hat{\theta}_K$ : for  $i = 1, \dots, n$ ,  $\hat{z}_{ik} = 1$  if  $k = \arg \max_{\ell} \hat{\tau}_{i\ell}$  and 0 otherwise. Here  $\hat{\tau}_{ik}$  indicates the conditional probability of  $y_i$  to arise from the  $k$ th component of the mixture with ml parameter  $\hat{\theta}_K$ . ICL is

$$\text{ICL}(K) = \text{BIC}(K) + \text{ENT}(K), \quad (2)$$

where

$$\text{ENT}(K) = - \sum_{i=1}^n \sum_{k=1}^K \hat{\tau}_{ik} \log \hat{\tau}_{ik}$$

is the entropy of the fuzzy clustering associated to the  $(\hat{\tau}_{ik})_{i=1, \dots, n; k=1, \dots, K}$ .

*The slope heuristic* It is important to recall that BIC and ICL are asymptotic criteria which behavior can be jeopardized when the number  $d$  of descriptors is not small with respect to the number  $n$  of observations. In such cases the slope heuristics of Birgé and Massart (2007), see also Baudry et al. (2011) can be a useful alternative. In the mixture context, it may be summarized as follows. Considering a criterion of the form

$$\text{SH}(K) = \log L(K) + \rho D_K, \quad (3)$$

it is assumed (i) that there exists a minimal  $\rho_{\min}$  such that any lighter penalty selects models with clearly too high complexities and such that heavier penalties select models with reasonable complexity; (ii)  $\rho_{\text{opt}} = 2\rho_{\min}$  provides us with an optimal penalty. In Baudry et al. (2011), it is proposed to directly estimate  $\rho_{\min}$  by the slope of  $\log L(K)$  with respect to  $D_K$  via a robust regression method. Obviously, this procedure requires that  $\log L(K)$  behaves linearly with respect to  $D_K$  at least for large enough values of  $K$ . Thus, an efficient use of the slope heuristics requires to get stable maximum likelihood values for a large amount of numbers of mixture components. SH is expected to minimize the risk (Birgé and Massart, 2007; Baudry et al., 2011).

In order to lead to an honest selection of  $K$ , all these criteria require reasonable evaluations of the consistent ml estimates of  $\theta_K$  for a large amount of values of  $K$ . As a consequence, solid initialization procedures for EM are needed.

### 3 Initializing the EM algorithm

The choice of  $\theta^0$  is decisive for the EM algorithm. Several strategies have been proposed to initiate EM for estimating mixture parameters and are available in mixture softwares. Initialization strategies can be distinguished by the importance they give to randomness. Some procedures available in the MIXMOD software (<http://www.mixmod.org>) use intensively random initializations and have been proposed in Biernacki et al. (2003) (see also Berchtold, 2004).

- Random** This basic procedure consists of running the EM algorithm until convergence from several random positions and to keep the solution providing the largest likelihood.
- Small EM** This procedure consists of using a large number of 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 after a few iterations (say five). Then EM is run from the parameter value providing the largest likelihood from these short runs of EM.
- SEM** The SEM algorithm is a stochastic algorithm incorporating between the E and M steps of EM a restoration of the unknown component labels  $z_i$ ,  $i = 1, \dots, n$  by drawing them at random from their current conditional distribution. The SEM algorithm is the same as the Monte Carlo EM algorithm with a single replication (see McLachlan and Krishnan, 2008, Chapter 6). This algorithm generates a Markov chain with a unique stationary distribution independent of its initial position. The SEM procedure consists of starting EM with the parameter value providing the largest likelihood from a long run of the SEM algorithm.
- CEM** The CEM algorithm, see for instance (Celeux and Govaert, 1992) is a classification algorithm incorporating a classification step between the E and M steps of EM. This classification step consists of assigning each observation  $y_i$  to the mixture component maximizing the current conditional probability  $(\tau_{ik})_{k \in \{1, \dots, K\}}$ . This algorithm converges in a small number of iterations. The CEM procedure consists of repeating a large number of times the CEM algorithm from random initial positions and to start EM with the parameter value providing the largest likelihood from these runs of the CEM algorithm.

Moreover, it is recommended to involve several starts from the procedures Small EM and CEM. Among these procedures, Small EM is often preferred. It is the default initialization procedure in the MIXMOD software. But none of them has shown to outperform the other procedures. Moreover, all of them can appear to be disappointing in a large dimension setting because the domain parameter to be explored becomes very large or when the number of mixture components is large.

At the opposite, deterministic initialization procedures such as the hierarchical procedure proposed in the software Mclust do not use random starting solutions (<http://www.stat.washington.edu/mclust>). Such hierarchical initialization can be outperformed by the above mentioned random procedures. But, they provide more stable results which can be viewed as an advantage in a high dimensional setting or for large numbers of mixture components.

Herafter, we propose recursive initialisation procedures which can be expected to take profit of the advantages of both worlds.

### 3.1 Recursive Initialization

A problem which often occurs in a mixture analysis with different numbers of components is that some solutions may be suboptimal or are spurious. This problem is common with the random initialization procedures of Section 3. It could prevent from selecting a sensible number of mixture components with the criteria presented in Section 2. The recursive procedures we now present aim to avoid irrelevant parameter estimates.

#### 3.1.1 Recursive algorithms

Assume that the user aims to choose a mixture component with a number of components belonging to  $\{K_{\min}, \dots, K_{\max}\}$ . Recursive initialization consists of splitting *at random* one of the  $K$  components into two components to get a  $(K + 1)$  solution.

- First the  $K_{\min}$  solution is thoroughly designed using for instance the Small EM procedure repeated a large number of times.
- From  $K = K_{\min}$ , the initial position of the  $(K + 1)$ -component mixture is obtained by splitting one of the components of the  $K$ -component mixture into two components.

This strategy can be related to Pelleg and Moore (2000), who introduce the *X-means* algorithm as an extension to the *k-means*. It enables to choose the number of clusters and to partially remedy to the initialization problem for *k-means*.

Three different strategies to choose the mixture component to be split will be considered.

- *Random Choice*: Choose at random the mixture component to be split (Papastamoulis et al., 2014).
- *Optimal Sequential Choice*: Choose the mixture component to be split by optimizing a splitting criterion.
- *Complete Choice*: Try to split all the  $K$  components and choose to split the component leading to the largest likelihood (Baudry, 2009).

Other splitting strategies such as the strategy in Fraley et al. (2005) especially conceived to deal with large data sets are possible.

#### 3.1.2 Splitting criteria

Possible criteria to choose the component to be split are now presented. They are derived from decompositions of the model likelihood  $L(\theta_K)$

$$\log L(\theta_K) = \sum_{i=1}^n \log \sum_{k=1}^K p_k \phi(y_i; \mu_k, \Sigma_k).$$



*Splitting the component with the weakest contribution to the likelihood* Since

$$\log L(\theta_K) = \sum_{\ell=1}^K \sum_{i=1}^n \tau_{i\ell} \log \sum_{k=1}^K p_k \phi(y_i; \mu_k, \Sigma_k),$$

it leads to split the component

$$\tilde{k}^L = \arg \min_{1 \leq \ell \leq K} \sum_{i=1}^n \tau_{i\ell} \log \sum_{k=1}^K p_k \phi(y_i; \mu_k, \Sigma_k). \quad (4)$$

An alternative criterion is the normalized contribution to the likelihood of a component

$$\tilde{k}^{Ln} = \arg \min_{1 \leq \ell \leq K} \frac{\sum_{i=1}^n \tau_{i\ell} \log \sum_{k=1}^K p_k \phi(y_i; \mu_k, \Sigma_k)}{\sum_{i=1}^n \tau_{i\ell}}. \quad (5)$$

*Splitting the component with the weakest contribution to the complete likelihood*  
We have

$$\log L(\theta_K) = \sum_{\ell=1}^K \sum_{i=1}^n \tau_{i\ell} \log p_\ell \phi(y_i; \eta_\ell) - \sum_{\ell=1}^K \sum_{i=1}^n \tau_{i\ell} \log \tau_{i\ell},$$

splitting the component with the weakest contribution to the complete likelihood leads to choose the component

$$\tilde{k}^{Lc} = \arg \min_{1 \leq \ell \leq K} \sum_{i=1}^n \tau_{i\ell} \log p_\ell \phi(y_i; \eta_\ell). \quad (6)$$

An alternative criterion is the normalized contribution to the complete likelihood of a component

$$\tilde{k}^{Lcn} = \arg \min_{1 \leq \ell \leq K} \frac{\sum_{i=1}^n \tau_{i\ell} \log p_\ell \phi(y_i; \eta_\ell)}{\sum_{i=1}^n \tau_{i\ell}}. \quad (7)$$

*Splitting the component with the largest contribution to the mixture entropy*  
It leads to split the component

$$\tilde{k}^E = \arg \max_{1 \leq \ell \leq K} - \sum_{i=1}^n \tau_{i\ell} \log \tau_{i\ell}. \quad (8)$$

If the normalized entropy is considered instead, it leads to split the component

$$\tilde{k}^{En} = \arg \max_{1 \leq \ell \leq K} - \frac{\sum_{i=1}^n \tau_{i\ell} \log \tau_{i\ell}}{\sum_{i=1}^n \tau_{i\ell}}. \quad (9)$$

#### 4 Avoiding degeneracies of the EM algorithm

Deriving the maximum likelihood parameter estimate of a Gaussian mixture faces an important difficulty since the likelihood function of Gaussian mixtures with unrestricted component covariance matrices is unbounded. Thus, the EM algorithm is jeopardized by degenerated solutions. The EM algorithm can fail with some starting values, models, and numbers of components (see for instance McLachlan and Peel, 2000 Section 3.10). To get rid of this issue in our numerical experiments, we opt for a Bayesian regularization, as in Fraley and Raftery (2007) or Ciuperca et al. (2003), and replace the ml estimate with the maximum a posteriori (MAP) estimate which maximizes the regularized loglikelihood  $\log L(\theta_K) + \log \pi(\theta_K)$ ,  $\pi$  being a prior distribution on the vector parameter  $\theta_K$ .

Since we make use of Bayesian inference merely to avoid degeneracies in the estimation of the component covariance matrices, we put no prior distribution on the mixing proportions and the component mean vectors. Following Fraley and Raftery (2007), we use an inverse Wishart( $\nu, \Lambda$ ) exchangeable conjugate prior distribution for the component covariance matrices  $\Sigma_k, k = 1, \dots, K$ . We choose the following prior hyperparameters:  $\nu = d + 2$  (as Fraley and Raftery, 2007) and  $\Lambda$  depends on the model. Then, the MAP estimate of  $\theta_K$  is derived with the EM algorithm.

For the general model,  $\Lambda = \frac{\sigma_0^{1/d} S}{|S|^{1/d}}$ ,  $S$  being the empirical covariance matrix of the data  $\mathbf{y}$  and  $\sigma_0$  a small positive number. We have  $|\Lambda| = \sigma_0$ . The greater  $\sigma_0$  the greater the regularization. Thus, this tuning parameter allows to control the regularization. It allows a weaker regularization than the hyperparameter  $\Lambda = \frac{S}{K^{1/d}}$ , proposed in Fraley and Raftery (2007). The formulas of the EM algorithm remain the same, except for the updating of the covariance matrices in the M step which becomes (at the  $(r + 1)$ th iteration) (see Fraley and Raftery, 2007 for details)

$$\Sigma_k^{r+1} = \frac{\Lambda + \sum_{i=1}^n \tau_{ik}^r (y_i - \mu_k^{r+1})(y_i - \mu_k^{r+1})'}{\nu + \sum_{i=1}^n \tau_{ik}^r + d + 2}.$$

We also consider Gaussian mixture models with diagonal component covariance matrices  $B_k$ , with  $B_k = \text{diag}(B_{kj}, j = 1, \dots, d)$  for  $k = 1, \dots, K$ . In this case, we use an inverseGamma( $\frac{\nu}{2}, \frac{\zeta_j}{2}$ ) distribution for  $B_{kj}$  with the prior hyperparameters  $\nu = d + 2$  and  $\zeta_j = (\sigma_0)^{1/d} \frac{s_j}{(s_1 \dots s_d)^{1/d}}$ ,  $s_j$  being the empirical variance of the variable  $j$  and  $\sigma_0$  a small positive number. The updating of the  $B_{kj}$ s at the  $(r + 1)$  iteration of the EM algorithm is

$$B_{kj}^{r+1} = \frac{\zeta_j + \sum_{i=1}^n \tau_{ik}^r (y_{ij} - \mu_{kj}^{r+1})^2}{\nu + \sum_{i=1}^n \tau_{ik}^r + 2}.$$

We also consider Gaussian mixture models with spherical covariance matrices  $\lambda_k I (1 \leq k \leq K)$  proportional to the identity matrix. In this case, we use an inverseGamma( $\frac{\nu}{2}, \frac{\zeta}{2}$ ) distribution for  $\lambda_k$  with the prior hyperparameters

$\nu = d + 2$  and  $\zeta = 2(\sigma_0)^{1/d}$ ,  $\sigma_0$  being a small positive number. The updating of the  $\lambda_k$ s at the  $(r + 1)$  iteration of the EM algorithm is

$$\lambda_k^{r+1} = \frac{\zeta + \sum_{i=1}^n \tau_{ik}^r (y_i - \mu_k^{r+1})' (y_i - \mu_k^{r+1})}{\nu + d \sum_{i=1}^n \tau_{ik}^r + d + 2}.$$

The choice of  $\sigma_0$  is obviously very important. A first idea is to make sure the chosen  $\sigma_0$  does not hide the data structure. With simulated data, it is possible to monitor the evolution of a regularized EM starting from the true parameter and to exclude values of  $\sigma_0$  such that the estimated parameter get too far away. For real data sets a non regularized ml parameter estimation can replace the true parameter...

## 5 Numerical experiments

The only way to assess the ability of EM initialization strategies to properly derive the ml estimate of a mixture model is to proceed to numerical experiments in practical situations. In this section, we present the results of numerical experiments on three different kinds of data sets. A six-component Gaussian mixture in  $\mathbb{R}^2$  is analyzed, first with “true” models, of which some contain the true distribution, and second with “false” models, of which none contains the true distribution. Then a simulated data set from a Gaussian mixture with a large number of components in  $\mathbb{R}^3$  is considered. Finally, numerical experiments are achieved on a real data set with a large number of observations to be clustered.

The initialization strategies considered are:

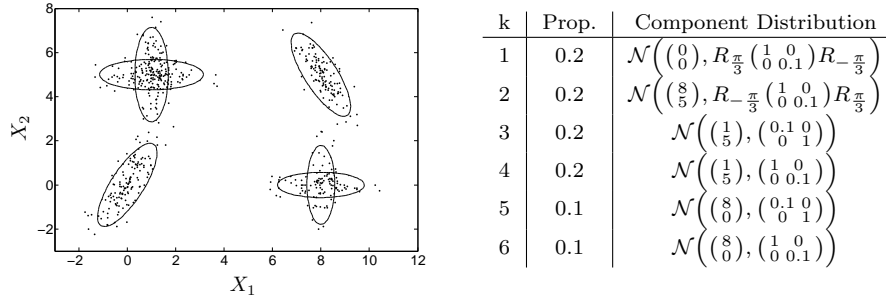
- Random. 13 runs of EM from random initializations.
- Small EM. Number of iterations of the small runs of EM: 5; number of small runs of EM: 50; repeat 10 times and keep the best.
- SEM. Number of iterations in the SEM runs: 500; number of SEM runs: 8.
- Mclust. Hierarchical initialization.
- KM1 Complete, L, Ln, etc. are KM1 strategies starting from the best  $K$ -component solution among those obtained from the initialization strategies Random, Small EM, SEM, Mclust and KM1 (for  $K > K_{\min}$ ).
- KM1 sComplete ( $s$  stands for *sequential*) is a KM1 strategy independent of the previous strategies (except for  $K = K_{\min}$ ).

The number of iterations of EM after the initialization step is the same for all: 1000. Moreover we consider the strategy “best without KM1” which consists of choosing the best solution after the final EM run, among those obtained from the strategies Random, Small EM, SEM, and Mclust.

For all simulated numerical experiments (Sections 5.1 and 5.2), 30 data sets were simulated.

### 5.1 True and wrong models

We consider 600 observations from a six-component Gaussian mixture in  $\mathbb{R}^2$  where two components have a non diagonal covariance matrix. The mixture characteristics are given in Figure 1. At first, general Gaussian mixture models with free variance matrices were fitted to this data set. Secondly, the fitted Gaussian mixture models were restricted to diagonal covariance matrices (the family denoted by  $[p_k \lambda_k B_k]$  in Celeux and Govaert, 1995).

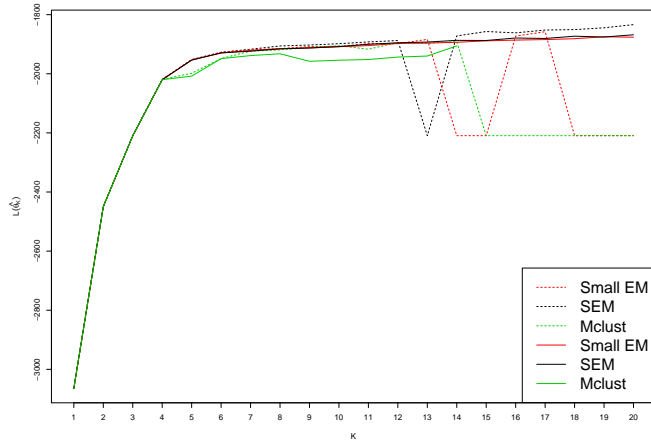


**Fig. 1** *True and Wrong Models* data set. 600 obs. from a six-component Gaussian mixture in  $\mathbb{R}^2$  (in the family denoted by  $[p_k \lambda_k C_k]$  in Celeux and Govaert, 1995).  $R_\theta$  denotes the  $\theta$  rotation matrix.

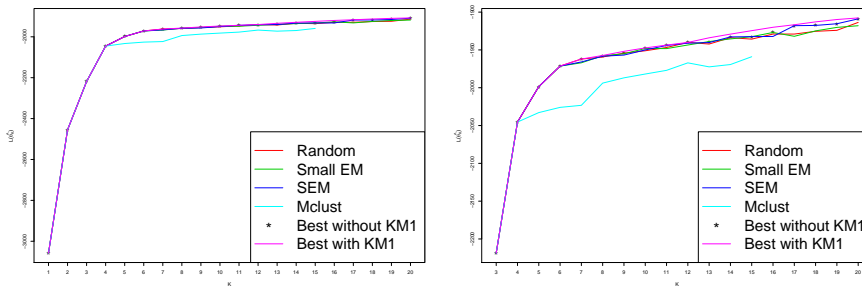
*True models* Despite the fact that it is a favorable situation for the Gaussian mixture model, the EM algorithms could be trapped when trying to maximize the non regularized likelihood as it is apparent from Figure 2. All the strategies encountered problems with the non regularized likelihood. Note that, by convention, when the EM algorithm is trapped, we set likelihood to the value obtained with  $K = 3$  components. We do so to highlight these situations in the graph. In the following we use only regularized EM algorithms.

Note that according to Figure 3 (i) the MClust strategy becomes disappointing here for  $K > 4$ , (ii) the SEM strategy seems to outperform the other random strategies for large numbers of components and (iii) the KM1 Complete strategy provides the highest maximum for all  $K$  values.

Figure 4 shows that not surprisingly BIC provides most often the true number of mixture components except for the Mclust strategy. SH has a tendency here to overestimate the number of components. For the strategies which do not involve KM1, SH can select surprisingly high numbers of components. It is noteworthy that ICL prefers almost always the clustering solution in four clusters visible in Figure 1. This experiment illustrates clearly the fact that ICL is not aiming to recover the true number of components but to select a stable clustering solution.

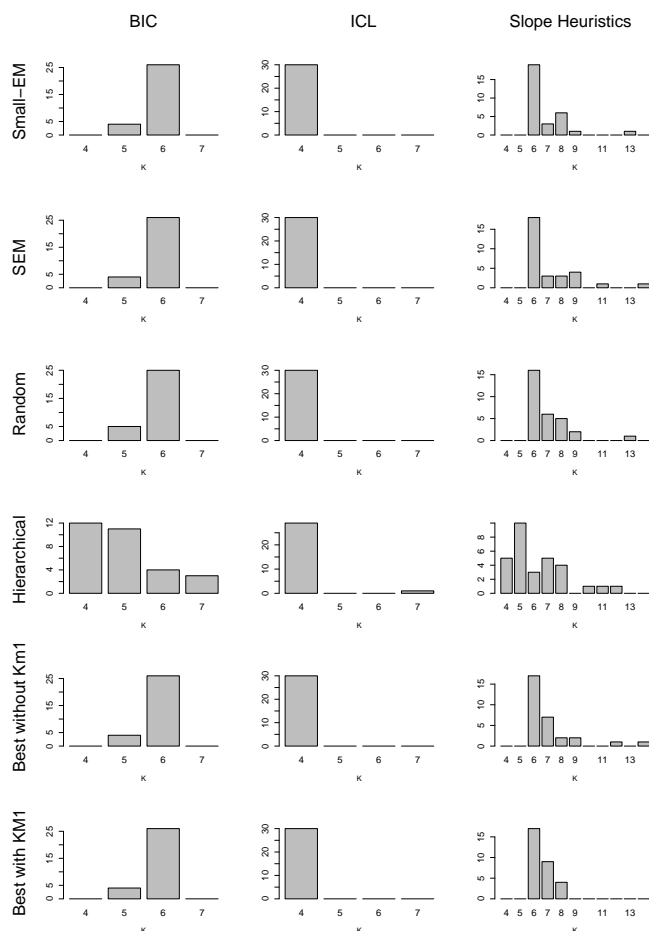


**Fig. 2** *True Model* experiment. Graphs of the maximized likelihood for different strategies, including non regularized strategies (dashed), as a function of the number of components with the “true” mixture models  $[p_k \lambda_k C_k]$



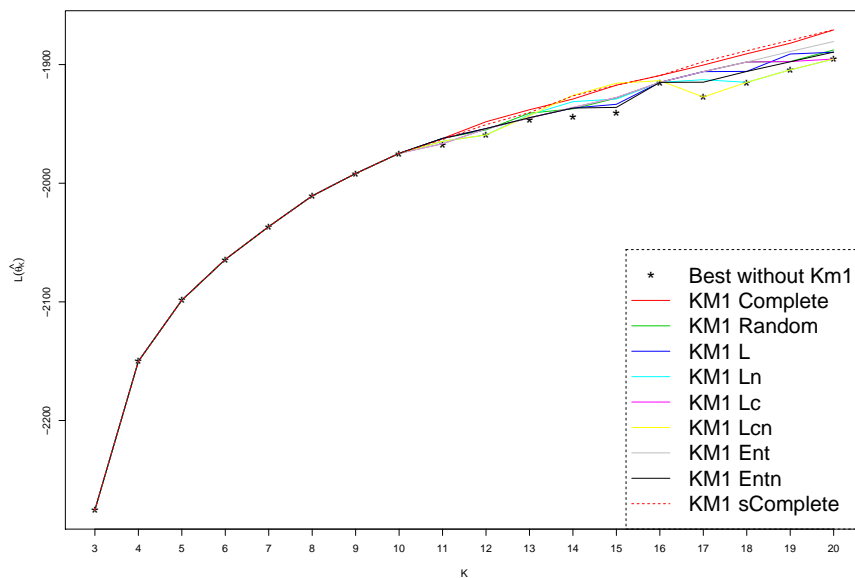
**Fig. 3** *True Model* experiment. Graphs of the maximized likelihood for different strategies as a function of the number of components with the “true” mixture model  $[p_k \lambda_k C_k]$ ; on the right figure, a zoom view

*Wrong models* First, for this more sensitive situation, we display the behavior of the KM1 strategies presented in Section 3.1.2 to get a sensible maximizer. It is apparent from Figure 5, that all the considered splitting criteria behave almost the same and are outperformed by the strategies KM1 Complete and KM1 sComplete. This behavior always occurs in all the situations of our numerical experiments. For simplicity, these splitting criteria will be omitted in the following. In practice, both “complete” KM1 strategies have analogous performances, but sometimes the KM1 sComplete strategy is outperformed by the KM1 Complete strategy. Since this strategy is also more natural (there is no reason not to choose the best solution among the  $K$  components), we only show the KM1 Complete strategy.

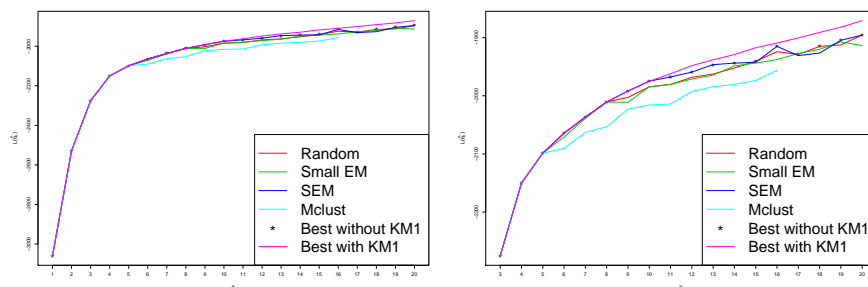


**Fig. 4** *True Model* experiment. Frequencies of choices of the number of mixture components with the criteria SH, BIC, and ICL for the true mixture model  $[p_k \lambda_k C_k]$

Figure 6 confirms the superiority of the KM1 strategy and the good behavior of the SEM strategy. It is also important to remark that all strategies but KM1 can lead to a non increasing likelihood as a function of  $K$ . Figure 7 illustrates the ability of ICL to select a sensible number of clusters despite the fact that the family of models is not quite relevant. Not surprisingly, in this misspecified setting, BIC and SH have a tendency to overestimate the number of components (see Biernacki et al., 2000).



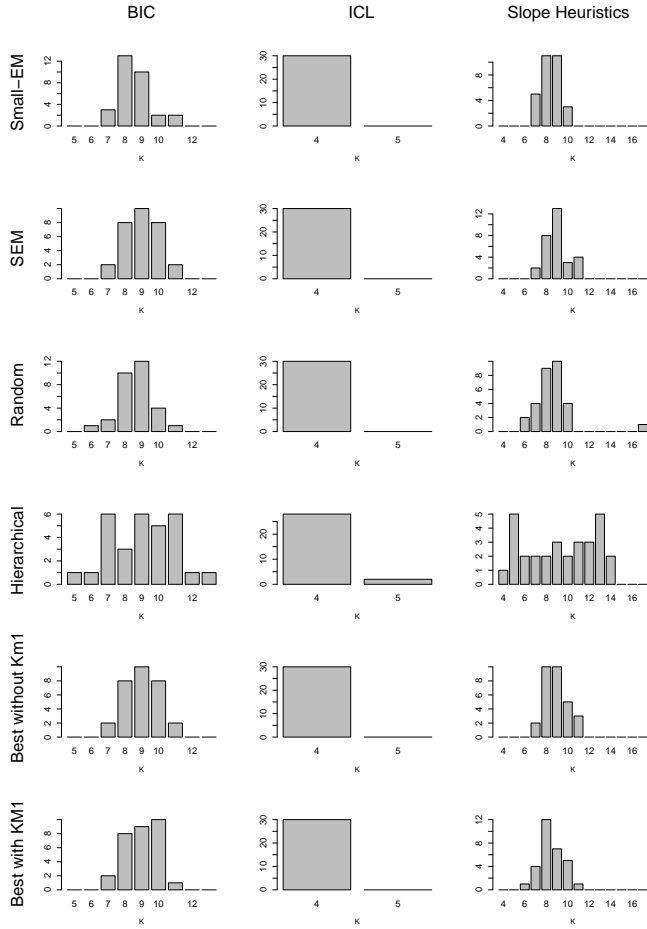
**Fig. 5** *Wrong Model* experiment. Behavior of the maximized likelihood for the different KM1 strategies as a function of the number of components with the wrong diagonal mixture model  $[p_k \lambda_k B_k]$



**Fig. 6** *Wrong Model* experiment. Behavior of the maximized likelihood for different strategies as a function of the number of components with the wrong diagonal mixture model  $[p_k \lambda_k B_k]$ ; on the right figure, a zoom view

## 5.2 Bubbles data set

This simulated data set, plotted in Figure 8, is composed of 1000 observations in  $\mathbb{R}^3$ . It consists of an equiprobable mixture of three large *bubble* groups centered at  $\nu_1 = (0, 0, 0)$ ,  $\nu_2 = (6, 0, 0)$  and  $\nu_3 = (0, 6, 0)$  respectively. Each bubble group  $j$  is simulated from a mixture of seven components according to



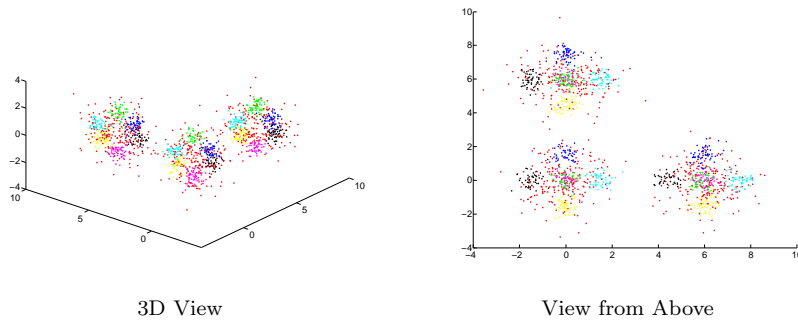
**Fig. 7** *Wrong Model* experiment. Frequencies of choices of the number of mixture components with the criteria SH, BIC, and ICL for the wrong diagonal mixture model  $[p_k \lambda_k B_k]$

the following density distribution:

$$y \in \mathbb{R}^3 \mapsto 0.4\Phi(y; \mu_1 + \nu_j, I_3) + \sum_{k=2}^7 0.1\Phi(y; \mu_k + \nu_j, 0.1I_3)$$

with  $\mu_1 = (0, 0, 0)$ ,  $\mu_2 = (0, 0, 1.5)$ ,  $\mu_3 = (0, 1.5, 0)$ ,  $\mu_4 = (1.5, 0, 0)$ ,  $\mu_5 = (0, 0, -1.5)$ ,  $\mu_6 = (0, -1.5, 0)$  and  $\mu_7 = (-1.5, 0, 0)$ . Thus the distribution of this data set is actually a 21-component Gaussian mixture. The reader is referred to Baudry (2009), Chapter 5 for more details. This is a challenging data set because of the high number of components which overlap and have small sizes. A model collection  $(S_m)_{1 \leq m \leq M_{\max}}$  of spherical Gaussian mixtures

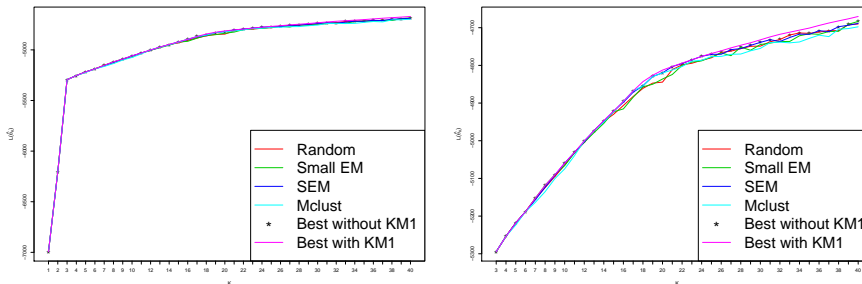




**Fig. 8** *Bubbles* Data Set. 1000 obs. from a 21-component spherical Gaussian mixture in  $\mathbb{R}^3$

with covariance matrices  $\Sigma_k = \lambda_k I_3$  with  $\lambda_k \in \mathbb{R}_+^*$  is fitted to the *Bubbles* data set.

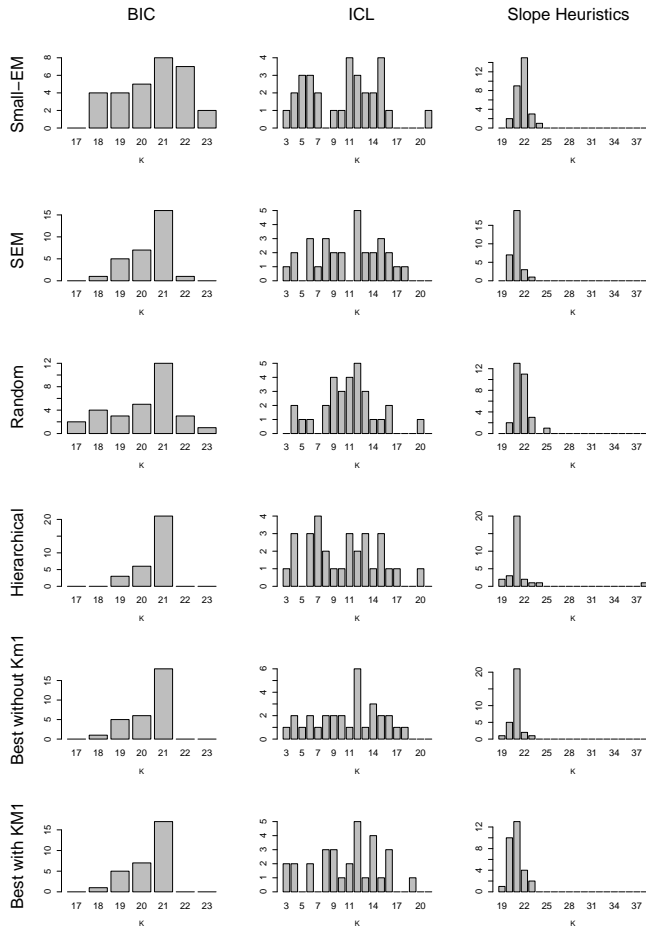
From Figures 9 and 10, again KM1 appears to be the best strategy and the Best without KM1 solution produces a non increasing loglikelihood in function of  $K$ . The SEM strategies behaves well. But, on the contrary to Mclust, the strategies Small EM and Random are amazingly suboptimal around  $K = 21$ , the true number of components. And actually, there is no sensitive difference between BIC and SH which have difficulties to recover the true number of clusters with these two strategies. Otherwise ICL does not select a stable number of clusters.



**Fig. 9** *Bubbles* experiment. Behavior of the maximized likelihood for different strategies as a function of the number of components with the spherical mixture model  $[p_k \lambda_k I]$  for the *Bubbles* data set; on the right figure, a zoom view.

### 5.3 Dynamic expression of the transcriptome in embryonic flies

As part of the modENCODE project, which aims to provide functional annotation of the *Drosophila melanogaster* genome, Graveley et al. (2011) charac-

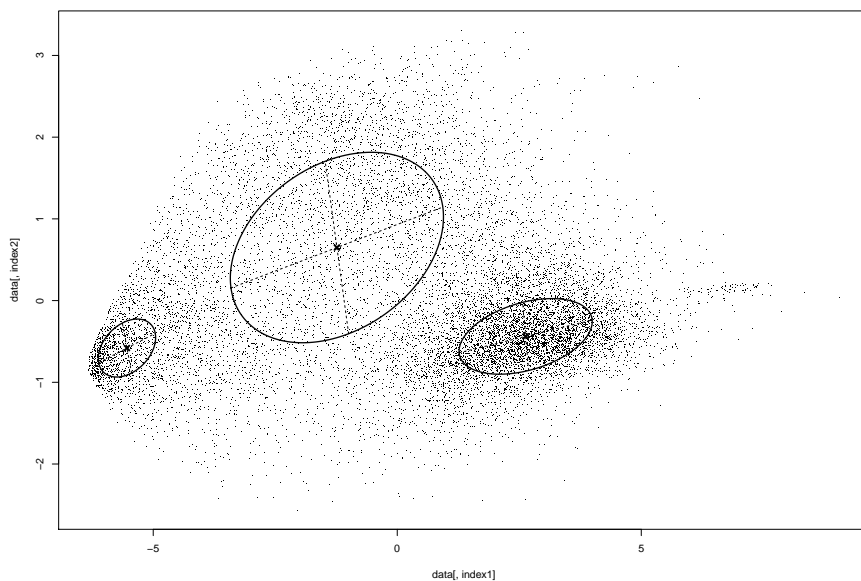


**Fig. 10** *Bubbles* experiment. Frequencies of choices of the number of mixture components with the criteria SH, BIC, and ICL for the spherical mixture model  $[p_k \lambda_k I]$  on the *Bubble* data set

terized the expression dynamics over 27 distinct stages of development during the life cycle of the fly using RNA-seq. Rau et al. (2015) use a subset of these data from 12 embryonic samples that were collected at two-hour intervals for 24 hours, with one biological replicate for each time-point to illustrate a clustering method derived from a Poisson mixture. The phenotype tables and raw read counts for the 13,164 genes with at least one non-zero count among the 12 time-points were obtained from the ReCount online resource (Frazee et al., 2011). In this paper, we consider this data set to illustrate initialization strategies of the EM algorithm for Gaussian mixture on a large data set. For this purpose, we transform the count data  $\mathbf{c}$  with the  $\log(c_j + 1)$ ,  $j = 1, \dots, 12$  transformation. Then we apply a PCA on the resulting data set in  $\mathbb{R}^{12}$  and

apply a Gaussian mixture model to the first two coordinates of the PCA. The analyzed data set is represented in Figure 11.

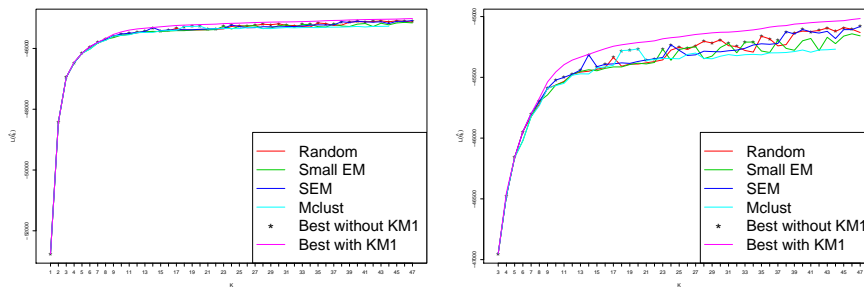
For this data set, the superiority of KM1 is more marked. Moreover, the strategies without KM1 produce clearly non increasing likelihoods in function of  $K$  (Figure 12). ICL with KM1 chooses a sensible number of clusters ( $K = 3$ ) depicted in Figure 11. The choice of  $K$  in a density estimation purpose seems more difficult. As a matter of fact, with the KM1 strategy, BIC chooses  $K = 12$  components while SH chooses  $K = 17$  components (see Table 1).



**Fig. 11** The *Drosophila* data set represented in the plane of its first principal components. The ellipses depict the three Gaussian components chosen by ICL with the KM1 strategy

Crit.	Small	SEM	Random	Mclust	Best without KM1	Best with KM1
SH	12	14	9	20	14	17
BIC	12	14	12	18	14	12
ICL	4	4	4	4	4	3

**Table 1** *Drosophila* data set. Number of mixture components selected by the model selection criteria SH, BIC and ICL.



**Fig. 12** *Drosophila* data set. Behavior of the maximized likelihood for different strategies as a function of the number of components ; on the right figure, a zoom view.

## 6 Discussion

We highlight some difficulties which can arise with the EM algorithm for maximizing the likelihood of a Gaussian mixture model. The degeneracies of the likelihood can be addressed by a Bayesian regularization. But the initialization issue remains and can be a most influential factor.

To avoid the traps associated to degeneracies, we replace the unbounded loglikelihood with a regularized loglikelihood derived from a Bayesian regularization (Fraley and Raftery, 2007). We modify their prior hyperparameter choices to get less regularized loglikelihoods, the strength of the regularization depending on a single scalar hyperparameter  $\sigma_0$ . Thus the degeneracies in maximum likelihood methods are avoided. But it is important to choose carefully  $\sigma_0$  to ensure stable results and meaningful estimates.

Biernacki et al. (2003) proposed random Search/Run/Select initialization strategies as Small EM and SEM. They stated that no strategy outperforms the others and indicated that likelihood singularities can make the task more difficult. The present study confirms the difficulty to favor a strategy against the others. Actually, these authors remark that in their experiments the Small EM strategy is slightly better than the others. But in our numerical experiments, SEM seems to be slightly better than all the others, including Small EM...

Beside random Search/Run/Select strategies, we propose recursive initialization strategies, so called KM1, which initialize the  $(K + 1)$  component mixture by the use of the  $K$  components solution. These recursive initialization strategies start from solutions which can be expected to be close to the ml estimate. Thus the EM algorithm is expected to converge faster with them than with fully random initializations. In this framework, we have defined several criteria to choose the component to be split. They all appear to be outperformed by the KM1 Complete procedure which consists of trying to split all the  $K$  components and splitting the component leading to the largest likelihood. Obviously this complete procedure is up to  $K$  times more expensive than the procedures using a criterion to split a component. But, this price

remains often reasonable and ensures almost consistently better results. It is noteworthy that, on the contrary to the other initialization strategies, KM1 provides increasing loglikelihood graphs.

The influence of the initialization procedures on the choice of the complexity of a model can be important. It is not so important with the clustering criterion ICL which provides quite stable results. It is more important with the slope heuristics SH and BIC. In particular, Small EM and KM1 procedures can lead to different choices. And, in such cases, the solutions provided with the KM1 procedure can be preferred. On the other hand, it is rather difficult from these numerical experiments to compare the behavior of SH and BIC criteria, but it is beyond the scope of this article. BIC may be more able to recover the true number of components when it exists and provide stabler selections. Notice also that the KM1 strategy is helpful to produce a consistent estimation of the slope of the maximized likelihood.

Finally let us stress the importance of the choice of the parameter  $\sigma_0$  to get an honest regularization. This is the topic of a current work.

## References

- Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.
- Baudry, J.-P. (2009). *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes*. PhD thesis, Université Paris-Sud.
- Baudry, J.-P., Maugis, C., and Michel, B. (2011). Slope heuristics: overview and implementation. *Statistics and Computing*, 22:455–470.
- Berchtold, A. (2004). Optimisation of mixture models: Comparison of different strategies. *Computational Statistics*, 19:385–406.
- 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*, 22:719–725.
- 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.
- Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73.
- 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.
- Ciuperca, G., Ridolfi, A., and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30:45–59.

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Fraley, C., Raftery, A., and Wehrens, R. J. (2005). Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics*, 14:529–546.
- Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631.
- Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24:155–181.
- Frazee, A. C. et al. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(449).
- Graveley, B. R. et al. (2011). The development transcriptome of *Drosophila melanogaster*. *Nature*, 471:473–479.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A*, 62(1):49–66.
- McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions, Second Edition*. Wiley.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
- Papastamoulis, P., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2014). On the estimation of mixtures of poisson regression models with large numbers of components. *Computational Statistics & Data Analysis*. To appear.
- Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In Langley, P., editor, *ICML*, pages 727–734. Morgan Kaufmann.
- Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*. (to appear).
- Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.