Sparse Bayesian Finite Mixtures

Gertraud Malsiner-Walli

WU Wirtschaftuniversität Wien

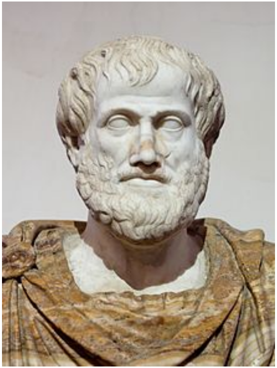
joint work with
Sylvia Frühwirth-Schnatter and Bettina Grün

Funded by the Austrian Science Fund (FWF P25850, V170 , P28740)
and Austrian National Bank (Jubiläumsfond 14663)

Brown Bag Seminar WU , January 18^{th} 2017

Outline

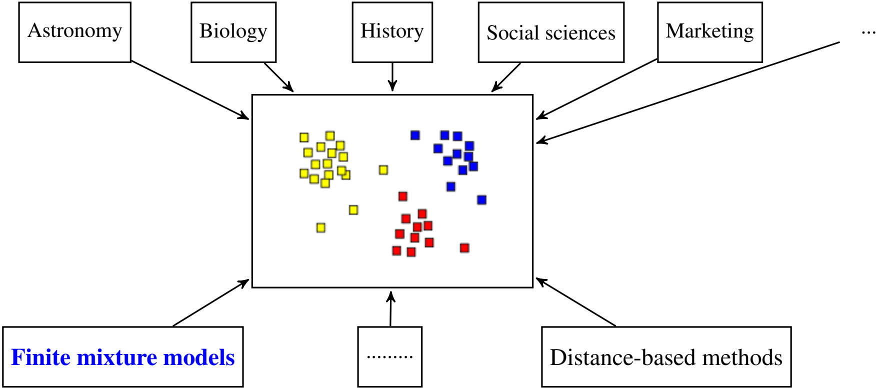
"Sapientis est Ordinare"



Aristotle,
384 – 322 BC

"It belongs to the wise person to create order"

Cluster analysis



Cluster analysis based on a finite mixture model I

Model

1. Observations $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)$ are a sample from a **mixture distribution** with $\boldsymbol{\vartheta} = (\boldsymbol{\eta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k p_k(\mathbf{y}_i|\boldsymbol{\theta}_k),$$

where

- the component densities $p_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ arise from the same parametric family,
- $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ are the component weights, $\sum_{k=1}^{K} \eta_k = 1$, $\eta_k \geq 0$,
- it is assumed that each **component** corresponds to a **data cluster**,
- usually the **group membership** $S_i \in \{1, \ldots, K\}$ is unknown:

  $\Rightarrow$ they are introduced as latent allocation variables $\mathbf{S} = (S_1, \ldots, S_N)$ to indicate the component from which each observation is drawn:

$$p(\mathbf{y}_i|S_i = k) = p_k(\mathbf{y}_i|\boldsymbol{\theta}_k), \quad \text{where } Pr(S_i = k) = \eta_k$$

Cluster analysis based on a finite mixture model II

Bayesian framework

2. The mixture likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ is combined with the **prior** $p(\boldsymbol{\vartheta})$ and the **posterior** $p(\boldsymbol{\vartheta}|\mathbf{y})$ is obtained:

$$p(\boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}).$$

3. Estimation of the posterior distribution through standard **MCMC methods** based on data augmentation and Gibbs sampling.

   Start with some classification $\mathbf{S} = (S_1, \dots, S_N)$ and iterate the following steps:

   3.1 Parameter simulation conditional on the classification $\mathbf{S}$:

      3.1.1 Sample $\boldsymbol{\eta}$.

      3.1.2 Sample the component-specific parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$.

   3.2 Classification simulation conditional on the parameters $\boldsymbol{\vartheta}$:

      3.2.1 Sample $\mathbf{S} = (S_1, \dots, S_N)$.

Issues and approach

- Challenges in model-based clustering:
  - (a) Estimation of the **number of components**: crucial and old problem!
  - (b) Capturing **(Non-Gaussian) data clusters**: normal components?

- Our approach: **”prior modelling”**:
  $\Rightarrow$ Specification of ”suitable priors” on the mixture parameters.

  - $\Rightarrow$ To induce **characteristics** in model estimation we are interested in.
  - $\Rightarrow$ Not a “new” kind of prior families, rather well-known conditional **conjugate priors**.
  - $\Rightarrow$ **Hyperparameters** of the priors are chosen carefully and in a prudential way.
  - $\Rightarrow$ Prior specifications work **simultaneously** (joint approach).
  - $\Rightarrow$ Data can overwhelm the prior information if they are informative enough
    $\Rightarrow$ **Flexible** way of modeling!

Bayesian normal mixture model

- Gaussian mixtures:

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} \eta_k f_N(\mathbf{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

- Priors:

$$\begin{aligned}
\boldsymbol{\eta} &\sim Dir(e_0, \dots, e_0), \\
\boldsymbol{\mu}_k &\sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0), \\
\boldsymbol{\Sigma}_k &\sim \mathcal{W}^{-1}(c_0, \mathbf{C}_0) \qquad (\Leftrightarrow \boldsymbol{\Sigma}_k^{-1} \sim \mathcal{W}(c_0, \mathbf{C}_0)).
\end{aligned}$$

- Hyperparameters $e_0, \mathbf{b}_0, \mathbf{B}_0, c_0, \mathbf{C}_0$?

Estimating K

Overfitting mixture

- Comparison of candidate models with different K (e.g. BIC, Bayes factors) to select the model with the best fit.

⇒ **Overfitting mixture**: At some point in the process, the **number of components** must be **overfitted** i.e. $K > K^{true}$

⇒ Overfitting: **non-identifiability** of the model.

Non-identifiability due to overfitting:

- Overfitting mixtures: irregular likelihood (Sylvia FS, 2006).
- If $K > K^{true}$, there are two possibilities how to handle a **superfluous component**:

1. **weight** of a superfluous component is shrunken toward zero (component-specific parameter vector not identified),
2. **component-specific parameters vector** of the superfluous component is equal to a 'true' one, splitted components (weights are not identified).

Dirichlet prior on the weights I

Posterior of an overfitting mixture: $K > K^{true}$

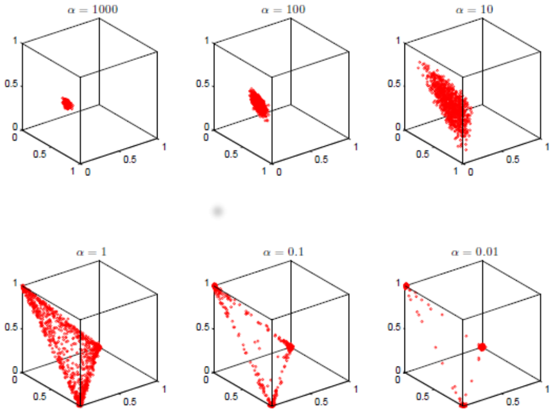
- Rousseau and Mengersen (2011) study the asymptotic behavior of the posterior distribution of an overfitting mixture model. They showed its shape depends on the prior on the weights:

$$\boldsymbol{\eta} \sim Dir(e_0, \dots, e_0)$$

  - If $e_0 < d/2$, $d = dim(\boldsymbol{\theta}_k)$, the posterior density handles overfitting by asymptotically shrunking weights of superfluous components towards 0, i.e. they are left **empty**.
  - If $e_0 > d/2$, the posterior density handles overfitting by forming at least two identical components, i.e. splitted components, '**filled**' components.

Dirichlet prior on the weights II

Dirichlet(α, α, α) distribution:



Plot by Chris Holmes and Chris Yau, Edinburgh, 2010, meeting ”Mixture estimation and Application”.

Dirichlet prior on the weights III

To select $\mathbf{K}^{\mathbf{true}}$:

"Decide through the Dirichlet prior whether you prefer **empty** components or **duplicated** components for overfitting mixtures" (Frühwirth-Schnatter, 2012).

- By calculating marginal likelihoods $p(\mathbf{y}|K)$ or the posterior $p(K|\mathbf{y})$ in RJMCMC:

$\Rightarrow$ Interest lies in **filling** all specified components

$\Rightarrow$ Specify a **redundant** prior on the mixture weights (i.e. $e_0 > d/2$).

- By estimating the number of non-empty components:

$\Rightarrow$ Interest lies in **emptying** superfluous components:

$\Rightarrow$ Specify a **sparse** prior on the mixture weights (i.e. $e_0 < d/2$)

Sparse finite mixtures (GMW, Sylvia FS, Bettina G, 2016)

Estimation of the number of mixture components:

$\Rightarrow$ Specify an **overfitting** mixture model ($K > K^{true}$).

$\Rightarrow$ Specify a **sparse prior on the weights** $\boldsymbol{\eta}$: choose e_0 small.

$\Rightarrow$ For each iteration m consider the number of **non-empty components** $K_+^{(m)}$.

$\Rightarrow$ Estimate K^{true} by the **most frequent number of non-empty components:**

$$\hat{K}_+ = mode\{p(K_+|\mathbf{y})\}$$

$\Rightarrow$ "Automatic" tool to select the number of components!

Mixture components versus data clusters

Note: Sparse finite mixtures

- make a distinction between
  - **K** (number of specified components) and
  - $\mathbf{K}_+$ (the number of non-empty components).

We assume that

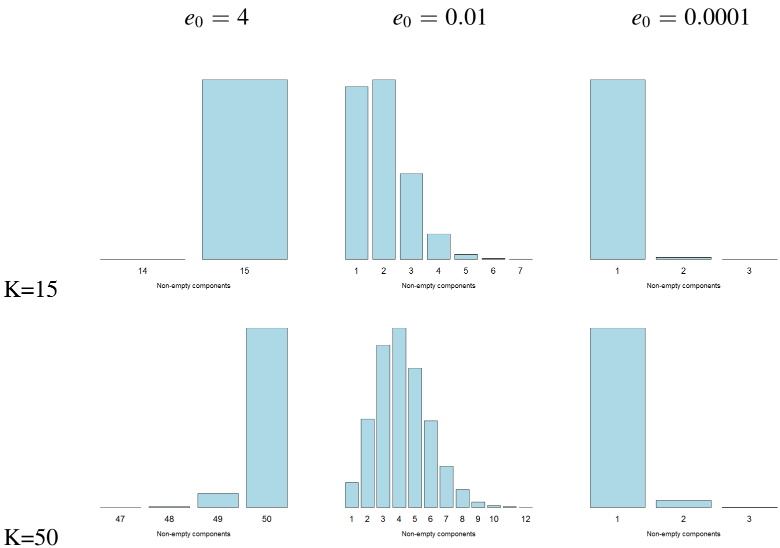
- K is fixed parameter,
- $\mathbf{K}_+$ is a **random variable**:
  - **a priori** the number K_+ depends on both e_0 and K (fixed parameters), i.e.

$$p(K_+|K, e_0),$$

  - **a posteriori** the number K_+ of non-empty groups can be estimated,

$$p(K_+|y).$$

Prior of K_+


$e_0 = 4$
$e_0 = 0.01$
$e_0 = 0.0001$
K=15
14
15
Non-empty components
1
2
3
4
5
6
7
Non-empty components
1
2
3
Non-empty components
K=50
47
48
49
50
Non-empty components
1
2
3
4
5
6
7
8
9
10
12
Non-empty components
1
2
3
Non-empty components

Simulation study I

Simulation study:

- Component means $\boldsymbol{\mu}_1 = (2, -2, 0, 0)'$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (2, 2, 0, 0)'$, and $\boldsymbol{\mu}_4 = -\boldsymbol{\mu}_3$ and isotropic covariance matrices $\boldsymbol{\Sigma}_k = \mathbf{I}_4$, $k = 1, \ldots, 4$.
- $\boldsymbol{\eta} = (0.25, 0.25, 0.25, 0.25)$.

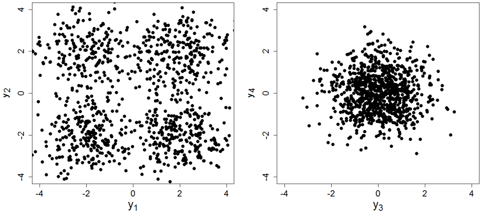


Figure: Scatter plots of one randomly selected data set.

Simulation study II

K	e_0 fixed	$\hat{K}_+$	MCR	MSE_μ
4	0.01	**4**	0.047	0.136
15	0.01	**4**	0.048	0.137
30	0.01	**4**(8)	0.048	0.136
30	0.001	**4**	0.048	0.136
30	0.00001	**4**	0.047	0.136

Table: Clustering results for different K.

Simulation study III

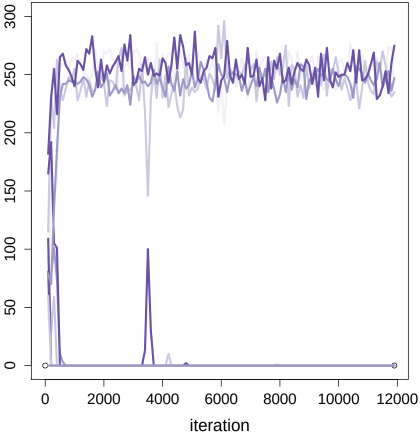


Figure: Number of observations allocated to the different components. MCMC run of a single data set, $K = 15$.

Simulation study IV

With a very **small component**: $\boldsymbol{\eta} = (0.02, 0.33, 0.33, 0.32)$:

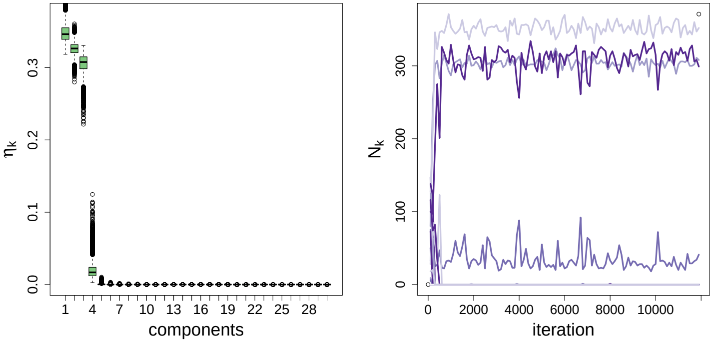


Figure: (unidentified) Posterior weight draws, sorted by size in each iteration, and trace plot of the number of observations allocated to the different mixture components.

Note: $K_+ \neq$ number of components with large(r) weights!

Sidestep: Relation to BNP approaches I

Bayesian Non-Parametrics (BNP) approach:

- Sparse finite mixtures are related to **infinite mixtures**, based on a Dirichlet process prior.
- A Dirichlet process prior $\mathcal{DP}(\alpha, \mathcal{G}_0)$ for $\mathbf{y}$ leads to **infinite mixture**

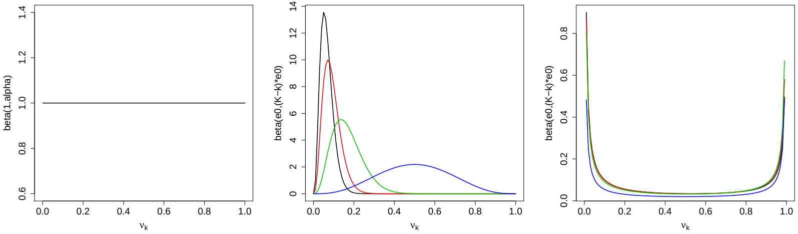
$$p(\mathbf{y}) = \sum_{k=1}^{\infty} \eta_k p_k(\mathbf{y}|\boldsymbol{\theta}_k).$$

- If the base measure $\boldsymbol{\theta} \sim \mathcal{G}_0$ is the same as the prior $p(\boldsymbol{\theta})$ in finite mixtures:
$\Rightarrow$ the only difference lies in the **prior of the weights** $\eta_1, \eta_2, \eta_3, \ldots$.
- The stick-breaking representation (Sethuraman, 1994) provides an connection in terms of the sticks $\nu_1, \nu_2, \nu_3, \ldots$:

$$\eta_1 = \nu_1, \quad \eta_2 = \nu_2(1-\nu_1), \quad \eta_k = \nu_k \prod_{j=1}^{k-1}(1-\nu_j), \quad \nu_k \sim Beta(a_k, b_k).$$

- For $\mathcal{DP}(\alpha, \mathcal{G}_0)$: $\quad \nu_k \sim Beta(1, \alpha)$.
For finite mixture: $\nu_k \sim Beta(e_0, (K-k)e_0), \quad \nu_K = 1$.

Sidestep: Relation to BNP approaches II



DP(α)	Finite mixture	Sparse finite mixture
$\nu_k \sim Beta(1,\alpha)$	$\nu_k \sim Beta(e_0, e_0(K-k))$	$\nu_k \sim Beta(e_0, e_0(K-k))$
$\alpha = 1$	$K = 15, e_0 = 4$	$K = 15, e_0 = 0.01$

Sidestep: Relation to BNP approaches III

- **Probability to create a new cluster**:
  DP mixture: $\frac{\alpha}{\alpha+N-1}$
  Finite mixture: $\frac{e_0(K-K_+^{-i})}{e_0K+N-1}$,
  K_+^{-1} is the number of non-empty clusters implied by $\mathbf{S}_{-i} = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_N)$.

- **Convergence:**
  A finite mixture with prior $\boldsymbol{\eta} \sim Dir(e_0)$ **converges** to a $\mathcal{DP}(\alpha)$ for $K \to \infty$ if

  $$e_0 = \alpha/K \quad \text{(Green and Richardson, 2001).}$$

- **Expected number of clusters**:
  DP mixture: $K_+ \propto \alpha log(N)$.
  Finite mixture: K_+ is asymptotically independent of N.

- **Conclusion**:
  - use infinite mixtures if you expect that the number of clusters **increases** for **increasing** data information,
  - use sparse finite mixtures if you do not!

Sidestep: Relation to BNP approaches IV

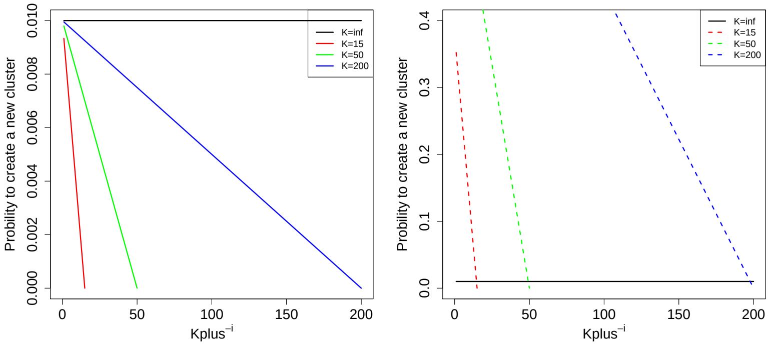


Figure: Probability to create a new cluster as a function of the already existing clusters K_+^{-i}:

- left: sparse finite mixtures with $e_0 = 1/K$,
- right: for finite mixtures with $e_0 = 4$,
- black line: for $K = \infty$.

Some benchmark data sets

Data set	N	r	$\boldsymbol{K_{true}}$	$\hat{K}_+$ for sparse finite mixtures ($K = 10, e_0 = 0.01$)
Iris	150	4	***3***	***3*** $adj = 0.92$, $er = 0.03$
Crabs	200	5	***4***	***4*** $adj = 0.80$, $er = 0.08$
Flea beetles	74	6	***3***	***3*** $adj = 1$, $er = 0.00$
AIS	202	3	***2***	***3*** $adj = 0.76$, $er = 0.11$
Wisconsin	569	3	***2***	***4*** $adj = 0.62$, $er = 0.21$
Yeast	626	3	***2***	***6*** $adj = 0.48$, $er = 0.23$

adj: adjusted Rand index (1 perfect classification), *er*: proportion of misclassified observations

Capturing non-Gaussian data clusters I

Problems with normal mixtures in model-based-clustering:

- If data clusters are **non-Gaussian**:

⇒ **number of estimated normal components** $\neq$ the **number of data clusters**, since: several normal components have to be **merged** to solve this misspecification.

- Recent research: non-Gaussian component densities such as **skew-normal** or **skew-t** distributions.

However:

- It may be difficult to decide which parametric distribution is **appropriate** to characterize a data cluster.

⇒ **"Mixture of mixtures"** (GMW, Sylvia FS, Bettina G., 2017):

  - models the non-Gaussian **cluster distributions** themselves as **Gaussian mixtures**.
  - Gaussian mixtures can approximate a wide class of probability distributions!

Capturing non-Gaussian data clusters II

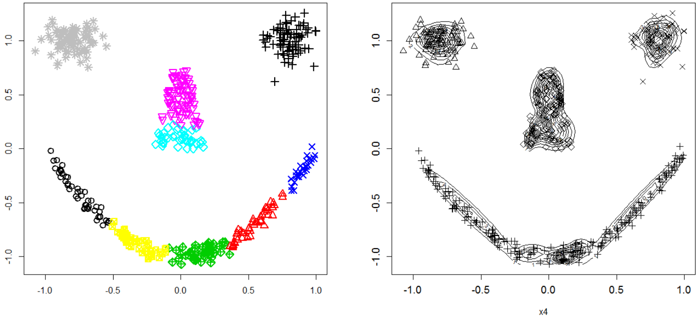


Figure: Smiley's data (Leisch, 2004)

Idea and strategy: Mixture of mixtures

- **Idea**: Specification of a mixture model where
  $\Rightarrow$ each cluster distribution is itself a mixture of normal subcomponents:

$$
\begin{aligned}
p(\mathbf{y}_i|\boldsymbol{\Theta}) &= \sum_{k=1}^{K} \eta_k p_k(\mathbf{y}_i|\boldsymbol{\theta}_k), \\
p_k(\mathbf{y}_i|\boldsymbol{\theta}_k) &= \sum_{l=1}^{L} w_{kl} f_{\mathcal{N}}(\mathbf{y}_i|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}).
\end{aligned}
$$

$\Rightarrow$ Highly over-parameterized mixture model!

- We specify **informative priors** for the parameters of the mixture of mixtures model in order to be able to
  - to estimate the number of data clusters,
  - to achieve a good approximation of the cluster density through the cluster mixture distribution.

Number of clusters

- Our strategy for $\boldsymbol{\eta} \sim Dir_K(e_0)$:
  - **"sparse finite mixture"**: specify an **overfitting** mixture of cluster distributions and define a sparse weight prior on the cluster weights.
- Our strategy for $\mathbf{w}_k \sim Dir_L(d_0)$:
  - We use the normal mixture to approximate an arbitrary cluster distribution in a semiparametric way.
    $\Rightarrow$ We are not interested in estimating the "true" number of subcomponents L.
  - We specify the same fixed **redundant number** of normal subcomponents L for each cluster.
  - We specify a **redundant prior** for the subcomponent weights in order to fill all subcomponents during MCMC sampling by choosing d_0 large, $d_0 > d/2$.

$\Rightarrow$ "Automatic" tool to get a good **density fit** of the cluster distribution!

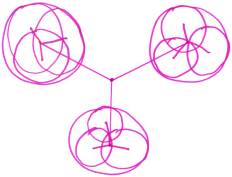
Modelling non-Gaussian cluster distributions I

- **Non-identifiability problem**: It cannot be decided by the likelihood which subcomponents build which cluster.
- **Strategy**: Specification of highly **informative priors** for the subcomponent parameters such that
  - within a cluster subcomponents have **strongly overlapping and flat** densities.

  $\Rightarrow$ **large** subcomponent covariance matrices.
  $\Rightarrow$ **strong shrinkage** of the subcomponent means toward the cluster mean.
- **Idea**: We specify the prior parameters through **variance-covariance decomposition** of the data.

Modelling non-Gaussian cluster distributions II

Variance-covariance decomposition of a mixture of mixtures:



$$
\begin{aligned}
Cov(\mathbf{Y}) &= \underbrace{\phi_B Cov(\mathbf{Y})}_{\text{by cluster means}} + \underbrace{(1-\phi_B)Cov(\mathbf{Y})}_{\text{within the clusters}} \\
&= \underbrace{\phi_B Cov(\mathbf{Y})}_{\text{by cluster means}} + \underbrace{(1-\phi_B)\phi_W Cov(\mathbf{Y})}_{\text{by the subcomponent means}} + \underbrace{(1-\phi_B)(1-\phi_W)Cov(\mathbf{Y})}_{\text{within the subcomponents}} \\
& \qquad\qquad\qquad\qquad\qquad\quad \Downarrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad \Downarrow \\
& \qquad\qquad\qquad\qquad\quad Cov(\boldsymbol{\mu}_{kl}) \qquad\qquad\qquad\qquad\qquad\quad \boldsymbol{\Sigma}_{kl}
\end{aligned}
$$

Modelling non-Gaussian cluster distributions III

To define the prior parameters for subcomponent means and covariance matrices:

1. Choose ϕ_W and ϕ_B, e.g. $\phi_B = 0.5, \phi_W = 0.1$.
2. Define the prior parameters in order that a priori

$$\begin{aligned} E(\mathbf{\Sigma}_{kl}) &= (1-\phi_W)(1-\phi_B)\mathbf{S}_y, \\ Cov(\boldsymbol{\mu}_{kl}) &= \phi_W(1-\phi_B)\mathbf{S}_y. \end{aligned}$$

Model identification

To **solve the label switching problem**:

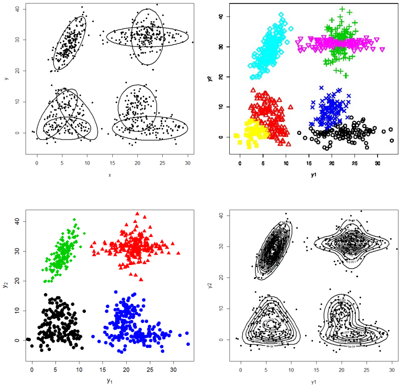
- On the **cluster level**:
  - **Cluster the draws in point process representation** to obtain a unique labeling.
  - Note: we clustered only a **functional** of the subcomponent means of a cluster in the point process representation.
- On the **subcomponent level**:
  - Actually: A lot of label switching occurs due the the strong overlapping subcomponent distributions, but it does not matter!

  $\Rightarrow$ It is not necessary to identify single subcomponents:
  we are only interested in the **whole cluster mixture distribution** of the cluster.

  $\Rightarrow$ we can **ignore** the label switching problem on this level!

Example: Simulated data I

- Data from a mixture of 8 bivariate normal distributions (left).
- Clustering using a sparse finite mixture (middle) compared to using a sparse finite mixture-ofmixtures model (right).



Example: Simulated data II

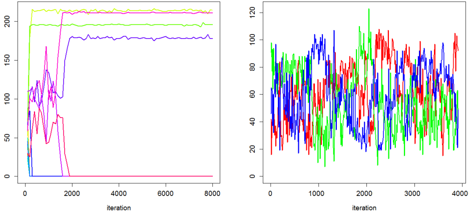


Figure: MCMC run with $K = 15$ and $L = 3$. Trace plot of number of observations allocated to different clusters (left) and trace plot of the subcomponents forming the L-shaped cluster.

Revisiting the benchmark data sets

Data set	K^{true}	$\boldsymbol{K_+}$ for `SparseMix` $L=1$	$\hat{K}_+$ **for `SparseMixMix`** $(K=10, e_0=0.001)$ $L=4$
AIS	2	***3*** $adj=0.76$, $er=0.11$	***2*** $adj=0.81$, $er=0.05$
Wisconsin	2	***4*** $adj=0.62$, $er=0.21$	***2*** $adj=0.82$, $er=0.05$
Yeast	2	***6*** $adj=0.48$, $er=0.23$	***2*** $adj=0.81$, $er=0.05$

adj: adjusted Rand index (1 perfect classification), *er*: proportion of misclassified observations.

$K^{true}=2$ ***recovered for all data sets***

Mixture of two *SAL* distributions (Franczak et al., 2012)

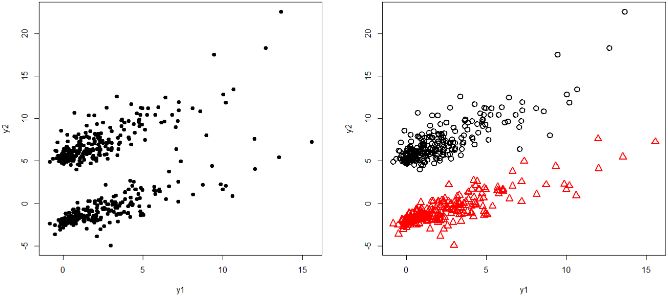


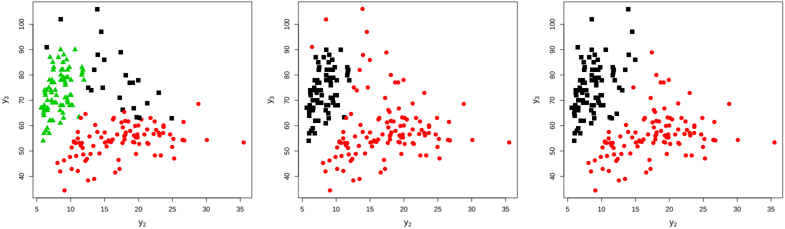
Figure: Samples from a mixture of two *SAL* distributions (left), the estimated clusters for $K = 10$, $L = 5$, $\phi_B = 0.4$, $\phi_W = 0.2$, with fixed hyperparameters $\mathbf{C}_{0k}$ and λ_{kl} (right-hand side).

Pitfalls of post-processing merging

AIS data sets, variables "X.Bfat" and "LBM".

Solutions:

- `Mclust` ($K = 3$), Fraley et al. (2012) (left),
- `combiClust` ($K = 2$), Baudry et al. (2010) (middle),
- **sparse finite mixture** ($K_+ = 2$), $K = 10, L = 4$ (right).



Flow cytometric data I

1. Flow cytometric data set DLBCL
   - $N = 7932$, $r = 3$, known labeling.
   - Sparse finite mixture of mixtures ($K = 30, L = 15, e_0 = 0.001$) yields $K_+ = 4$, error rate=0.03.

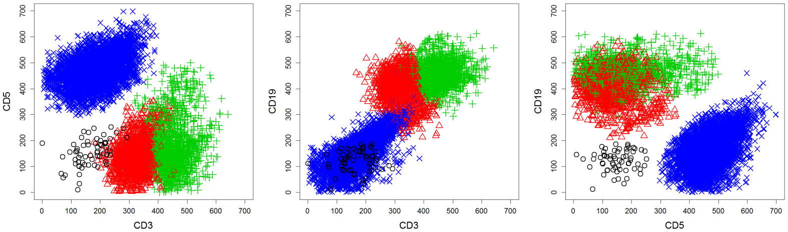


Figure: Flow cytometry data set DLBCL. Scatterplot of the clustering results.

Flow cytometric data II

2. Flow cytometric data set GvHD
   - $N = 12442$, $r = 6$, unknown labeling.
   - Sparse finite mixture of mixtures ($K = 30, L = 15, e_0 = 0.001$) yields $K_+ = 8$.

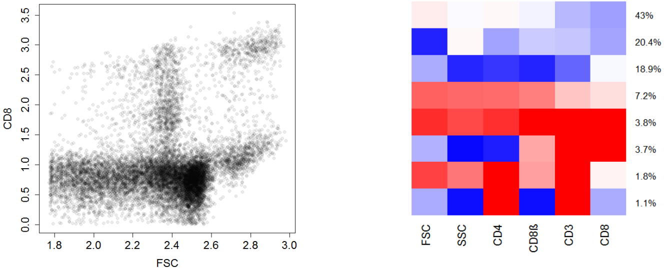


Figure: Flow cytometric data set GvHD. Scatter plot of two variables ("FSC", "CD8") (left-hand side), and heatmap of the clustering results by fitting a sparse hierarchical mixture of mixtures model (right-hand side).

Summary

Sparse finite mixtures

- Estimates the number of data clusters through the number of **non-empty components** (random a priori).

$\Rightarrow$ In an overfitting mixture specification of a **Dirichlet prior with e_0 very small.**

Mixtures of mixtures

- Flexible modelling of unknown cluster distributions.
- Prior specification crucial: strongly overlapping subcomponent densities.

Extensions

- *Sparse finite mixtures*: Extension to other **non-Gaussian component densities**, e.g. mixtures of t-distributions, Poisson distributions, topic model? ...
- *Mixtures of mixture*s: for **latent class models**: overcome the local independence assumption?
- *Computational issues*: for large N, p: MCMC tends to get stuck

  $\Rightarrow$ Work in progress: develop another sampling scheme to overcome this issue!

References I

Baudry, J.-P., A. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combing mixture components for clustering. *Journal of Computational and Graphical Statistics 19*(2), 332–353.

Fraley, C., A. Raftery, T. Murphy, and L. Scrucca (2012). Technical report 597. *Department of Statistics, University of Washington (http://www. stat. washington. edu/mclust/)*.

Franczak, B. C., R. P. Browne, and P. D. McNicholas (2012). Mixtures of shifted asymmetric Laplace distributions. *eprint arXiv:1207.1727*.

Frühwirth-Schnatter, S. (2012). Flexible econometric modelling based on sparse finite mixtures. Presentation at the ISBA 2012, 11th World Meeting of the International Society of Bayesian Analysis.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing 26*, 303–324.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society B 73*(5), 689–710.