JOHANNES KEPLER
UNIVERSITY LINZ | JKU

Flexible and Sparse Bayesian Model-Based Clustering

Bettina Grün

Joint work with
S. Frühwirth-Schnatter and G. Malsiner-Walli

Bozen 2014

This research is supported by the Austrian Science Fund (FWF): V170-N18.

Finite mixture models

- We will focus on finite mixtures of Gaussians where the density is given by

$$h(\boldsymbol{y}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \eta_k f_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

  where $f_{\mathcal{N}}()$ is the multivariate normal distribution and

$$\eta_k \geq 0, \qquad \sum_{k=1}^{K} \eta_k = 1.$$

Open issues in model-based clustering

- Selecting a suitable number of components K.
- Identifying cluster-relevant variables.
- Dealing with non-normal cluster shapes.

$\Rightarrow$ We investigate how to resolve these issues in a Bayesian estimation context.

Bayesian parameter estimation: Motivation

- Prior information can be included in the model fitting process.
- Smoothing and regularization effect on the mixture likelihood function (Fraley and Raftery, 2007).
- Parameter uncertainty can be easily assessed using the whole posterior distribution.
- No reliance on asymptotic normality allowing for valid inference in cases where regularity conditions are violated, e.g., small data sets and mixtures with small component weights.
- The posterior distribution for N iid observations from the mixture model is given by

$$p(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \propto p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N | \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

  where $p(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the prior distribution.

Prior choice: General considerations

We will only focus on priors with the following characteristics:

- A-priori the different sets of parameters are independent and also the parameters between components are independent:

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(\boldsymbol{\eta}) p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) \\ &= p(\boldsymbol{\eta}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{k=1}^{K} p(\boldsymbol{\Sigma}_k) \end{aligned}$$

- Symmetric priors, i.e., they are invariant to relabeling of the components.
- Proper priors to avoid improper posteriors.
- Conditional conjugate priors to allow for Gibbs sampling after data augmentation (if possible).
- Consider the use of hyperpriors to reduce sensitivity of a specific choice of the prior parameters.

Prior choices for sparse modeling

We will investigate the choice of

- **Priors on the weights:**
  In particular for the case of overfitting mixtures, where the likelihood is problematic.
- **Priors on the component means:**
  Assuming the presence of cluster-irrelevant variables we investigate priors which allow to distinguish between cluster-relevant and cluster-irrelevant variables.

Prior on the weights

- Conjugate prior: Dirichlet prior

$$\boldsymbol{\eta} \sim \mathcal{D}(e_1, \ldots, e_K)$$

- The exchangeable Dirichlet prior is assumed with

$$e_k \equiv e_0, \quad k = 1, \ldots, K.$$

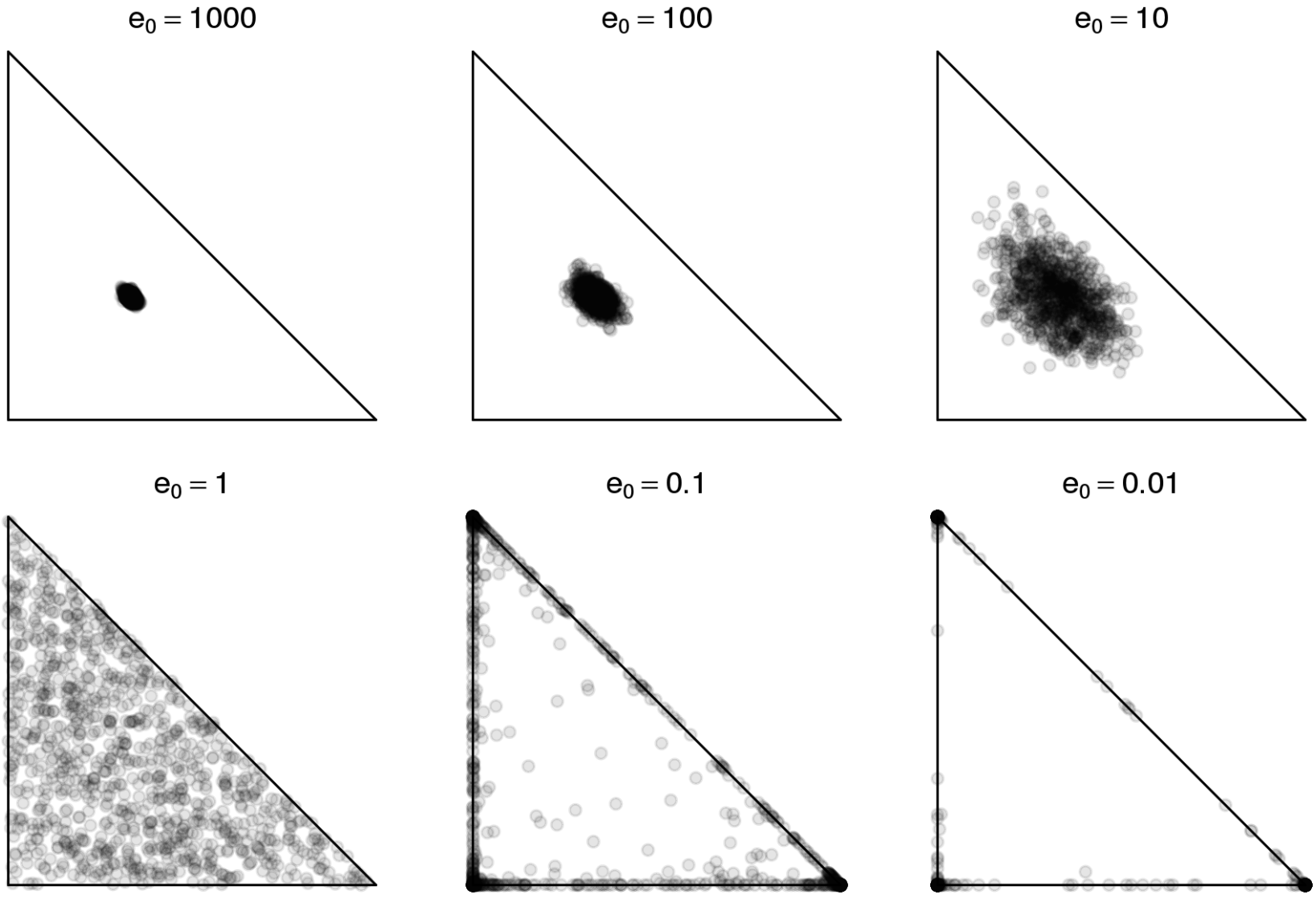
  This implies:

  - The prior expectation is

$$\mathbb{E}[\eta_k | e_0] = \frac{1}{K}$$

    regardless of the specific value of e_0.
  - The prior variance depends on the size of e_0.

Prior on the weights / 2



Prior on the weights / 3

- Gibbs sampling step:
  - Draw $\boldsymbol{\eta}$ from the following Dirichlet distribution

$$\boldsymbol{\eta}|\boldsymbol{S} \sim \mathcal{D}(n_1 + e_0, \ldots, n_K + e_0),$$

    where n_k are the number of observations assigned to component k, i.e., the number of observations, where $S_i = k$.
- The mean of this conditional posterior is

$$\mathbb{E}[\eta_k|\boldsymbol{S}, e_0] = \frac{n_k + e_0}{N + Ke_0}.$$

- The choice of e_0 is rather uncontroversial if the number of components is assumed to be known.
- Under model uncertainty, the choice of e_0 is crucial.
  $\Rightarrow$ A suitable value needs to be selected depending on the strategy used to determine the true number of components K^{true}.

Dirichlet prior for overfitting mixtures

- Overfitting mixtures are mixtures where the fitted number of components K exceeds the true number of components K^{true}.
- The likelihood reflects the two possible ways of dealing with the superfluous components:
  - **Empty components:**
    - η_k is shrunken towards 0.
    - The component-specific parameters are identified only through their prior.
  - **Duplicated components:**
    - The difference of the component-specific parameters are shrunken towards 0.
    - Only the sum of the corresponding component weights is identified.
- The likelihood is multimodal, because it mixes these two unidentifiability modes.

Dirichlet prior for overfitting mixtures / 2

- Recent research by Rousseau and Mengersen (2011) indicates that the value of e_0 strongly influences the posterior density for overfitting mixtures.
- They show the following asymptotic result:
  - If $e_0 < d/2$, then asymptotically the posterior density concentrates over regions where $K - K^{\text{true}}$ groups are left empty.
  - If $e_0 > d/2$, then asymptotically the posterior density concentrates over regions with duplicated components.

  d denotes the dimension of the component-specific parameters.
- Consequence for empirical applications:
  - Decide through the Dirichlet prior whether you prefer empty groups or duplicated components for overfitting mixtures.
  - This decision helps to interpret the draws from the posterior distribution of an overfitting mixture.

Identifying the number of components

We distinguish the following model selection approaches:

- Marginal likelihoods and Reversible Jump MCMC (RJMCMC):
  - Use overfitting mixtures with duplicated components (e_0 large).
    $\Rightarrow$ Avoids overestimating K^{true}.
- Non-empty components:
  Determine the number of non-empty components for each sweep m of the sampler

$$K_0^{(m)} = K - \sum_{k=1}^{K} \boldsymbol{I}\{n_k^{(m)} = 0\}$$

  and use the most frequently visited value as estimate for K^{true}.
  - Use overfitting mixtures with empty components (e_0 small).

See Nobile (2004).

Prior on the component means

- Conjugate prior: multivariate normal distribution.

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0).$$

- Gibbs sampling step:
  - Draw $\boldsymbol{\mu}_k$ from the following multivariate normal distribution:

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{b}_k, \boldsymbol{B}_k),$$

where

$$\begin{aligned}\boldsymbol{B}_k &= \left(\boldsymbol{B}_0^{-1} + n_k \Sigma_k^{-1}\right)^{-1}, \\ \boldsymbol{b}_k &= \boldsymbol{B}_k \left(\boldsymbol{B}_0^{-1}\boldsymbol{b}_0 + n_k \Sigma_k^{-1} \bar{\boldsymbol{y}}_k\right),\end{aligned}$$

where $\bar{\boldsymbol{y}}_k$ is the sample mean in group k.

  - Proper priors pull the component means toward prior mean.
  - The amount is governed by the prior variance.

Identifying cluster-irrelevant variables

- Inclusion of cluster-irrelevant variables can:
  - Mask the cluster structure.
  - Reduce the accuracy of the parameter estimates.
- Proposed approaches:
  - Variable selection using stepwise procedures or stochastic model search (Raftery and Dean, 2006).
  - Shrinking of component means towards a common mean (Yau and Holmes, 2011; Frühwirth-Schnatter, 2011).

Shrinkage priors

- We consider shrinkage priors which can be represented as a scale mixture of normals.
- Assuming $y \sim \mathcal{N}(\mu, \sigma^2)$, the prior distribution for the location parameter μ is specified as

$$\pi(\mu) = \int f_{\mathcal{N}}(\mu|0, \lambda) d\pi(\lambda),$$

  where $\pi(\lambda)$ is a mixing distribution.
- This prior can also be expressed in hierarchical form as

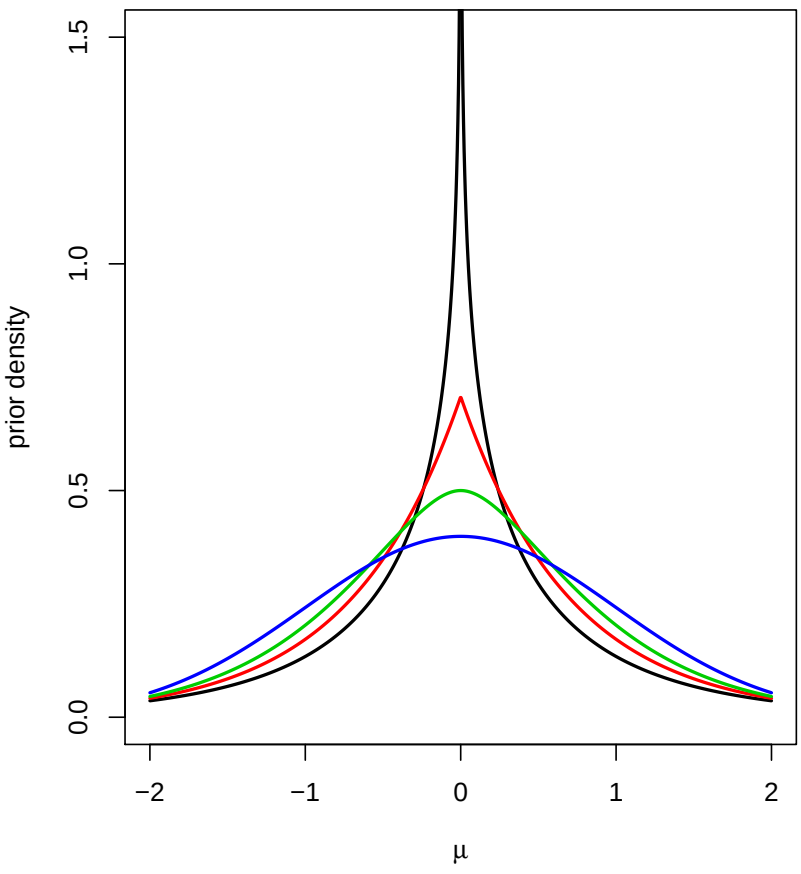
$$\begin{aligned} \mu &\sim \mathcal{N}(0, \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

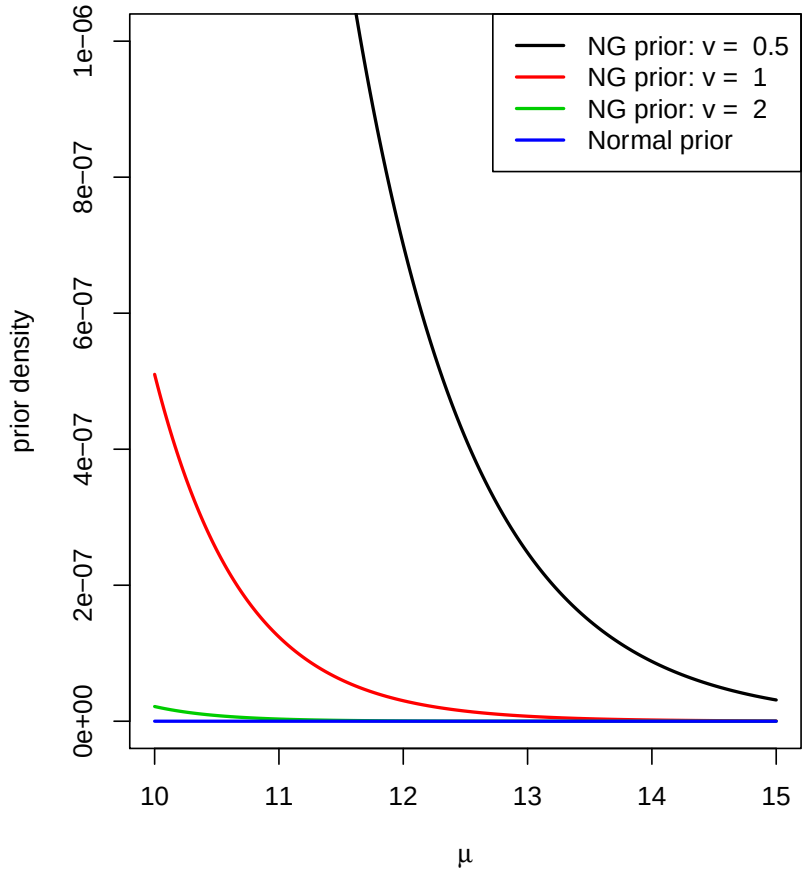
  $\Rightarrow$ Easy to implement for MCMC sampling.

Shrinkage priors / 2

- Some examples:
  - If $\pi(\lambda) \sim \mathcal{G}(1, \nu_2)$, the marginal distribution $\pi(\mu)$ is the double-exponential prior.
    $\Rightarrow$ Lasso (Yau and Holmes, 2011)
  - If $\pi(\lambda) \sim \mathcal{G}(\nu_1, \nu_2)$, the marginal distribution $\pi(\mu)$ is called the **normal gamma prior** (Griffin and Brown, 2010).
- If $\nu_1 = \nu_2$:
  - $\mathbb{E}(\lambda_j) = 1$.
    $\Rightarrow$ The expected variance of μ_{kj} is as a-priori specified.
  - $\mathbb{V}(\lambda_j) = 1/\nu_1$.
    $\Rightarrow$ Choose $\nu_1 < 1$.
- The normal gamma prior puts more weight around zero and has heavier tails than the double-exponential distribution.

Shrinkage priors / 3





Model identification

- The likelihood is invariant with respect to a permutation of the components.
- The use of symmetric priors implies that this invariance also holds for the posterior.
- Component-specific inference is impossible based on the MCMC output due to **label switching** (Redner and Walker, 1984).
- Several strategies have been proposed to determine an identified model (for an overview see Jasra et al., 2005).
- We suggest to cluster (part of) the component-specific parameters of the MCMC draws in the point process representation, e.g., using k-means, to obtain a unique labeling and to discard draws where this is not achieved.

Modeling strategy

- Use a large value for K and a small e_0 in order to allow for automatic selection of a suitable number of clusters using the most frequent number of non-empty clusters during MCMC sampling.
  - e_0 can be either set very small and fixed.
  - Alternatively, we also investigate the use of a hyperprior

$$e_0 \sim \mathcal{G}(a, a \cdot K)$$

    with $a = 10$.
- Use a normal gamma prior for the component means with $\nu_1 = \nu_2 = 0.5 < 1$.
  - If component means are pulled together with a shrinkage prior, a hyperprior needs to be specified for the prior mean $\boldsymbol{b}_0$.
- Note that for the normal gamma prior e_0 needs to be selected smaller than for the standard prior and fixing it gives better results.

Example: Simulation

- Simple setup: 2 cluster-generating variables & 2 noisy variables with 4 components and mean values:

$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4) = \begin{pmatrix} -2 & -2 & 2 & 2 \\ -2 & 2 & 2 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- $\boldsymbol{\eta} = (0.25, 0.25, 0.25, 0.25)$
- $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \mathit{diag}(1, 1, 1, 1)$
- $N = 1000$, 10 data sets, averaged results.
- Priors: $\nu_1 = 0.5$, $e_0 \sim \mathcal{G}(10, 10 \cdot 15)$.
- MCMC: 10000 draws after a burn-in of 2000 draws.

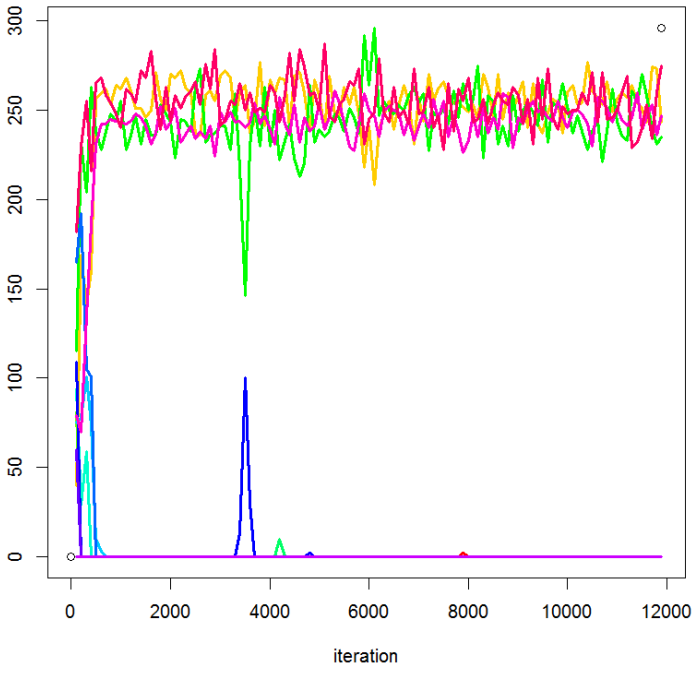
Example: Simulation / 2

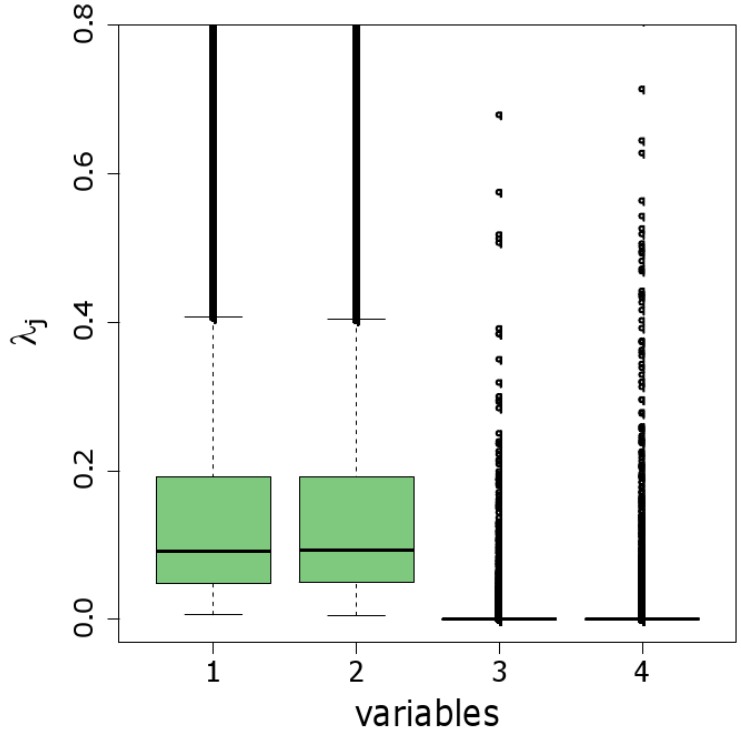
Results for different K under the standard (Sta) and normal gamma prior (NG), averaged over 10 data sets.

prior	K	$\hat{e}_0$	e_0 fixed	$\hat{\boldsymbol{K}}_0$	M_0	$M_{0,\rho}$	MCR	MSE_μ
Sta	4	0.27		**4**	10000	0	0.049	0.184
	15	0.05		**4**	9709	0	0.049	0.184
	30	0.03		**4**	9786	0	0.048	0.185
Ng	4		0.01	**4**	10000	0	0.048	0.155
	15		0.01	**4**	7620	0	0.048	0.156
	30		0.01	**4**(9)	5294	0	0.048	0.159
	30		0.001	**4**	9224	0	0.048	0.154

Example: Simulation / 3

1 data set, $K = 15$, standard prior, traces of the number of observations allocated to the different components.





Mixtures of Gaussian mixtures

- To account for non-normal shapes of the cluster distributions in the finite mixture model

$$h(\boldsymbol{y}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{y}|\boldsymbol{\theta}_k),$$

each cluster distribution can be semi-parametrically estimated using a finite mixture of Gaussians

$$f_k(\boldsymbol{y}|\boldsymbol{\theta}_k) = \sum_{l=1}^{L_k} w_{kl} f_{\mathcal{N}}(\boldsymbol{y}|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}).$$

Mixtures of Gaussian mixtures / 2

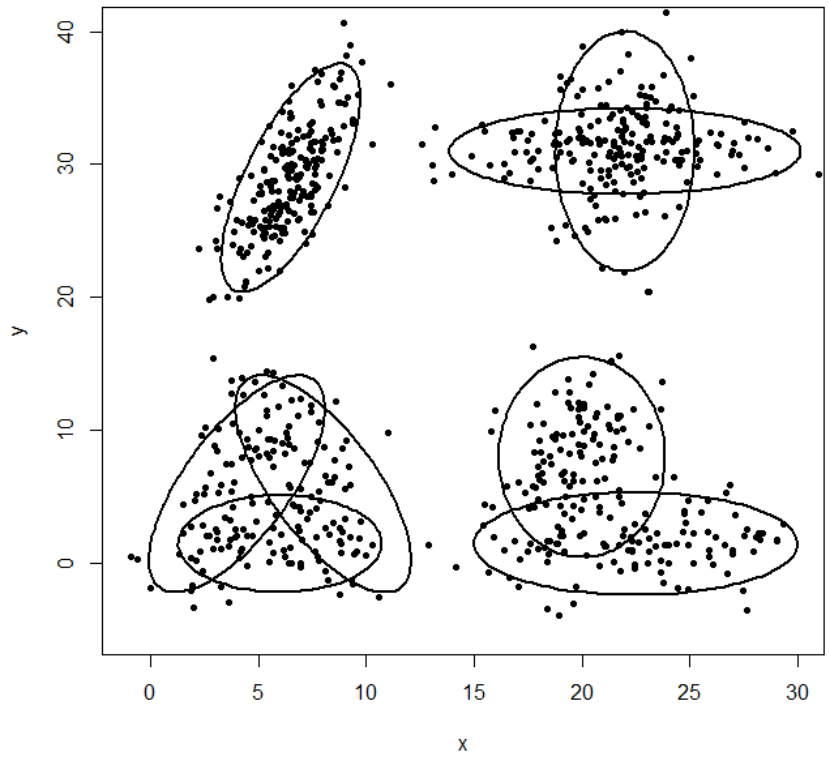
- Using the finite mixture of Gaussians model for density estimation implies:
  - The number of subcomponents L_k is less crucial and only needs to be sufficiently high. So we assume $L_k \equiv L$ for all k.
  - Identification of the subcomponent-specific parameters is not necessary.
- Using the likelihood only the mixture of Gaussian mixtures model is not identifiable. Several post-processing methods have been proposed to merge components into clusters (Baudry et al., 2010; Hennig, 2010).

Mixtures of Gaussian mixtures: Prior choice

- We use the prior specification in the Bayesian estimation to allow for automatic distinction between subcomponents from the same or different clusters.
- We aim at finding density clusters of convex shapes with gaps between the cluster densities.
- The priors on the cluster level:
  - Sparse prior for the weights to allow for automatic selection of number of clusters.
  - No shrinkage on cluster means.
- The priors on the subcomponent level:
  - Prior on the weights which ensures that all subcomponents are filled.
  - Shrinkage of subcomponent means toward the cluster mean.
  - The prior on the variance-covariance matrix tends to increase their volumes.

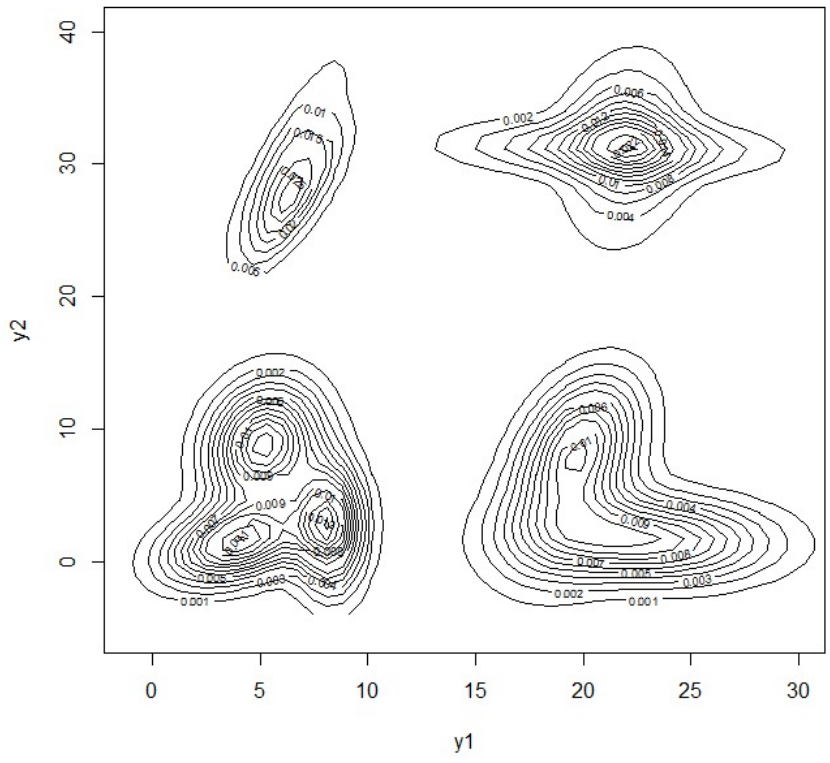
Example: Simulation

$K_{\text{true}} = 4$ – true density:



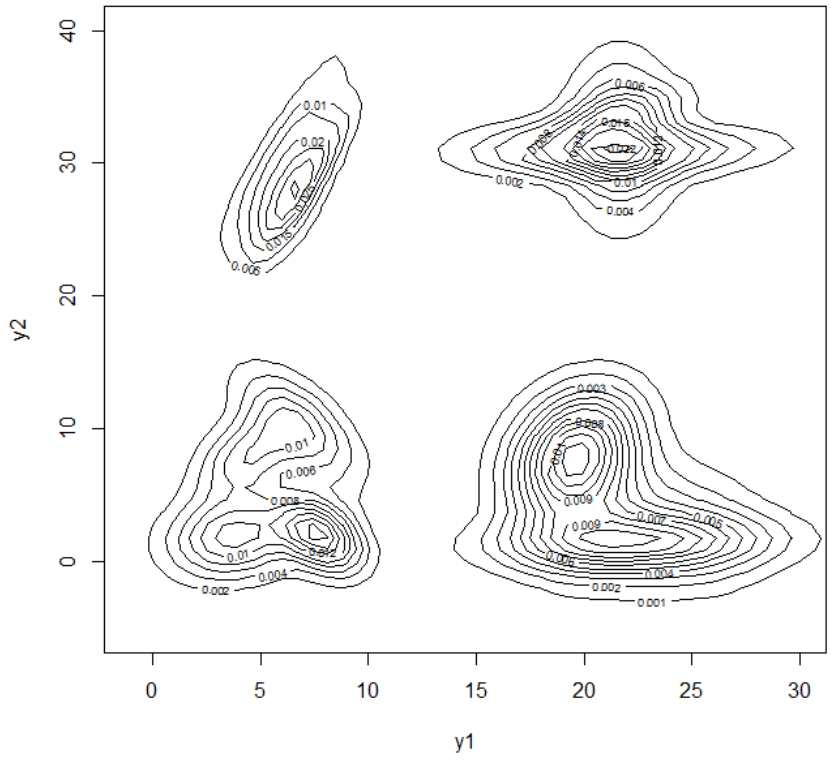
Example: Simulation / 2

$K_{\max} = 10, L = 3$ – fitted density; $\hat{K}_0 = 4$; $K_{\text{true}} = 4$:



Example: Simulation / 3

$K_{\max} = 10, L = 4$ – fitted density; $\hat{K}_0 = 4$; $K_{\text{true}} = 4$:



Summary & future work

- Summary:
  - Bayesian estimation of finite mixture models can help to deal with unresolved issues in model-based clustering.
  - Suitable prior choice helps to identify:
    - Number of components.
    - Cluster-relevant and cluster-irrelevant variables.
    - Subcomponents and clusters in a mixture of mixtures.
- Future work:
  - Priors to induce parsimonious mixture models with respect to the variance-covariance matrices.
  - Further variants are possible when relaxing some of the general considerations such as the choice of symmetric priors.

References

J. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. **Journal of Computational and Graphical Statistics**, 19:332–353, 2010.

C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. **Journal of Classification**, 24(2): 155–181, Sept. 2007.

S. Frühwirth-Schnatter. Label switching under model uncertainty. In K. Mengerson, C. Robert, and D. Titterington, editors, **Mixtures: Estimation and Application**, pages 213–239. Wiley, 2011.

J. E. Griffin and P. J. Brown. Inference with normal-gamma prior distributions in regression problems. **Bayesian Analysis**, 5(1):171–188, 2010.

C. Hennig. Methods for merging Gaussian mixture components. **Advances in Data Analysis and Classification**, 4:3–34, 2010.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. **Statistical Science**, 20(1):50–67, 2005.

References / 2

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite Gaussian mixtures. **Statistics and Computing**, in press.

A. Nobile. On the posterior distribution of the number of components in a finite mixture. **The Annals of Statistics**, 32:2044–2073, 2004.

A. E. Raftery and N. Dean. Variable selection for model-based clustering. **Journal of the American Statistical Association**, 101(473):168–178, 2006.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. **SIAM Review**, 26(2):195–239, Apr. 1984.

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

C. Yau and C. Holmes. Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. **Bayesian Analysis**, 6 (2):329–352, 2011.