Different points of view for selecting a latent structure model

Gilles Celeux

Inria Saclay-Île-de-France, Université Paris-Sud

Latent structure models: two different point of views

Density estimation

- LSM can be regarded as a versatile semi parametric tool for estimating density.
- It allows to deal with heterogeneity in the data. The latent structure is not of particular interest.

Cluster analysis

- In this perspective, the latent structure is of primary interest.
- For instance Gaussian mixture is the most used model in Model-based clustering (MBC).
- The aim is to estimate and interpret the hidden structure in the data.

Estimating LSM parameters

Two approaches:

Standard statistical inference

- Estimating the mixture parameters (through maximum likelihood or Bayesian inference)

Clustering inference

- Simultaneous estimation of both the model parameters and the latent structure

Quantitative data: multivariate Gaussian Mixture (MGM)

Multidimensional observations $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ in $\mathbb{R}^d$ are assumed to be a sample from a probability distribution with density

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_k \pi_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

where

- π_k : mixing proportions
- $\phi(.|\boldsymbol{\mu}_k, \Sigma_k)$: Gaussian density with mean $\boldsymbol{\mu}_k$ and variance matrix Σ_k.

This is the most popular model for clustering of quantitative data.

Qualitative Data: latent class model (LCM)

- Observations to be classified are described with d qualitative variables.
- Each variable j has m_j response levels.

Data $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are defined by

$$\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$$

with

$$\begin{cases} x_i^{jh} = 1 & \text{if } i \text{ has response level } h \text{ for variable } j \\ x_i^{jh} = 0 & \text{otherwise.} \end{cases}$$

The standard latent class model (LCM)

Data are supposed to arise from a mixture of g multivariate multinomial distributions with pdf

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k m_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \sum_k \pi_k \prod_{j,h} (\alpha_k^{jh})^{x_i^{jh}}$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \alpha_1^{11}, \ldots, \alpha_g^{dm_d})$ is the parameter of the latent class model to be estimated :

- α_k^{jh} : probability that variable j has level h in cluster k,
- π_k : mixing proportions

Latent class model is assuming that the variables are conditionnally independent knowing the latent clusters.

EM algorithm (maximum likelihood estimation)

Algorithm

- **Initial Step** : initial solution $\boldsymbol{\theta}^0$
- **E step**: Compute the conditional probabilities t_{ik} that observation i arises from the kth component for the current value of the mixture parameters:

$$t_{ik}^m = \frac{\pi_k^m \varphi_k(\boldsymbol{x}_i; \alpha_k^m)}{\sum_\ell \pi_\ell^m \varphi_\ell(\boldsymbol{x}_i; \alpha_\ell^m)}$$

- **M step**: Update the mixture parameter estimates maximising the expected value of the completed likelihood. It leads to weight the observation i for group k with the conditional probability t_{ik}.
  - $\pi_k^{m+1} = \frac{1}{n} \sum_i t_{ik}^m$
  - α_k^{m+1} : Solving the Likelihood Equations

Features of EM

- EM is increasing the likelihood at each iteration
- Under regularity conditions, convergence towards the unique consistent solution of likelihood equations
- Easy to program
- Good practical behaviour
- Slow convergence situations (especially for mixtures with overlapping components)
- Many local maxima or even saddle points
- Quite popular: see the McLachlan and Krishnan book (1997)

Classification EM

The CEM algorithm, clustering version of EM, estimate both the mixture parameters and the labels by maximising the completed likelihood

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{k,i} z_{ik} \log \pi_k f(\mathbf{x}_i; \alpha_k)$$

Algorithm

- **E step**: Compute the conditional probabilities t_{ik} that observation i arises from the kth component for the current value of the mixture parameters.
- **C step**: Assign each observation i to the component maximising the conditional probability t_{ik} (MAP principle)
- **M step**: Update the mixture parameter estimates maximising the completed likelihood.

Features of CEM

- CEM aims maximising the complete likelihood where the component label of each sample point is included in the data set.
- Contrary to EM, CEM converges in a finite number of iterations
- CEM provides biased estimates of the mixture parameters.
- CEM is a *K-means*-like algorithm.

Model-based clustering via EM

Relevant clustering can be deduced from EM

- Estimating the mixture parameters with EM
- Computing of t_{ik}, conditional probability that observation $\boldsymbol{x}_i$ comes from cluster k using the estimated parameters.
- Assigning each observation to the cluster maximising t_{ik} (MAP : Maximum a posteriori)

This strategy could be preferred since CEM provides biased estimates of the mixture parameters.
But CEM is doing the job for well-separated mixture components.

Choosing the number of components

A model selection problem

- All models are wrong but some are useful (G. Box)
- The problem does not restrict to solve the bias-variance dilemma
- The problem is to choose a useful number of components
- This choice cannot be independent of the modelling purpose

Criteria for choosing g in a density estimation context

The AIC criterion

AIC is approximating the expected deviance of a model m with ν_m free parameters. Assuming that the data arose from a distribution belonging to the collection of models in competition, AIC is

$$\mathrm{AIC}(m) = 2 \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m) - 2\nu_m.$$

The BIC criterion

BIC is a pseudo-Bayesian criterion. It is approximating the integrated likelihood of the model m

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|m, \theta_m)\pi(\theta_m)d\theta_m,$$

$\pi(\theta_m)$ being a prior distribution for parameter θ_m,

$$\mathrm{BIC}(m) = \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m) - \frac{\nu_m}{2} \log(n).$$

Practical behaviour of BIC

Despite theoretical difficulties in the mixture context

- Simulation experiments (see Roeder & Wasserman 1997) show that BIC works well at a practical level to choose a sensible Gaussian mixture model,
- See also the good performances of a cross-validated likelihood criterion proposed by Smyth (2000).

Choosing a clustering model

Since BIC does not take into account the clustering purpose for assessing m, BIC has a tendency to overestimate g regardless of the separation of the clusters.

Choosing g in a clustering perspective

The ICL criterion

The integrated completed log-likelihood is

$$\log \mathbf{p}(\mathbf{x}, \mathbf{z} \mid m) = \log \int_{\Theta_m} \mathbf{p}(\mathbf{x}, \mathbf{z} \mid m, \theta) \pi(\theta \mid m) d\theta,$$

It is closed form from conjugate non informative prior for the LCM. For GMM, its BIC-like approximation is

$$\text{ICL-BIC}(m) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} | m, \hat{\theta}) - \frac{\nu_m}{2} \log n,$$

where the missing data have been replaced by their most probable value for parameter estimate $\hat{\theta}$.

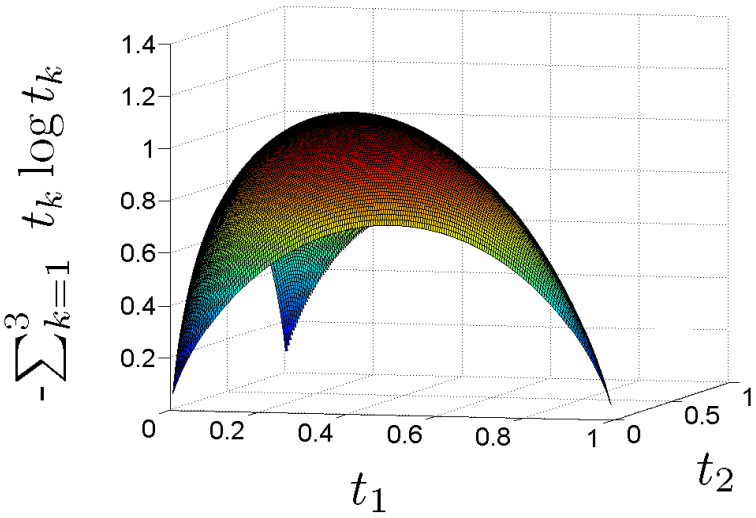
Roughly speaking criterion ICL-BIC is the criterion BIC penalized by the estimated mean entropy

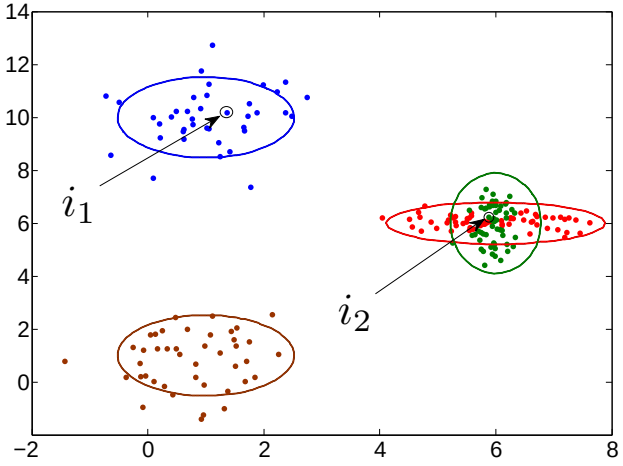
$$E(m) = -\sum_{ik} t_{ik}^m \log t_{ik}^m \geq 0.$$

The entropy: measure of the clustering confidence

$$\mathrm{ENT}(\theta; x) = -\sum_{k=1}^{K} t_{ik}(x;\theta) \log t_{ik}(x;\theta) \in [0, \log K].$$

$$\mathrm{ENT}(\theta) = \sum_{i=1}^{n} \mathrm{ENT}(\theta; x_i).$$





$\mathrm{ENT}(\widehat{\theta}_4^{\mathrm{MLE}}; x_{i_1})$ near 0.

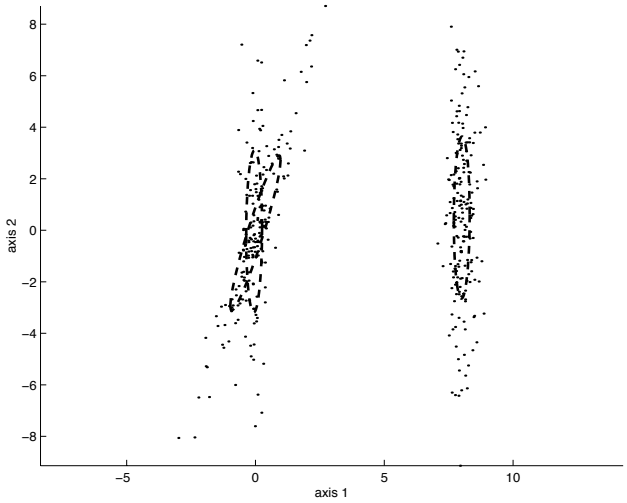
$\mathrm{ENT}(\widehat{\theta}_4^{\mathrm{MLE}}; x_{i_2})$ near $\log 2$.

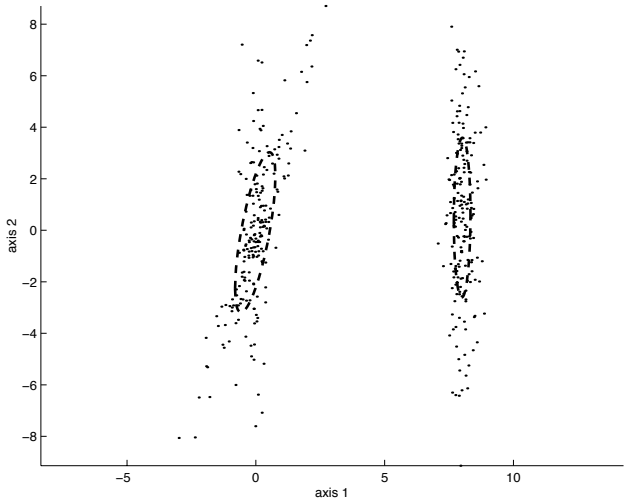
Behaviour of the ICL criterion

Because of this additional entropy term, ICL favors model giving rise to partitioning the data with the greatest evidence.

- ICL appears to provide a stable and reliable estimate of g for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping.
- But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components.

Contrasting BIC and ICL





Typical solutions proposed by BIC (left) (92%) and ICL (right) (88%) with the following features: Gaussian mixture with free variance matrices, $n = 400$.
The criteria select g and the form of the variance matrices from their eigenvalue decomposition.

- BIC outperforms ICL from the density estimation point of view...
- But from the cluster analysis point of view ?...

Contrast Minimisation in MBC context

The classification loglikelihood with the completed data $(\mathbf{x}, \mathbf{z})$ for model $\mathcal{M}_g$:

$$\log \mathrm{L_c}\big(\theta; (\mathbf{x}, \mathbf{z})\big) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log \pi_k \phi(\mathbf{x}_i; \mu_k, \Sigma_k).$$

An important relation is

$$\log \mathrm{L_c}(\theta) = \log \mathrm{L}(\theta) + \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log t_{ik}(\mathbf{x}_i; \theta).$$

Taking the conditional expectation of this relation leads to

$$\log \mathrm{L_{cc}}(\theta) = \log \mathrm{L}(\theta) - \mathrm{ENT}(\theta).$$

and $\log \mathrm{L_{cc}}(\theta)$ the conditional expectation of the complete loglikelihood is an alternative criterion to maximum likelihood.

ICL revisited

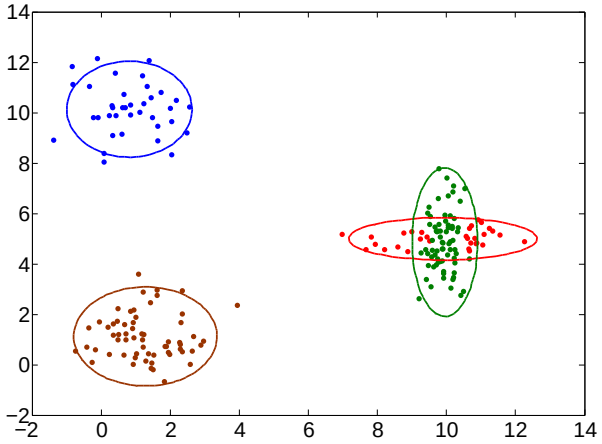
- By analogy with BIC, we get the consistent criterion $\mathrm{L_{cc}}$-ICL

$$\widehat{g}^{\mathrm{L_{cc}\text{-}ICL}} = \underset{g \in \{1,\ldots,g_M\}}{\operatorname{argmin}} \left\{ -\log \mathrm{L_{cc}}(\widehat{\theta}_g^{\mathrm{MLccE}}) + \frac{\nu_g}{2} \log n \right\}.$$

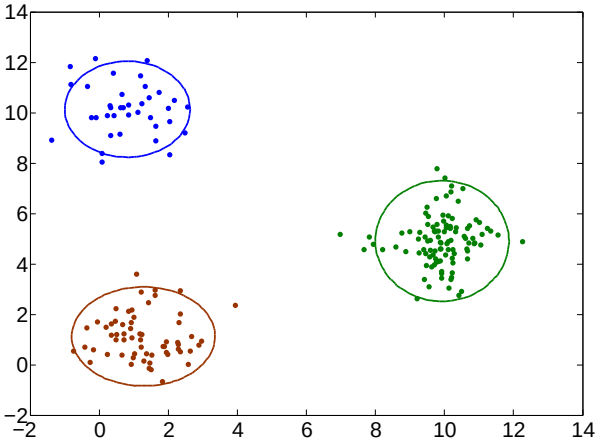
- ICL can be regarded as an approximation of $\mathrm{L_{cc}}$-ICL :

$$\widehat{g}^{\mathrm{ICL}} = \underset{g \in \{1,\ldots,g_M\}}{\operatorname{argmin}} \left\{ -\log \mathrm{L_{cc}}(\widehat{\theta}_g^{\mathrm{MLE}}) + \frac{\nu_g}{2} \log n \right\}.$$

"Cluster is not mixture component"

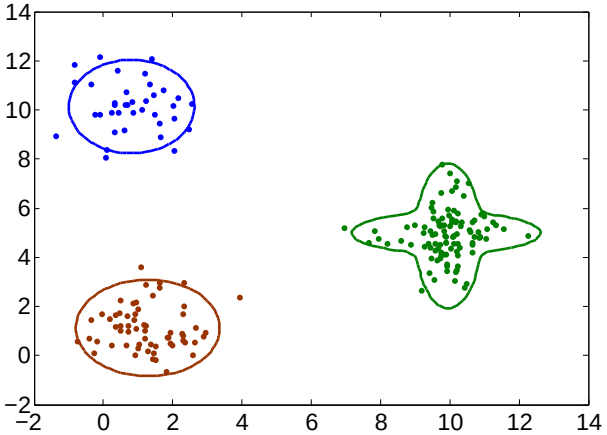


Solution selected by BIC



Solution selected by ICL

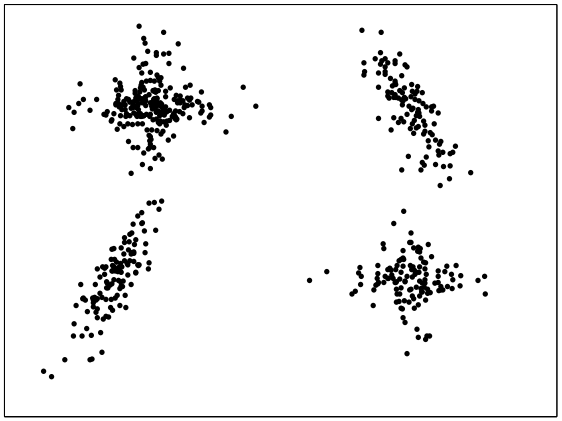
Combiner deux classes



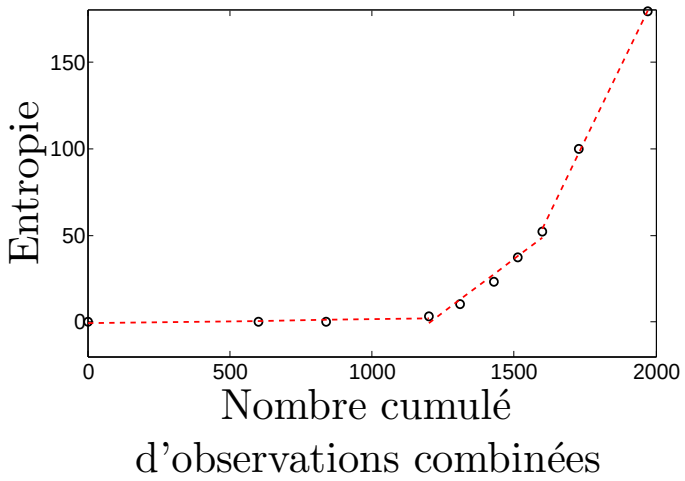
Combined Solution

Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).
The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Simulated Data

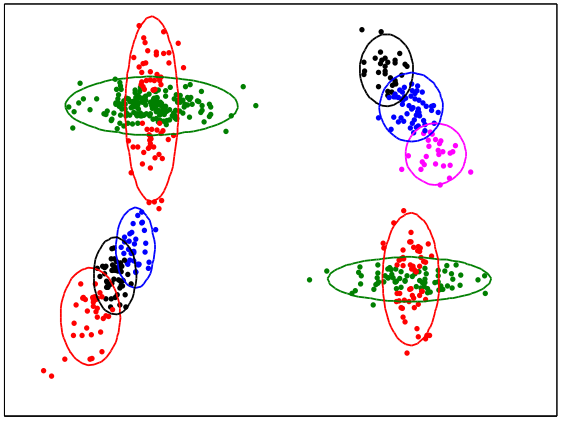


Combined solutions entropy

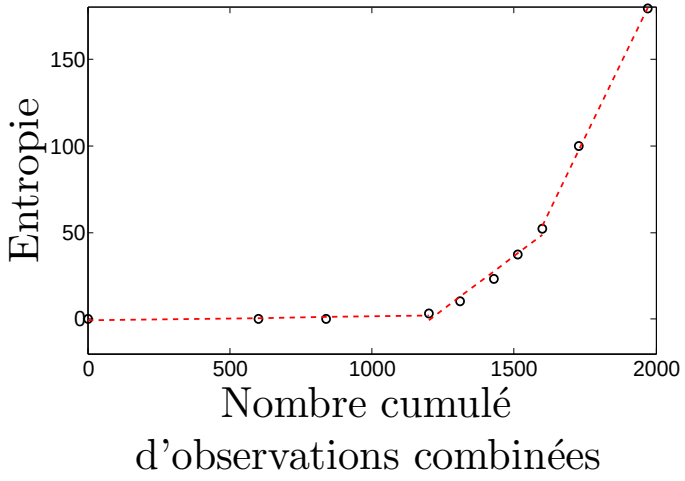
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



BIC (K=10)

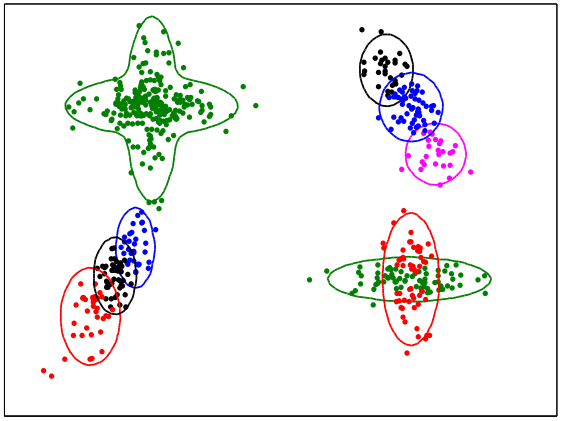


Combined solutions entropy

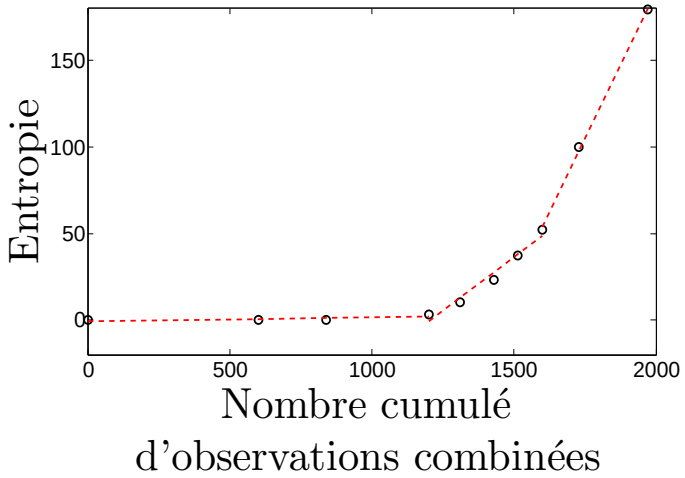
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=9)

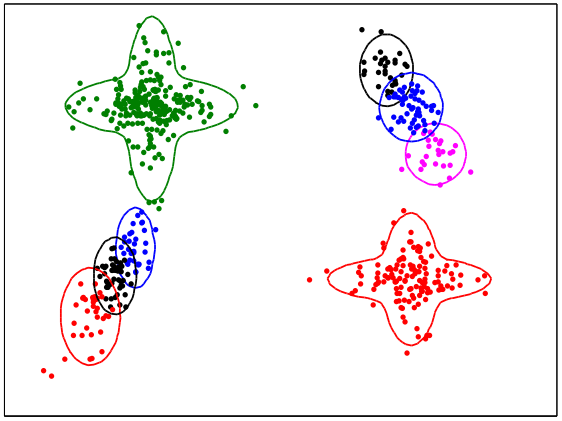


Combined solutions entropy

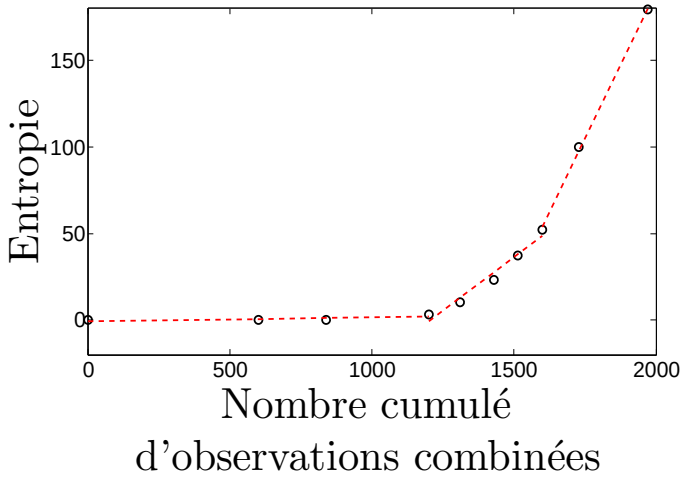
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=8)

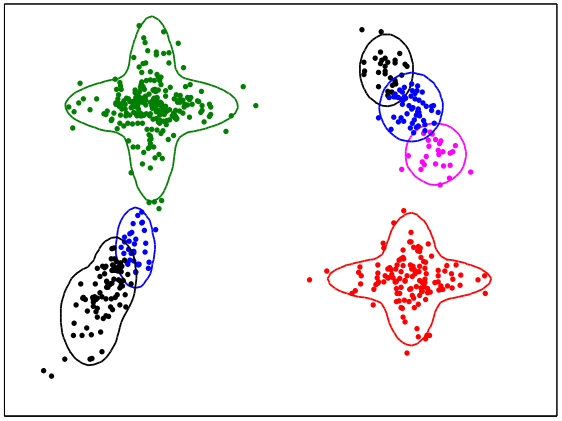


Combined solutions entropy

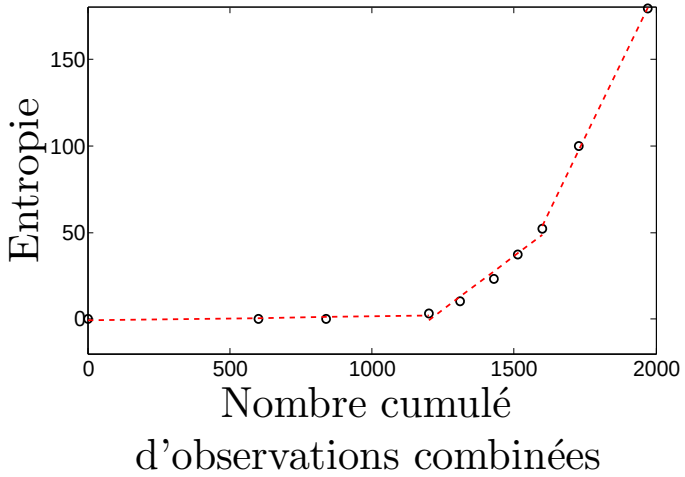
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=7)

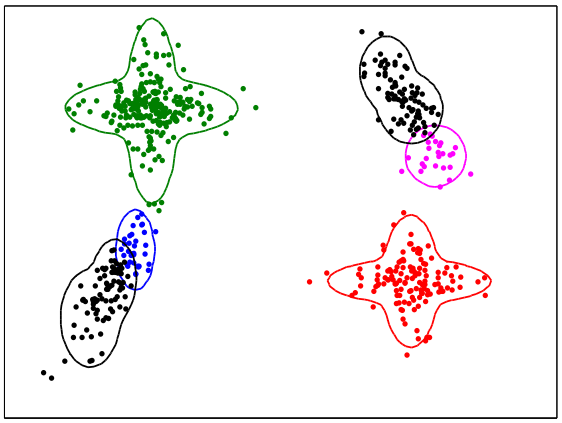


Combined solutions entropy

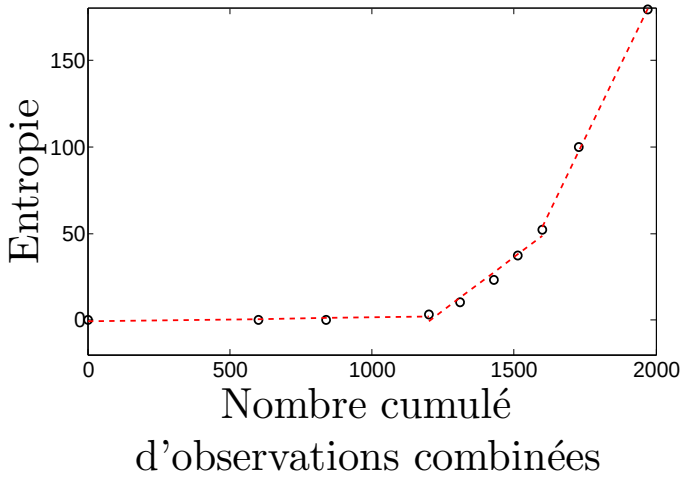
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=6)

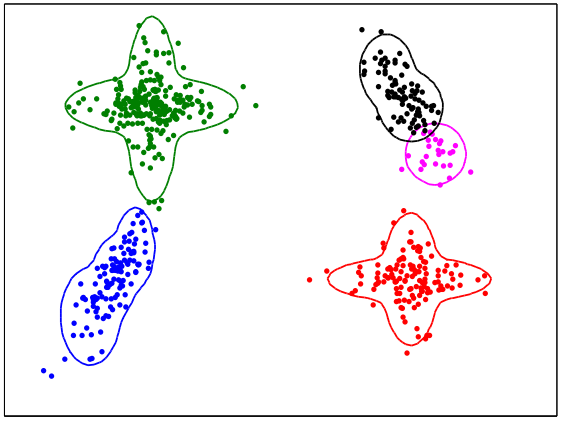


Combined solutions entropy

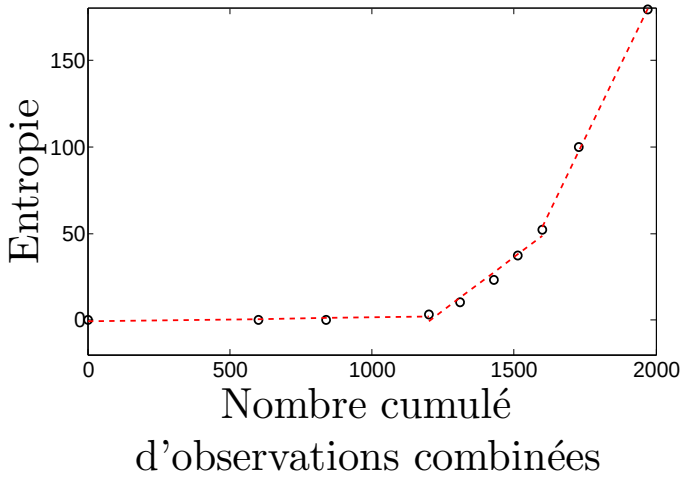
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=5)

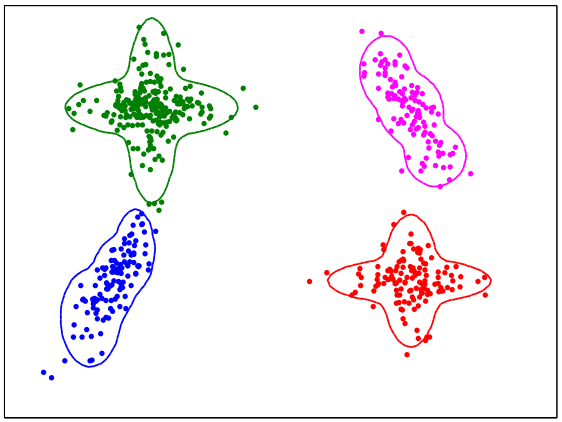


Combined solutions entropy

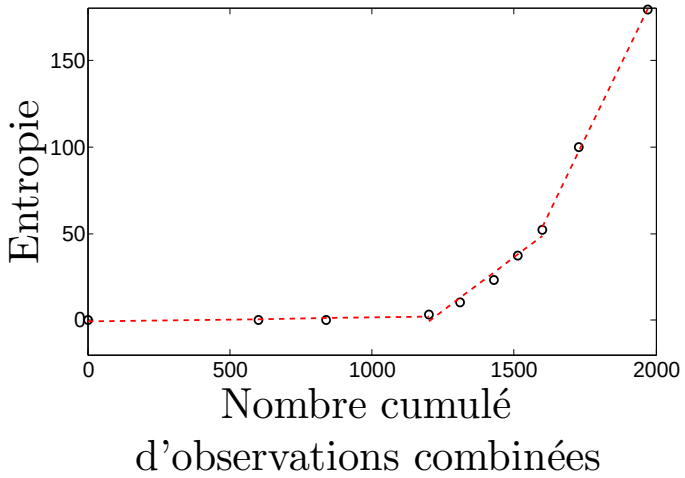
Combining mixture components

It leads to a hierarchical combining of mixtures components by merging at each step the two components maximising the decrease of the entropy starting from the BIC solution (Baudry et al., JCGS 2010).

The graph of the entropy in function of the number of clusters is helpful to choose a sensible number of clusters.



Combined solution (K=4)



Combined solutions entropy

Bayesian inference

Choosing conjugate priors

- Mixing proportions π: Dirichlet priors $\mathcal{D}(a, \ldots, a)$ are non informative priors, with $a = 1/2$ we get the Jeffreys prior.
- The choice of a has been proved to be quite sensitive (Frühwirth-Schnatter, 2011) to select the number of components.
- Dirichlet non informative priors are also possible for qualitative models.
- In the continuous case, the conjugate priors for $\alpha = (\mu, \sigma^2)$ are weakly informative.

Priors for the number of components

This sensitive choice jeopardizes Bayesian inference for mixtures (Aitken 2000, 2010).
Choosing truncated Poisson $\mathcal{P}(1)$ priors over the range $1, \ldots, g_{\max}$ is often a reasonable choice (Nobile 2005).

Standard MCMC

The first task is to approximate the posterior distribution of the LSM parameters.

Gibbs sampling

- With fixed g, Gibbs sampling can be regarded as a reference method to derive Bayesian estimates for latent structure models.
- With unknown g, the possibility to estimate it in the same exercise exists thanks to Reversible Jump MCMC.
- But, in my opinion, RJMCMC algorithms remain unattractive despite efforts to improve them.

Collapsed model

The clustering view point

Considering $\mathbf{z}$ as a parameter, leads to computing the collapsed joint posterior

$$P(g, \mathbf{z}|\mathbf{x}) = P(g) CF(.) \prod_{k=1}^{g} M_k$$

where $CF(.)$ is a closed form function made of Gamma functions and

$$M_k = \int P(\alpha_k) \prod_{i/z_i=k} p(\mathbf{x}_i|\alpha_k) \mathrm{d}\alpha_\mathrm{k}.$$

The allocation sampler

The point of the allocation sampler of Nobile and Fearnside (2007) is to use a (RJ)MCMC algorithm on the collapsed model.

Moves with fixed numbers of clusters

- Updating the label of unit i in cluster k:

$$P(\tilde{z}_i = k') \propto \frac{n'_k + 1}{n_k} \frac{M_{k'}^{+i} M_k^{-i}}{M_{k'} M_k}, k' \neq k.$$

- Other moves are possible (Nobile and Fearnside 2007).

Moves to split or combine clusters

Two reversible moves to split a cluster or combine two clusters analogous to the RJMCMC moves of R & G'97 are defined. But, thanks to collapsing, those moves are of fixed dimension. Integrating out the parameters leads to reduce the sampling variability.

The allocation sampler: label switching

Following Nobile, Fearnside (2007), Friel and Wyse (2010) used a post-processing procedure with the cost function

$$C(k_1, k_2) = \sum_{t=1}^{C-1} \sum_{i=1}^{n} I\left\{z_i^{(c)} \neq k_1, z_i^{(C)} = k_2\right\}.$$

1. The $\mathbf{z}^{(c)}$ MCMC sequence has been rearranged such that for $s < c$, $\mathbf{z}^{(s)}$ uses less or the same number of components than $\mathbf{z}^{(c)}$.
2. An algorithm returns the permutation $\sigma(.)$ of the labels in $\mathbf{z}^{(c)}$ which minimises the total cost $\sum_{k=1}^{g_{C-1}} C(k, \sigma(k))$.
3. $\mathbf{z}^{(c)}$ is relabelled using the permutation $\sigma(.)$.

Remarks on the procedure to deal with label switching

- Due to collapsing, the cost function does not involve sampled model parameters.
- Simple algebra lead to an efficient on-line post-processing procedure.
- When g is large, $g!$ is tremendous.

Summarizing MCMC output

Using the modal cluster model

- The $\hat{g}$ which appeared most often is chosen,
- the N label vectors $\mathbf{z}$ are extracted from the MCMC sample and post processed to undo label switching,
- then, the posterior distributions of cluster membership $(t_{i1}, \ldots, t_{i\hat{g}})$ are estimated by their frequencies in the MCMC sample,
- and, i is assigned to cluster $\text{argmax}_k \, t_{ik}$.

Using the MAP

The maximum a posteriori model is the visited $(g, \mathbf{z})$ having highest probability a posteriori from the MCMC sample.

A case study

ML and Bayesian aprroaches are compared on a real data set from a clustering point of view using the Latent Block Model.

The Latent Block Model is mixture model with two latent structures, one for the rows, one for the columns.

The data set records the votes of 435 members (267 democrats, 168 republicans) of the 98th Congress on 16 different key issues.

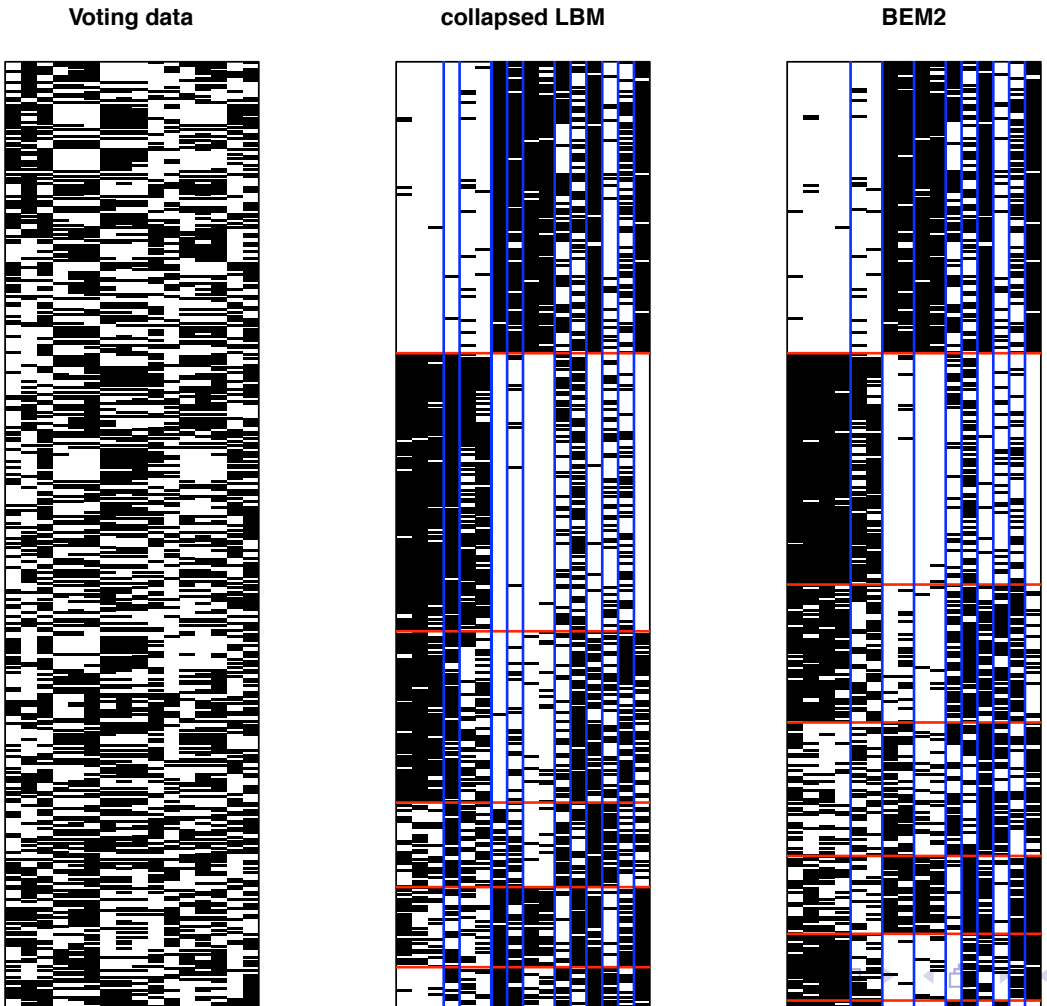
For each issue, the possible votes were yes, no and abstention.

To restrict the analysis to binary variables, the votes no and abstention have been grouped.

Bayesian Analysis

Wyse and Friel (2010) used non informative priors. The sampler has been run 220,000 iterations with 20,000 for burn-in.

It leads them to select a $(g = 7, m = 12)$ solution.



Maximum Likelihood Analysis

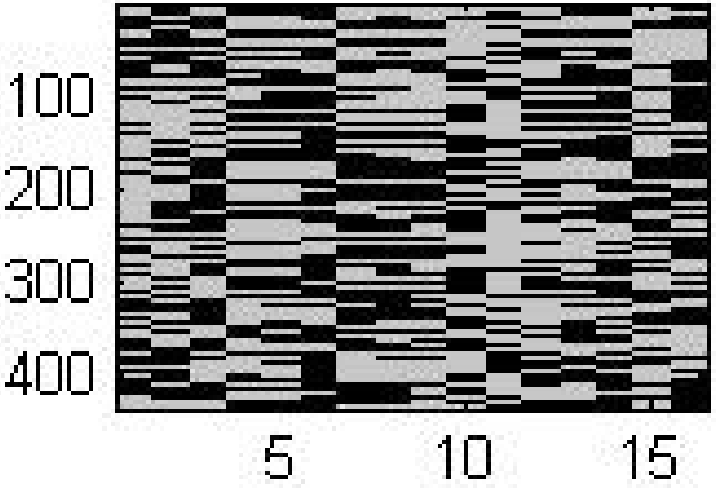
- We use a Stochastic EM algorithm with $g = 2, ..., 8$ and $m = 2, ..., 12$.
- Using ICL, the best solution was obtained with $(g = 5, m = 4)$ clusters with the following distribution for the two political parties

	Republicans	Democrats
1	39	38
2	0	139
3	121	7
4	8	83

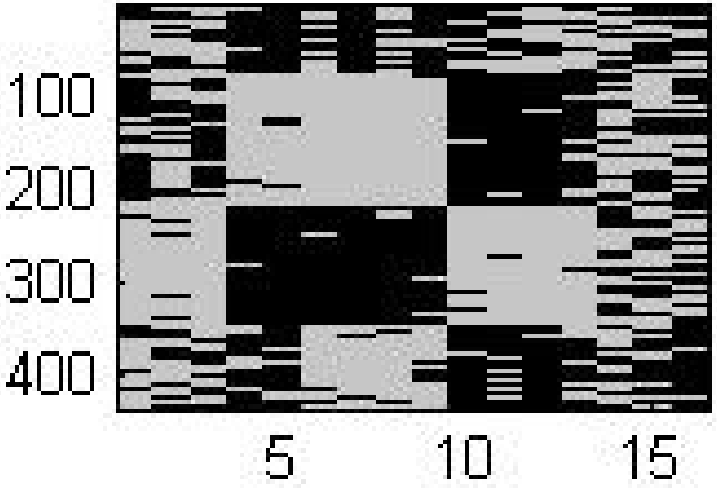
- Using the pseudo BIC criterion, the best solution was obtained with $(g = 3, m = 6)$ clusters. with the following distribution for the two political parties

	Republicans	Democrats
1	134	24
2	32	79
3	2	164

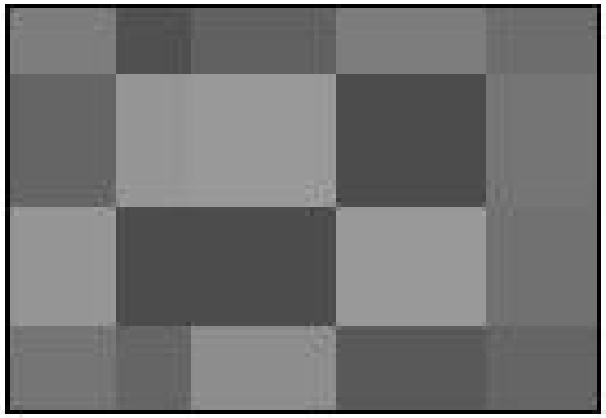
The ICL solution



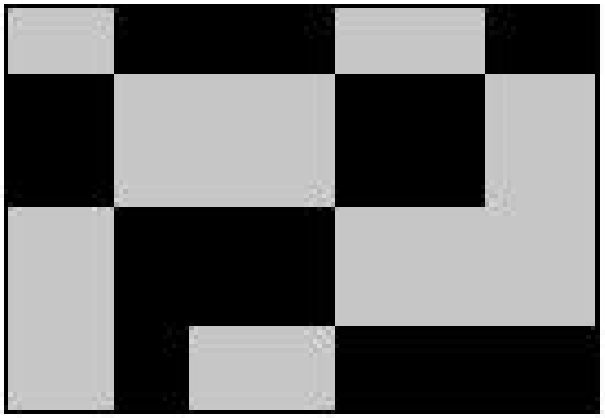
Initial data



Reorganized data



Reorganized and averaged data



Summary

Concluding remarks

- The focus of the analysis is of primary importance for selecting a LSM.
- This focus could also influence the choice of the algorithm to estimate LSM with large data set (CEM...)
- Bayesian analysis for LSM suffers several drawbacks: the more important one is the label switching problem.
- Standard methods to deal with this problem require to identify $g!$ clusters.
- Frühwirth-Schnatter (2005, 2011) proposed a k-means clustering method in the point process representation of the MCMC draws to identify g clusters instead of $g!$ clusters.
- Collapsing leads to promising Bayesian methods for LSM in a clustering context.
- But, in a high dimensional setting, Bayesian analysis of LSM remains difficult...

References

Aitkin (2011) How many components in a finite mixture? In *Mixtures: Estimation and Applications*, Chapter13, Wiley

Baudry, Raftery, Celeux, Lo and Gottardo (2010) Combining mixture components for Clustering, *Journal of Computational and Graphical Statistics*, **19**, 332-353.

Biernacki, Celeux and Govaert (2000) Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Trans. on PAMI*, **22**, 719-725.

Frühwirth-Schnatter (2006) *Finite Mixture and Markov Switching Models*. Springer

Frühwirth-Schnatter (2011) Dealing with label switching under model uncertainty. In *Mixtures: Estimation and Applications*, Chapter 10 , Wiley

McLachlan and Peel (2000) *Finite Mixture Models*. Wiley

Nobile and Fearnside (2007) Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, **17**, 147-162.

Wyse and Friel (2010) Block clustering with collapsed latent block models. *Statistics and Computing* **22**, 415-428.

A software: MIXMOD http://www.mixmod.org