Finite Mixture and Markov Switching Models

Sylvia Frühwirth-Schnatter

Western Swiss Doctoral School in Statistics and Probability, February 2020

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Part I

Finite Mixture Models and Model-based Clustering

Outline

Part I: Finite Mixture Models and Model-based Clustering

- Finite mixture distributions
- Unsupervised Clustering
- Bayesian Approach toward Estimation
- Mixture-of-experts models
- Overfitting mixtures
- Sparse finite mixtures in action
- Model selection for finite mixtures

Alternative models for prior allocation

- A second look at the standard finite mixture model:

$$\begin{aligned}
&\mathrm{Pr}(y_i = j) = \eta_1 \mathrm{Pr}(y_i = j|S_i = 1, \boldsymbol{\theta}_1) + \eta_2 \mathrm{Pr}(y_i = j|S_i = 2, \boldsymbol{\theta}_2),\\
&\mathrm{Pr}(S_i = 1) = \eta_1,\\
&\mathrm{Pr}(S_i = 2) = \eta_2.
\end{aligned}$$

- ***The prior probability of belonging to class k is the same for all persons.***
- Could this be true?

Costumers are not alike!

- Costumer are heterogeneous with respect to brand and price of beverages (see e.g. [Frühwirth-Schnatter et al., 2004])
- Some of them are price sensitive, some of them are brand sensitive

- Do they have the same prior probability to belong to a group?

Mixtures-of-experts models

- The prior probability of belonging to a certain group is not the same for all persons i, but depends on characteristics $\mathbf{x}_i$ of the person, e.g. age and income.

Mixtures-of-experts for 2 classes:

$$\begin{aligned}
\Pr(y_i = j) &= \eta_{i1}\Pr(y_i = j|S_i = 1, \boldsymbol{\theta}_1) + \eta_{i2}\Pr(y_i = j|S_i = 2, \boldsymbol{\theta}_2),\\
\Pr(S_i = 1|\mathbf{x}_i) &= \eta_{i1} = F(\mathbf{x}_i\beta),\\
\Pr(S_i = 2|\mathbf{x}_i) &= \eta_{i2} = 1 - \eta_{i1},
\end{aligned}$$

where $F(z)$ is the cdf of the logistic distribution.

Mixtures-of-experts models

Extension to more than two classes $K > 2$:

$$\Pr(S_i = k|\mathbf{x}_i) = \eta_{ik} = F(\mathbf{x}_i\beta_k), \qquad k = 2, \ldots, K,$$

where $F(\lambda_{ik})$ is the link function of a multinomial logistic model, i.e. $F(\lambda_{ik}) = \exp(\lambda_{ik})/(1 + \sum_{l=2}^{K} \exp(\lambda_{il}))$.

- Many applications, e.g.
  - Speech recognition [Peng et al., 1996]
  - Modeling the voting behavior [Gormley and Murphy, 2008]
  - Analyzing labour market data [Frühwirth-Schnatter et al., 2012]
- MCMC: auxiliary mixture sampling [Frühwirth-Schnatter and Frühwirth, 2010], Polya Gamma sampler [Polson et al., 2013].
- **Review**: [Gormley and Frühwirth-Schnatter, 2019]

Example: Effect of plant closure

[Frühwirth-Schnatter et al., 2018]:

- Analysing plant closure effects for a **Cohort study:** male workers (aged between 35 and 55) employed in 1982–1988
- Individual quarterly data for 10 year after plant closure
- Panel of $N = 5,841$ male workers with $T_i = 40$ quarterly data on labour market states (employed/sick/out of labour force/retired)
- Research question:
  - What is the **effect of plant closure** on the employment career?
  - Is there a **difference** between workers facing plant closure and those who did not?
- Time-varying mixture-of-experts Markov chain clustering
- Economic interpretability led us to choose **5 clusters**

Time-varying Markov chain clustering

- Time-inhomogeneity present the plant closure data.
- **Generalized transition matrices:** inhomogeneous transition matrix depending on a history $\mathcal{H}_{it}$ [Frühwirth-Schnatter, 2011b]:

$$\Pr(y_{it} = j|\mathcal{H}_{it}, S_i = k),$$

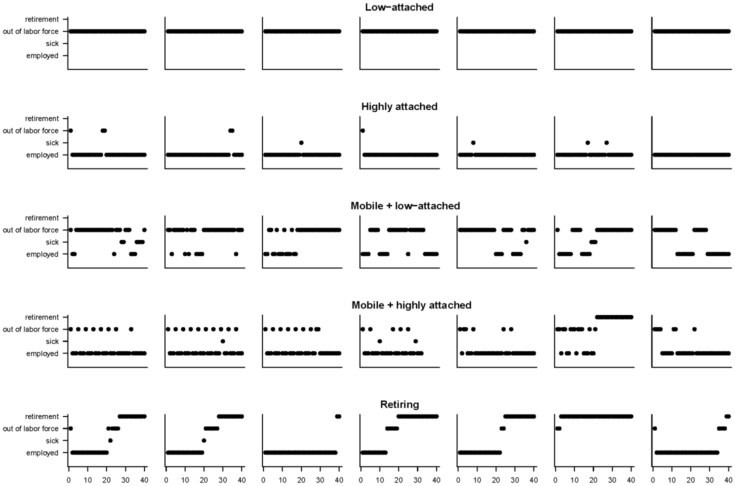
  where $\mathcal{H}_{it} = \{y_{i,t-1}, \mathbf{x}_{it}\}$.
- Typically $\mathbf{x}_{it}$ is some discrete covariate, e.g. the year after plant closure:

$$\boldsymbol{\vartheta}_k = (\boldsymbol{\pi}_k, \boldsymbol{\xi}_{k,1}, \boldsymbol{\xi}_{k,2}, \dots, \boldsymbol{\xi}_{k,10})$$

- We could include addition information (age group, . . .) in $\mathbf{x}_{it}$

Model-based clustering of plant closure data

Employment profiles of cluster members ranked 10th, 25th, 50th, 70th, 100th, 200th, 350th

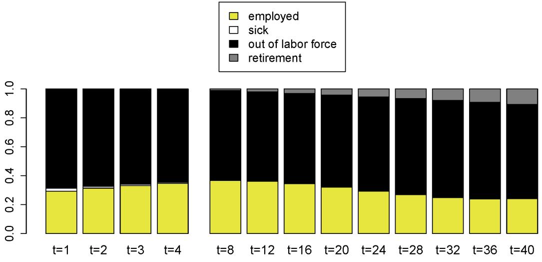


Analysing dynamic effects

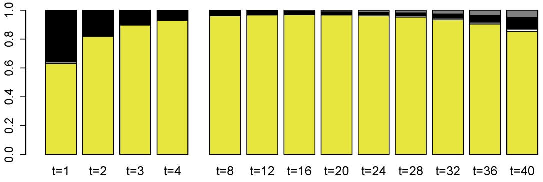
State distribution $\pi_{k,t}$, where

$$\pi_{k,t} = \pi_k \xi_{k,1\to t}, \qquad \xi_{k,1\to t} = \xi_{k,1\to(t-1)}\xi_{ky}; \quad \xi_{k,1\to 2} := \xi_{k1}.$$

over distance $t = 4(y-1) + q$ from plant closure (quarters), for cluster k:



Cluster 1: low attached



Cluster 2: highly attached

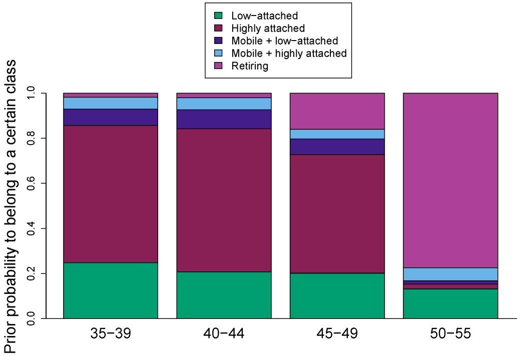
Mixtures-of-experts approach

- **Mixture-of-experts model:** multinomial logit model

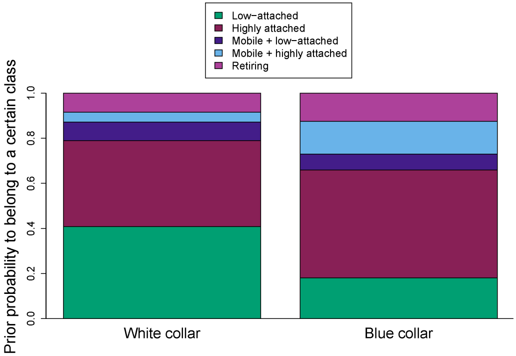
$$\Pr(S_i = k|\mathbf{x}_i) = F(\mathbf{x}_i\beta_k)$$

- Covariates $\mathbf{x}_i$ based on **individual characteristics**:
  - age at the time of plant closure (five age groups: 35-39, 40-44, 45-49, 50-55)
  - levels of experience (low, medium, high)
  - broad occupational status (blue versus white collar)
  - income before plant closure (low, medium, high) based on the tertiles of the general income distribution at time of plant closure
- . . . and on **firm characteristics**:
  - three categories of firm size (1-10, 11-100, and more than 100 employees)
  - four broad economic sectors (service, industry, seasonal business outside of hotel and construction, unknown)

Prior probabilities to belong to a cluster



Impact of age



Impact of white versus blue collar

Outline

Part I: Finite Mixture Models and Model-based Clustering

Overfitting mixtures

- **Overfitting** finite mixture distributions,

$$p(\mathbf{y}) = \sum_{k=1}^{K} \eta_k f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k),$$

  where K is larger than the true number of components K_{tr} in the data.
- Likelihood function is highly irregular
- Specify a Dirichlet prior on the weights:

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \sim \mathcal{D}_K(e_0).$$

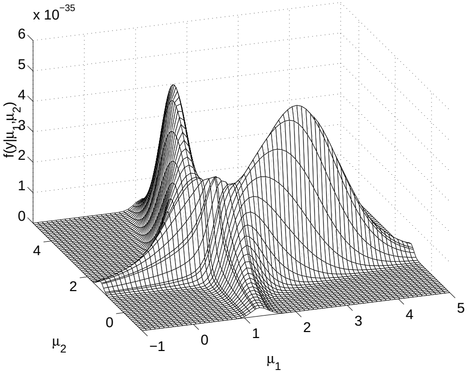
- The hyperparameter e_0 has, again, a regularizing effect on the likelihood function.

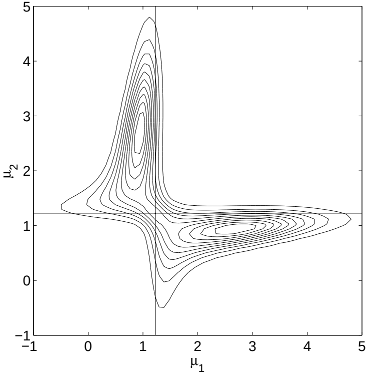
The likelihood for overfitting mixtures

- The likelihood is highly irregular for overfitting mixtures because it reflects two possible ways of dealing with overfitting mixtures with $K > K_{tr}$:
  - Empty components: η_k is shrunken toward 0; $\boldsymbol{\theta}_k$ is identified only through the prior $p(\boldsymbol{\theta}_k)$
  - Duplicated components: $\boldsymbol{\theta}_k - \boldsymbol{\theta}_j$ is shrunken toward 0; only the sum of the components weights $\eta_k + \eta_j$ is identified.
- The likelihood is multimodal, because it mixes these two unidentifiability modes.

Example

Simulated data with $\mu_1 = 1$, $\mu_2 = 1.5$, $\sigma_1^2 = \sigma_2^2 = 1$, $N = 100$; surface and contours of the integrated mixture likelihood $p(\mathbf{y}|\mu_1, \mu_2)$





Prior choices for finite mixtures

- Formulate a prior on the components parameters $\boldsymbol{\theta}_k \sim \mathcal{G}_0$ (typically conditionally conjugate)
- The prior distribution on the weights $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ is a ***Dirichlet distribution*** $\mathcal{D}(e_1, \dots, e_K)$.
- The seemingly non-informative ***uniform prior on the unit simplex***, i.e. the $\mathcal{D}(1, \dots, 1)$-distribution is very informative in unexpected places.
- The hyperparameter $e_1, \dots, e_K$ are informative in particular for overfitting mixtures with $K > K_{tr}$.

The role of the Dirichlet prior

- Consider $\boldsymbol{\eta} \sim \mathcal{D}(e_1, \ldots, e_K)$ where $e_k \equiv e_0$, denoted by $\boldsymbol{\eta} \sim \mathcal{D}_K(e_0)$.
- Let $d = \dim \boldsymbol{\theta}_k$.
- An important paper by [Rousseau and Mengersen, 2011] shows the following asymptotic result:
  - If $e_0 < d/2$, then asymptotically the posterior density concentrates over regions where the total sum of the weights corresponding to $K - K_{tr}$ superfluous groups is 0.
  - If $e_0 > d/2$, then asymptotically the posterior density concentrates over regions with duplicated components.

The role of the Dirichlet prior, ctd.

Consequence for empirical applications [Frühwirth-Schnatter, 2011a]:

- ***decide through the Dirichlet prior*** whether you prefer empty groups or duplicated components for overfitting mixtures;
- making this decision helps to interpret the draws from the posterior distribution of an overfitting mixture;
- making this decision ***facilitates estimating the number of non-empty, non-identical components.***
- to obtain sparsity, e_0 very often has to be much smaller than $d/2$ in finite samples.

Partitions implied by finite mixtures

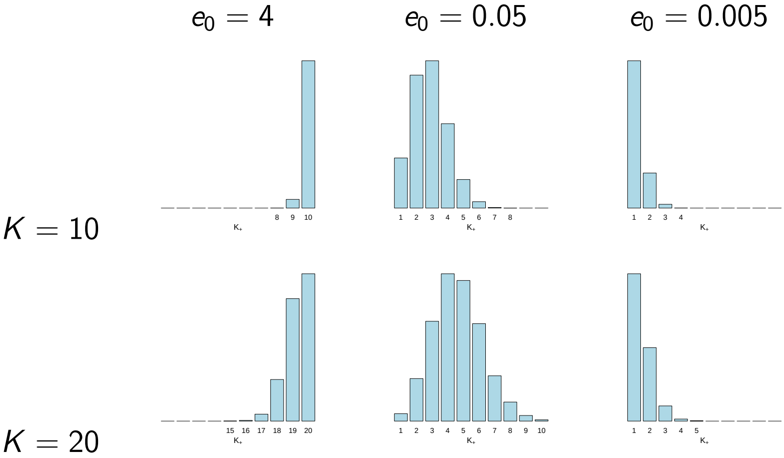
- Clustering arises naturally trough the indicator S_i of the component generating $\mathbf{y}_i|S_i \sim \mathcal{T}(\boldsymbol{\theta}_{S_i})$. $\mathbf{S} = (S_1, \dots, S_N)$ defines a **partition of the data**.
- With N_k is the number of observations allocated to component k, we obtain:

$$N_1, \dots, N_K|\boldsymbol{\eta} \sim \mathrm{MulNom}\left(N; \eta_1, \dots, \eta_K\right). \tag{3}$$

- Depending on the weight distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$, multinomial sampling according to (3) may lead to
  - partitions with **empty groups ($N_k = 0$)**.
  - fewer than K mixture components were used to generate the N data points.
  - the data contain $K_+ < K$ **non-empty clusters**:

$$K_+ = K - \sum_{k=1}^{K} \mathbb{I}\{N_k = 0\}.$$

The **choice of** e_0 of prior $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \sim \mathcal{D}_K(e_0)$ determines whether the number K_+ of clusters in N data points is fixed ($K_+ = K$) or random apriori.



Sparse finite mixtures

- **Overfitting** finite mixture distributions,

$$p(\mathbf{y}) = \sum_{k=1}^{K} \eta_k f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k),$$

  where K is larger than the true number of components K_{tr} in the data.
- Specify a Dirichlet prior on the weights, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \sim \mathcal{D}_K(e_0)$, with
  - e_0 **very small, e.g.** $e_0 = 0.01$;
  - $e_0 \sim \mathcal{G}(a_e, b_e)$ **with** $\mathrm{E}(e_0) = a_e/b_e$ **very small**.
- $\mathbf{y}$ can be univariate or multivariate, continuous, discrete-valued, mixed-type, time series data, outcomes of a regression model, ...

Mixture components versus data clusters

- Sparse finite mixtures make a distinction between K (the order of the mixture distribution) and K_+, the number of clusters in the partition of the data!
- For a sparse finite mixture, the number K_+ of clusters in N data points is random a priori. The prior depends both on e_0 and K, where K is a fixed hyperparameter.
- **Allows to estimate the number K_+ of non-empty groups** aposteriori, given the data, using posterior (MCMC) draws of the indicators $\mathbf{S}$ and the corresponding partitions.
- Is related to Bayesian non-parametric approaches (BNP), where $K = \infty$.

Outline

Part I: Finite Mixture Models and Model-based Clustering

- Finite mixture distributions
- Unsupervised Clustering
- Bayesian Approach toward Estimation
- Mixture-of-experts models
- Overfitting mixtures
- Sparse finite mixtures in action
- Model selection for finite mixtures

Sparse finite mixtures in action

- [Malsiner Walli et al., 2016]: Model-based clustering based on sparse finite Gaussian mixtures
- [Malsiner Walli et al., 2017]: Sparse mixtures of mixtures using Bayesian estimation
- [Frühwirth-Schnatter and Malsiner-Walli, 2019]: "From here to infinity"- sparse finite versus Dirichlet process mixtures in model-based clustering:
  - **Sparse finite mixture for discrete-valued data:** Poisson and negative binomial mixture for count data; sparse latent class models; sparse finite mixtures of GLM regression models;
  - **Sparse finite mixtures of skew-N and skew-t distributions**

Sparse Gaussian Mixtures: some benchmark data sets

Data set	N	r	K_{tr}	Frequentistic (`mclust`)	Sparse Gaussian Mixtures ($K = 10$)
Iris	150	4	**3**	**2** $adj = 0.57,\ er = 0.33$	**3** $adj = 0.92,\ er = 0.03$
Crabs	200	5	**4**	**9** $adj = 0.48,\ er = 0.46$	**4** $adj = 0.80,\ er = 0.08$
Flea beetles	74	6	**3**	**5** $adj = 0.77,\ er = 0.18$	**3** $adj = 1,\ er = 0.00$
AIS	202	3	**2**	**3** $adj = 0.73,\ er = 0.13$	**3** $adj = 0.76,\ er = 0.11$
Wisconsin	569	3	**2**	**4** $adj = 0.55,\ er = 0.30$	**4** $adj = 0.62,\ er = 0.21$
Yeast	626	3	**2**	**8** $adj = 0.50,\ er = 0.20$	**6** $adj = 0.48,\ er = 0.23$

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

Problems with clustering via Gaussian mixtures

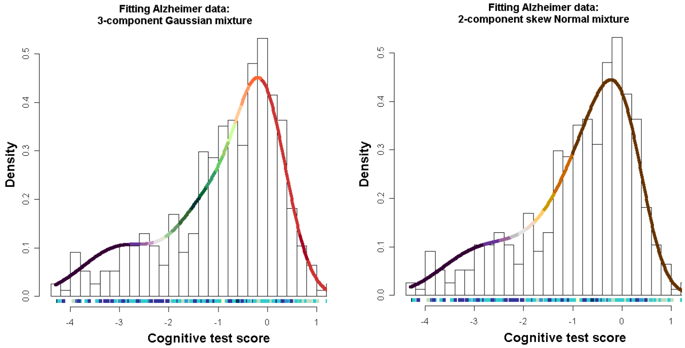
- **Clustering kernel in the mixture model essential for classification**.
- $f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k)$ should describe the variation of the observations $\mathbf{y}_i$ in cluster k by a **realistic probabilistic model**.
- If multivariate normal distributions are used as clustering kernel, i.e. $f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k) \sim \mathcal{N}_r(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, a problem might arise, if the **component density has been misspecified**.
- In this case, it is problematical to identify the order K of the mixture distribution with the number of clusters in the data, since several Gaussian components have to be merged to address this misspecification.

Example: Alzheimer's Disease Data

- Alzheimer's disease (AD) is a complex disease that has multiple genetic as well as environmental risk factors. It is commonly characterized by loss of a wide range of cognitive abilities with aging.
- For the present analysis, the data set consists of 451 subjects from the cohorts of the Religious Orders Study (ROS), see [Wilson et al., 2004] and the Memory and Aging Project (MAP), see [Bennett et al., 2005].
- The level of cognition of the subjects was clinically evaluated proximate to their death based on tests of cognitive functions and summarized by a mean global cognition score, with higher scores suggesting better cognition capabilities.
- The genetic risk factor Apolipoprotein E (ApoE) polymorphism was determined by genotyping the DNA from the subjects' blood.

Example: Alzheimer's Disease Data

- [Frühwirth-Schnatter and Pyne, 2010], $N = 415$ patient



- Apply sparse finite mixtures of skew-N and skew-t distributions

Finite mixtures of skew-N and skew-t distributions

- Clustering kernel: parametric non-Gaussian distributions
- Uni-/multivariate mixtures of **skew normal and the skew-t distribution** [Frühwirth-Schnatter and Pyne, 2010, Lee and McLachlan, 2013]

Standard skew-N distribution

A univariate random variable Y follows a standard skew-N distribution with skewness parameter α, if the pdf takes the form

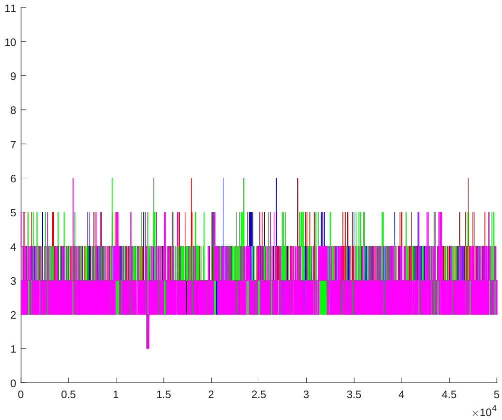
$$p(y) = 2\phi(y)\Phi(\alpha y),$$

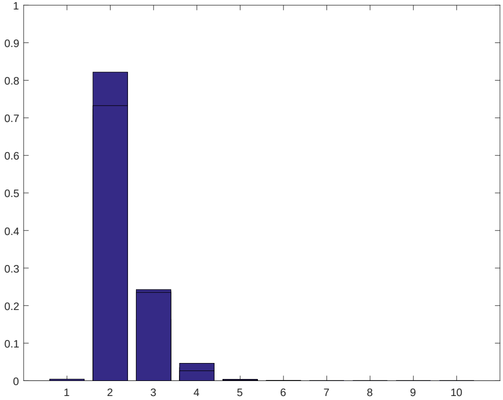
where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and the cdf of the standard normal distribution.

  - **Left-skewed ($\alpha < 0$)** or **right-skewed ($\alpha > 0$)**; $\alpha = 0$: standard normal
- **Standard skew-t with ν degrees of freedom:** $\phi(y)$ and $\Phi(\alpha y)$ are, respectively, the cdf and the pdf of a standard t_ν-distribution.

Sparse skew-N mixtures for Alzheimer data

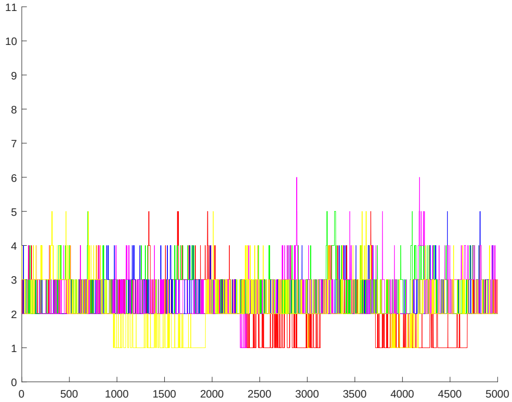
$K = 10$, $e_0 = 0.01$ $\quad \Rightarrow \hat{\mathbf{K}}_+ = \mathbf{2}$ was selected for various priors

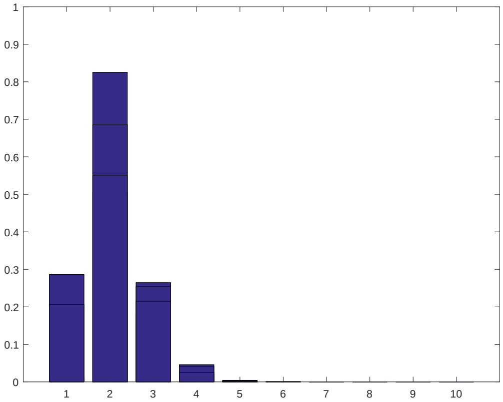




Sparse skew-t mixtures

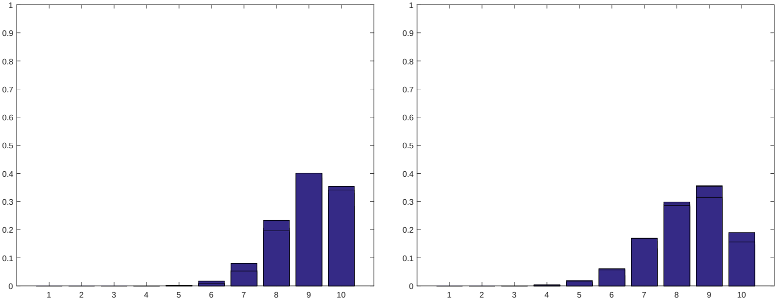
$K = 10$, $e_0 = 0.01$ $\Rightarrow$ $\widehat{\mathbf{K}}_+ = \mathbf{2}$ was selected for all priors





Skew-N and skew-t with Jeffrey's prior on the weights

- Because of $d = 3$ (Skew-N) and $d = 4$ (Skew-t), [Rousseau and Mengersen, 2011] would allow $e_0 = 0.5$ (Jeffrey's prior, $K = 10$)
- However, **strong overfitting** $\Rightarrow \widehat{\mathsf{K}}_+ = \mathbf{9}$ both for Skew-N (left) and Skew-t (right)



Posterior distribution $\Pr(K_+|\mathbf{y})$ for Alzheimer data

Skew normal	$K_+ = 1$	$K_+ = 2$	$K_+ = 3$	$K_+ = 4$	$K_+ = 5$	$K_+ = 6$	$K_+ \geq 7$
SFM ($K = 10$)							
$e_0 \sim \mathcal{G}(1, 200)$	0.0127	**0.76**	0.193	0.0285	0.00512	0.00032	0
DPM							
$\alpha \sim \mathcal{G}(2, 4)$	0	0.181	**0.302**	0.214	0.139	0.0827	0.0819
Skew-t							
SFM ($K = 10$)							
$e_0 \sim \mathcal{G}(1, 200)$	0.263	**0.597**	0.124	0.0152	0.00124	2e-05	0.
DPM							
$\alpha \sim \mathcal{G}(2, 4)$	0.0028	**0.29**	0.275	0.206	0.124	0.0583	0.0445
$\log \hat{p}(\mathbf{y}|K)$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
Skew normal	-689.62	**-682.37**	-684.45	-690.41	-696.12	-	-
Skew-t	-692.29	**-688.98**	-690.31	-694.11	-699.85	-	-

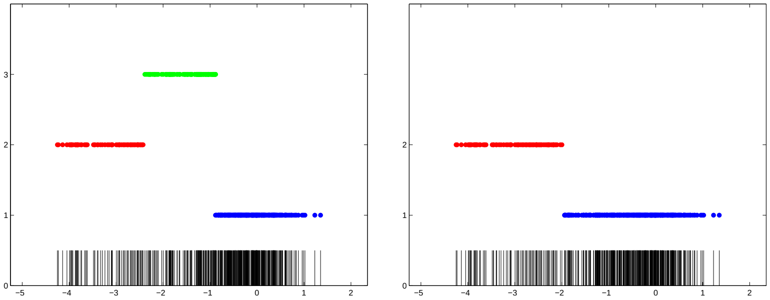
Parameter Estimation

- Two component skew normal mixture modeling of Alzheimer's disease data set.
- Parameter estimation using posterior means (posterior standard deviations in parenthesis)

k	ξ_k	ω_k^2	α_k	$\mu_k = \mathrm{E}(Y\|S_i = k)$	η_k
1	0.36 (0.11)	1.26 (0.37)	-2.61 (0.78)	-0.46 (0.10)	0.767 (0.061)
2	-3.55 (0.43)	2.20 (1.3)	2.06 (1.48)	-2.65 (0.34)	0.233 (0.061)

  - The first component has a much higher expected cognitive score μ_k than the second one;
  - The first component exhibit considerable negative skewness, while the skewness parameter is positive for the second component.

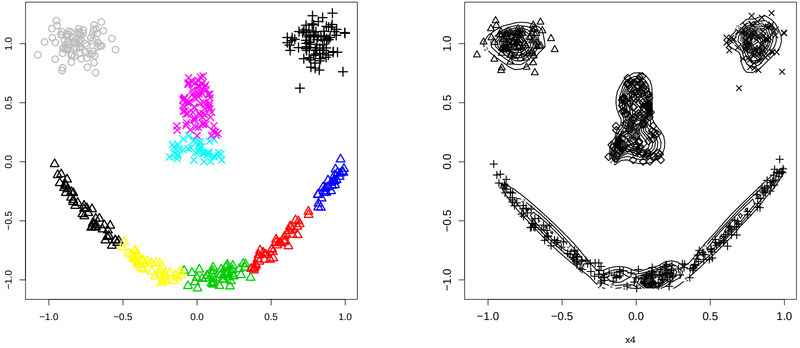
Clustering



Clustering of the data based on a mixture of three normal distributions (left hand side) and on a mixture of two skew normal distributions (right hand side)

Smiley's Data [Leisch, 2004]

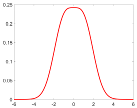
Sparse Gaussian finite mixture approach [Malsiner Walli et al., 2016] yields 9 "clusters"

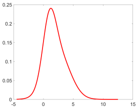


Resulting clustering solution (left) and fitted density (right)

Semi-parameteric non-Gaussian cluster kernels

- It may be difficult to decide which parametric distribution is appropriate to characterize a data cluster, especially in higher dimensions.
- [Malsiner Walli et al., 2017] pursue a sparse mixture of Gaussian mixtures approach:
  - exploits the ability of normal mixtures to accurately approximate a wide class of probability distributions: **models the non-Gaussian cluster distributions themselves by Gaussian mixtures**





  - use the concept of **sparse finite mixtures to select the number of clusters**.

Sparse Gaussian mixtures-of-mixtures

- Consider an overfitting finite mixture with a sparse Dirichlet prior on $\boldsymbol{\eta} \sim \mathcal{D}_K(e_0)$, i.e.

$$p(\mathbf{y}) = \sum_{k=1}^{K} \eta_k f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k),$$

- where each cluster distribution $f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k)$ is assumed to be a mixture of L multivariate normal distributions (subcomponents):

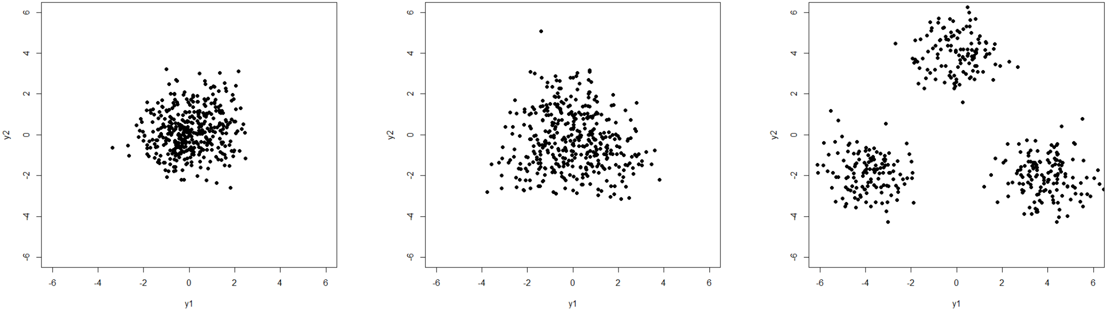
$$f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k) = \sum_{l=1}^{L} w_{kl} f_N(\mathbf{y}|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}). \tag{4}$$

- The Gaussian mixture (4) provides a **semi-parametric density fit** to possibly asymmetric, heavy-tailed cluster distributions.
- The sparse Dirichlet prior $\mathcal{D}_K(e_0)$ allows to **estimate the number of these (non-Gaussian) clusters**.

Variance decomposition of a mixture of normals

Fraction of variance explained by differences in the means:

$$\sum_{k=1}^{K} \eta_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})' = \phi Cov(\mathbf{y})$$



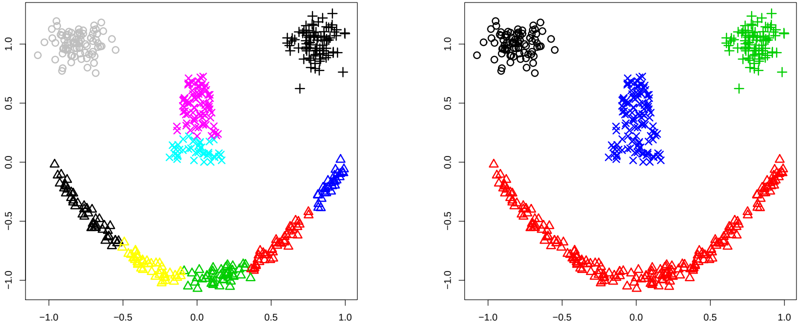
$\phi = 0.1$ (left), $\phi = 0.5$ (middle), $\phi = 0.9$ (right)

Bayesian inference

- Within each cluster k, the Gaussian mixture provides a **semi-parametric density fit** to a possibly asymmetric, heavy tailed cluster distribution with strong prior overlap of the component densities:
  - $\mathbf{w}_k = (w_{k1}, \dots, w_{kL}) \sim \mathcal{D}_L(f_0)$ with $f_0 = d/2 + 2$;
  - $p(\boldsymbol{\theta}_k|\psi) = p(\boldsymbol{\mu}_{k1}, \dots, \boldsymbol{\mu}_{kL}|\psi_1)p(\boldsymbol{\Sigma}_{k1}, \dots, \boldsymbol{\Sigma}_{kL}|\psi_2)$
  - $\boldsymbol{\mu}_{kl}|\mathbf{b}_k \overset{iid}{\sim} \mathcal{N}(\mathbf{b}_k, \mathbf{B}_0), l = 1, \dots, L$ with small prior variation (ϕ_W)
  - $\mathbf{b}_k \sim \mathcal{N}(\mathbf{b}_0, \mathbf{M}_0)$ with large prior variation (ϕ_B)
  - $\boldsymbol{\Sigma}_{kl}^{-1}|c_0, \mathbf{C}_{0k} \overset{iid}{\sim} \mathcal{W}_r(c_0, \mathbf{C}_{0k}), l = 1, \dots, L$ and $\mathbf{C}_{0k}|g_0, \mathbf{G}_0 \overset{iid}{\sim} \mathcal{W}_r(g_0, \mathbf{G}_0)$.
- Related to the infinite mixture of infinite Gaussian mixtures [Yerebakan et al., 2014].

Revisiting Smiley

Right-hand side: sparse Gaussian mixture-of-mixture model
$(K = 10, L = 10, e_0 = 0.01) \quad \Rightarrow \mathbf{\widehat{K}_+ = 4}$



Invariance of the likelihood

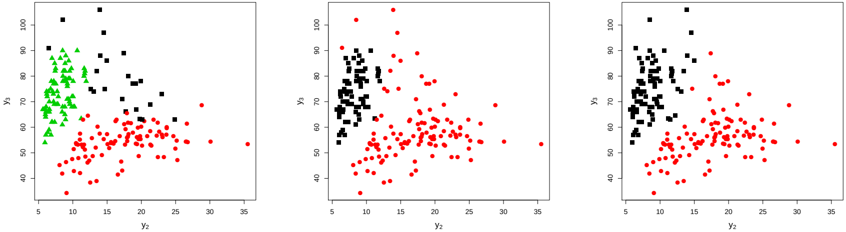
- The likelihood is completely ignorant concerning the issue which of the KL components belong together:

$$p(\mathbf{y}|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k f_T(\mathbf{y}|\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \sum_{l=1}^{L} \tilde{w}_{kl} f_N(\mathbf{y}|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}),$$

  because only $\tilde{w}_{kl} = \eta_k w_{kl}$ is identified.
- Components are often **merged in a post-processing fashion** [Li, 2005, Baudry et al., 2010, Hennig, 2010, Melnykov, 2016].
- Identification achieved through the carefully designed hierarchical prior.

Pitfalls of post-processing merging

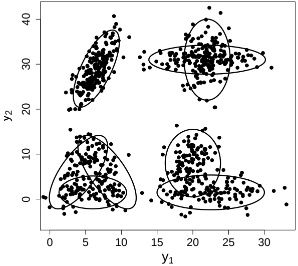
- AIS data set, variables "X.Bfat" and "LBM"; scatter plots of the observations with different estimated classifications

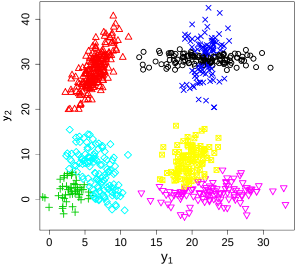


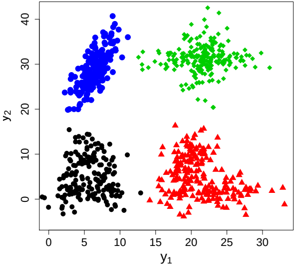
  - left-hand side: `Mclust` with $K = 3$ [Fraley et al., 2012]
  - middle: `combiClust` [Baudry et al., 2010]
  - right-hand side: **sparse mixture of mixtures** approach $(K = 10, L = 4) \Rightarrow \mathbf{\widehat{K}_+ = 2}$

Simulated data

- Data simulated from a mixture of 8 bivariate normal distributions (left)







- Clustering using a sparse Gaussian mixture ($K = 10, e_0 = 0.001$; middle)
- Clustering using a **sparse Gaussian mixture-of-mixture model** ($K = 10, L = 4, e_0 = 0.001$; right)

Revisiting the benchmark data sets

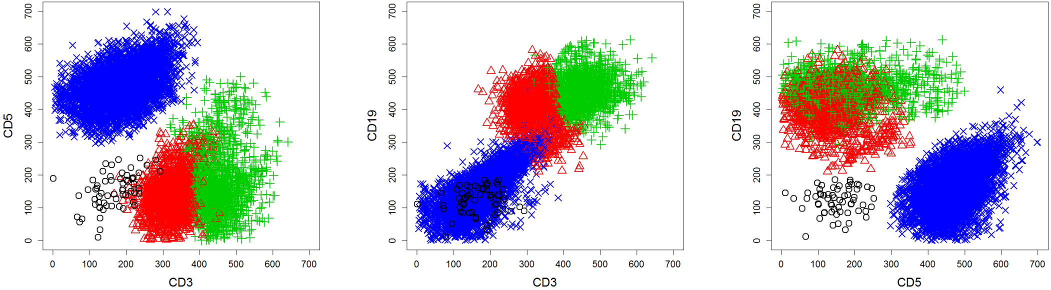
Data set	$\mathbf{K_+}$ for `SparseMix` $L = 1$	$\hat{K}_+$ **for** `SparseMixMix` $(K = 10, e_0 = 0.001)$ $L = 4$	$L = 5$
AIS	**3** $adj = 0.76, \ er = 0.11$	**2** $adj = 0.81, \ er = 0.05$	**2** $adj = 0.76, \ er = 0.06$
Wisconsin	**4** $adj = 0.62, \ er = 0.21$	**2** $adj = 0.82, \ er = 0.05$	**2** $adj = 0.82, \ er = 0.05$
Yeast	**6** $adj = 0.48, \ er = 0.23$	**2** $adj = 0.81, \ er = 0.05$	**2** $adj = 0.76, \ er = 0.06$

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

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

Example: flow cytometric data

- $N = 7932$ data points ($d = 4$)
- sparse mixture of mixtures ($K = 30$, $e_0 = 0.001$; $L = 15$) yields $\widehat{\mathbf{K}}_+ = \mathbf{4}$ ($\mathbf{K_{tr}} = \mathbf{4}$)



- Error rate (0.03) outperforms the error rate of 0.056 reported by [Lee and McLachlan, 2013]

Example: Fabric fault data

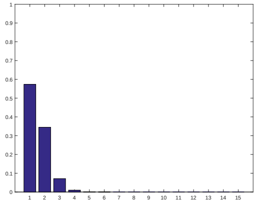
- Regression analysis of data on fabric faults [Aitkin, 1996, McLachlan and Peel, 2000].
- The response variable y_i is the number of faults in a bolt of length l_i
- Log marginal likelihoods of various mixtures of regression models based on the regressor $x_i = (1 \;\; \log l_i)$ [Frühwirth-Schnatter et al., 2009]

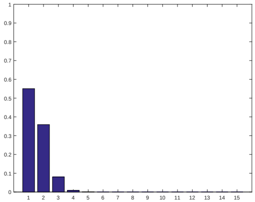
Model	$K = 1$	$K = 2$	$K = 3$	$K = 4$
Poisson	−101.79	−99.21	−100.74	−103.21
Poisson (fixed slope)	−101.79	−97.46	−97.65	−98.60
Negative Binomial	−96.04	−99.05	−102.21	−104.95
Negative Binomial (fixed slope)	−96.04	−97.25	−98.76	−99.97

- Marginal likelihood (based on $e_0 = 4$) points to a **homogeneous model** based on the negative binomial distribution

Sparse mixture of binomial regression models

- $K = 10 \quad \Rightarrow \widehat{\mathbf{K}}_{+} = \mathbf{1}$ is selected for $e_0 = 0.01$ (left hand side) and $e_0 = 0.1$ (right hand side)





- Sparse finite mixtures are also **useful for “testing” homogeneity**

Outline

Part I: Finite Mixture Models and Model-based Clustering

Model selection criteria

- **Marginal likelihoods** - model selection including K, the clustering kernel, etc.
- **One-sweep Bayesian methods**:
  - RJMCMC [Richardson and Green, 1997] - selection of K (K_+ as a by-product)
  - Sparse finite mixtures (SFS and Malsiner-Walli, 2019, ADAC) - selection of K_+
- **Statistical (information) criteria**:
  - BIC - model selection including K, the clustering kernel, etc.
  - DIC [Spiegelhalter et al., 2002] - application to finite mixture models is not without problems [Celeux et al., 2006]
  - Entropy-based criteria: penalize the failure of the model to provide a classification into well-separated clusters
- See [Celeux et al., 2019] for a review.

Marginal likelihoods

- Definition of the **marginal likelihood** $p(\mathbf{y}|K)$:

$$p(\mathbf{y}|K) = \int p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)d\boldsymbol{\vartheta}. \tag{5}$$

- Computational challenge:
  - Marginal likelihoods difficult to compute [Celeux et al., 2019]
  - Keeping the balance across multiple modes important (SFS, 2019, BJPS)
- Interpretation of marginal likelihoods:
  - What are we actually estimating? K or K_+?
  - Again, the choice of the prior distribution $\boldsymbol{\eta} \sim \mathcal{D}_K(\gamma)$ on the weight distribution is important.

Sampling based approximation of the marginal likelihood

- **Importance sampling** is based on rewriting (5) as

$$p(\mathbf{y}|K) = \int \frac{p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)}{q_K(\boldsymbol{\vartheta})} q_K(\boldsymbol{\vartheta}) \; \boldsymbol{\vartheta},$$

- Determine a sample $\boldsymbol{\vartheta}^{(l)}, l = 1, \dots, L$ from the importance density $q_K(\boldsymbol{\vartheta})$.
- The importance sampling estimator of the marginal likelihood is given by:

$$\hat{p}_{IS}(\mathbf{y}|K) = \frac{1}{L} \sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)}, K)p(\boldsymbol{\vartheta}^{(l)}|K)}{q_K(\boldsymbol{\vartheta}^{(l)})}. \tag{6}$$

Computational challenges

- **Tail behaviour**
  - $\hat{p}_{IS}(\mathbf{y}|K)$ has high standard errors, if $q_K(\boldsymbol{\vartheta})$ has thin tails compared to the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$.
  - **Bridge sampling estimators** are robust to the tail behaviour [Meng and Wong, 1996, Frühwirth-Schnatter, 2004]
- **Keeping the balance**
  - Importance density $q_K(\boldsymbol{\vartheta})$ has to mimic the multimodality of the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ which results from invariance to label switching.
  - Various strategies to ensure balanced importance densities (SFS, 2019, BJPS).

The bridge sampling estimator

- Choose two functions:
  - an importance function $q_K(\boldsymbol{\vartheta})$ (approximation to the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$)
  - a positive function $\alpha(\boldsymbol{\vartheta})$ such that $\int \alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}, K)\,\mathrm{d}\,\boldsymbol{\vartheta} > 0$
- **General bridge sampling estimator of the marginal likelihood**:
  - The identity:

$$\int \alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}, K)\,\mathrm{d}\,\boldsymbol{\vartheta} = \int \alpha(\boldsymbol{\vartheta})\frac{p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)}{p(\mathbf{y}|K)}q_K(\boldsymbol{\vartheta})\,\mathrm{d}\,\boldsymbol{\vartheta}$$

  - yields following estimator of the marginal likelihood:

$$p(\mathbf{y}|K) = \frac{\mathrm{E}_{q_K(\boldsymbol{\vartheta})}\left(\alpha(\boldsymbol{\vartheta})p(\mathbf{y}|\boldsymbol{\vartheta}, K)p(\boldsymbol{\vartheta}|K)\right)}{\mathrm{E}_{p(\boldsymbol{\vartheta}|\mathbf{y},K)}\left(\alpha(\boldsymbol{\vartheta})q_K(\boldsymbol{\vartheta})\right)}.$$

The bridge sampling estimator

- [Meng and Wong, 1996] derive an optimal choice for $\alpha(\boldsymbol{\vartheta})$ which yields a bridge sampling estimator that requires i.i.d. draws $\boldsymbol{\vartheta}^{(l)}, l = 1, \ldots, L$ from the importance density $q_K(\boldsymbol{\vartheta})$ and i.i.d. draws from the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$.
- Markov chain Monte Carlo (MCMC) draws $\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M$ from the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$ are typically autocorrelated.
- [Meng and Schilling, 1996] define an alternative optimal bridge sampling estimator $p_{BS}(\mathbf{y}|K)$ based on following function $\alpha(\boldsymbol{\vartheta})$:

$$\alpha(\boldsymbol{\vartheta}) = 1/\left(L \cdot q_K(\boldsymbol{\vartheta}) + M_\star \cdot p(\boldsymbol{\vartheta}|\mathbf{y}, K)\right).$$

  - $M_\star$ is the effective sample size, estimated as $\hat{M}_\star = \min(M, M/\hat{\rho})$, where $\hat{\rho}$ is an estimator of the inefficiency factor of the posterior draws $f^{(m)} = p(\mathbf{y}|\boldsymbol{\vartheta}^{(m)}, K)p(\boldsymbol{\vartheta}^{(m)}|K)$.
  - This definition of $\alpha(\boldsymbol{\vartheta})$ requires knowledge of the (unknown) normalizing constant $p(\mathbf{y}|K)$ to evaluate $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$.

Computing the (optimal) bridge sampling estimator

- Derive two sets of independent draws:
  - MCMC draws $\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M$ from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y}, K)$;
  - independent draws $\boldsymbol{\vartheta}^{(l)}, l = 1, \ldots, L$ from the importance density $q_K(\boldsymbol{\vartheta})$.
- The following recursion is applied until convergence:

$$\hat{p}_{BS}(\mathbf{y}|K) = \lim_{t\to\infty} \hat{p}_{BS,t}(\mathbf{y}|K).$$

  - Use the IS estimator $\hat{p}_{IS}(\mathbf{y}|K)$ (6) as a starting value $\hat{p}_{BS,0}(\mathbf{y}|K)$.
  - Define $\hat{p}_{BS,t}(\mathbf{y}|K)$ recursively:

$$\hat{p}_{BS,t}(\mathbf{y}|K) = \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)}, K)p(\boldsymbol{\vartheta}^{(l)}|K)}{Lq_K(\boldsymbol{\vartheta}^{(l)}) + \hat{M}_\star \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(l)}, K)p(\boldsymbol{\vartheta}^{(l)}|K)}{\hat{p}_{BS,t-1}(\mathbf{y}|K)}}}{\frac{1}{M}\sum_{m=1}^{M} \frac{q_K(\boldsymbol{\vartheta}^{(m)})}{Lq_K(\boldsymbol{\vartheta}^{(m)}) + \hat{M}_\star \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{(m)}, K)p(\boldsymbol{\vartheta}^{(m)}|K)}{\hat{p}_{BS,t-1}(\mathbf{y}|K)}}}.$$

Constructing the importance density

- $\mathbf{S}^{(m)}, m = 1, \ldots, M$ are M posterior draws of the latent allocations $\mathbf{S}$.
- Rao–Blackwellised approximation of the posterior distribution of $\boldsymbol{\vartheta}$ based on introducing the latent allocations $\mathbf{S}$ as missing data yields:

$$p(\boldsymbol{\vartheta}|\mathbf{y}, K) = \sum_{\mathbf{S}} p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{y}, K)p(\mathbf{S}|\mathbf{y}, K) \approx q_K(\boldsymbol{\vartheta}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)}, \mathbf{y}, K) \qquad (7)$$

  - Conditional density $p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{y}, K)$ often from a well-known family e.g. Poisson mixtures: $\mu_k|\mathbf{S}, \mathbf{y} \sim \mathcal{G}\left(a_0 + \overline{y}_k, b_0 + N_k\right)$
  - Gibbs sampling might lead to imbalanced label switching
  - Enforce label switching to ensure that importance density is (nearly) balanced.

Achieving balance

Double random permutation bridge sampling estimators:

- (Nearly) balanced label switching of the MCMC draws $(\boldsymbol{\vartheta}^{(m)}, \mathbf{S}^{(m)}), m = 1, \ldots, M$ through random permutation of the labels [Frühwirth-Schnatter, 2001b]
- Independent random permutation when constructing $q_K(\boldsymbol{\vartheta})$ in (7)

Achieving balance

Full permutation bridge sampling estimators:

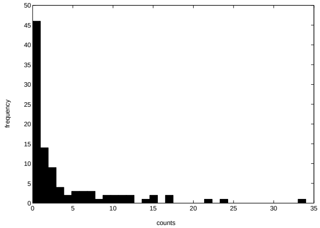
- Construct a fully symmetric importance density:
  - choose $q = 1, \ldots, Q$ MCMC draws,
  - for each q, define $K!$ expanded component densities by applying all possible permutations $\rho \in \mathcal{S}_K$:

$$q_K(\boldsymbol{\vartheta}) = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{K!} \sum_{\rho \in \mathcal{S}_K} p(\boldsymbol{\vartheta} | \rho(\mathbf{S}^{(q)}), \mathbf{y}, K).$$

- Robust to unbalanced label switching in the MCMC draws

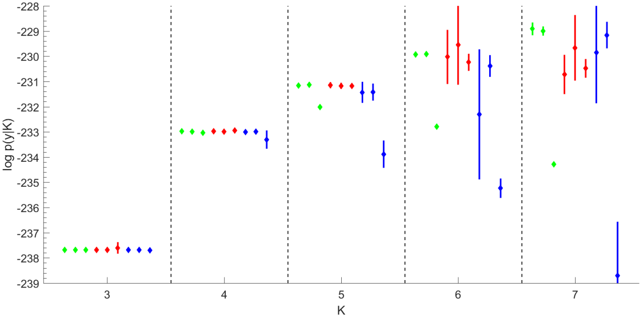
Example: Eye Tracking Data

- For illustration, consider the count data on eye tracking anomalies in 101 schizophrenic patients studied by [Escobar and West, 1998]
- empirical distribution of the observations:



- fit a mixture of Poisson distributions with unknown number of components

EYE TRACKING DATA: log marginal likelihoods with $e_0 = 4$



For each K, nine estimators $\log \hat{p}_{\bullet}(\mathbf{y}|K) \pm 3\,\mathrm{SE}$ are given (from left to right):
$\log \hat{p}_{BS,F}(\mathbf{y}|K)$, $\log \hat{p}_{BS,D}(\mathbf{y}|K)$, $\log \hat{p}_{BS,R}(\mathbf{y}|K)$ (green); $\log \hat{p}_{IS,F}(\mathbf{y}|K)$, $\log \hat{p}_{IS,D}(\mathbf{y}|K)$, $\log \hat{p}_{IS,R}(\mathbf{y}|K)$ (red); $\log \hat{p}_{RI,F}(\mathbf{y}|K)$, $\log \hat{p}_{RI,D}(\mathbf{y}|K)$, $\log \hat{p}_{RI,R}(\mathbf{y}|K)$ (blue).

Finite mixture models with a prior on K

- [Richardson and Green, 1997] consider finite mixture models with a discrete prior on K (K is random apriori)
  - $p(K)$ is a truncated uniform
  - $\boldsymbol{\eta}_K | K \sim \mathcal{D}_K(1)$ is uniform
  - RJMCMC for a one-sweep sampler
- [Nobile, 2004] shows that a proper prior on K is needed to obtain a proper posterior $p(K|\mathbf{y})$
- [Miller and Harrison, 2018] show that sampler from BNP mixtures can be used
- SFS, Malsiner-Walli, Grün (coming soon): learn K and K_+ under sensible priors on $p(K)$ and $\boldsymbol{\eta}_K | K \sim \mathcal{D}_K(\gamma_K)$

Reversible Jump MCMC

- Start with a certain mixture model with K components and select classifications $\mathbf{S} = (S_1, \ldots, S_N)$ where S_i assign a certain observation to a certain component ($S_i = k \Rightarrow$ assign y_i to component k).
- Repeat the following steps for $m = 1, \ldots, M$:
  (a) Perform the following **dimension-preserving move**:
    (a-1) Update the model-specific parameter $\boldsymbol{\vartheta}_K = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \eta_1, \ldots, \eta_K)$
    (a-2) Update the current allocation $\mathbf{S}$.
  (b) Perform the following **dimension-changing moves**:
    (b-1) split one mixture component into two components or merge two components into one.
    (b-2) delete or add empty components

Moving to a Mixture with $K+1$ Components

Assume that the current model $\mathcal{M}_K$ is a mixture with K components, the model parameter being equal to $\boldsymbol{\vartheta}_K$. To jump to a mixture model $\mathcal{M}_{K+1}$ with $K+1$ components, proceed in the following way.

(a) Match the dimensions between the models: propose $\mathbf{u}$, where $\dim(\boldsymbol{\vartheta}_{K+1}) = \dim(\boldsymbol{\vartheta}_K) + \dim(\mathbf{u})$, from a proposal density $q_{K,K+1}(\mathbf{u})$, and determine $\boldsymbol{\vartheta}_{K+1}$ from $\boldsymbol{\vartheta}_{K+1} = g_{K,K+1}(\boldsymbol{\vartheta}_K, \mathbf{u})$.

(b) Reallocate the observations according to a proposal $q(\mathbf{S}^{\text{new}}|\mathbf{S}, \boldsymbol{\vartheta}_{K+1})$.

(c) Move to the finite mixture model $\mathcal{M}_{K+1}$ with component parameter $\boldsymbol{\vartheta}_{K+1}$ and allocations $\mathbf{S}^{\text{new}}$ with probability $\min(1, A)$.

Acceptance probability

The acceptance probability A depends on $\boldsymbol{\vartheta}_K$, $\boldsymbol{\vartheta}_{K+1}$, $\mathbf{S}$ and $\mathbf{S}^{\text{new}}$:

$$A = (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times |\text{Jacobian}|,$$

$$\text{likelihood ratio} = \prod_{i:S_i^{\text{new}} \neq S_i} \frac{p(\mathbf{y}_i|\boldsymbol{\theta}_{S_i^{\text{new}}})}{p(\mathbf{y}_i|\boldsymbol{\theta}_{S_i})}$$

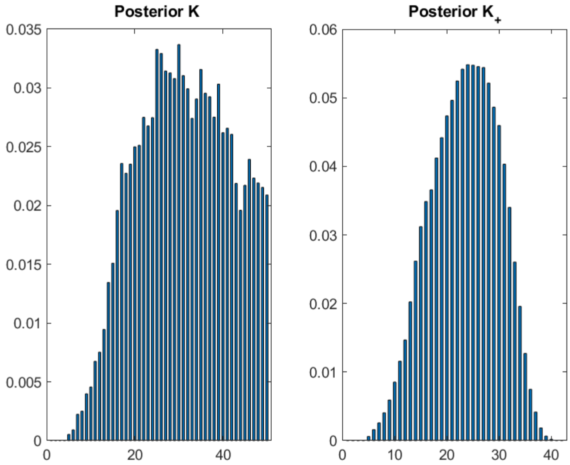
$$\text{prior ratio} = \frac{p(\mathbf{S}^{\text{new}}|\boldsymbol{\vartheta}_{K+1}, \mathcal{M}_{K+1})p(\boldsymbol{\vartheta}_{K+1}|\mathcal{M}_{K+1})\Pr(\mathcal{M}_{K+1})}{p(\mathbf{S}|\boldsymbol{\vartheta}_K, \mathcal{M}_K)p(\boldsymbol{\vartheta}_K|\mathcal{M}_K)\Pr(\mathcal{M}_K)}$$

$$\text{proposal ratio} = \frac{m_h(\boldsymbol{\vartheta}_{K+1}, \mathcal{M}_{K+1})}{q(\mathbf{S}^{\text{new}}|\mathbf{S}, \boldsymbol{\vartheta}_{K+1})q_{K,K+1}(\mathbf{u})m_h(\boldsymbol{\vartheta}_K, \mathcal{M}_K)}$$

$$|\text{Jacobian}| = \left|\frac{\partial g_{K,K+1}(\boldsymbol{\vartheta}_K, \mathbf{u})}{\partial(\boldsymbol{\vartheta}_K, \mathbf{u})}\right|.$$

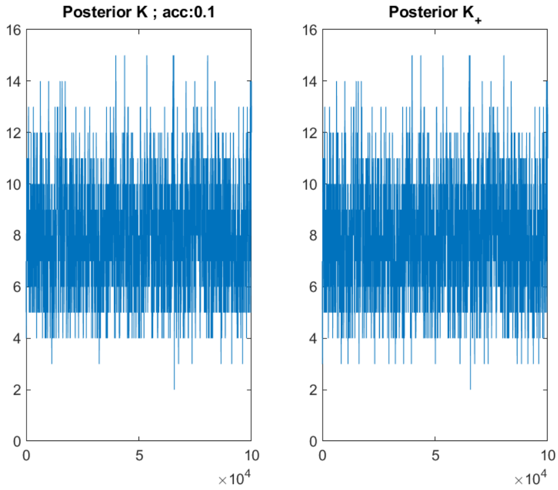
Eye Tracking Data, RJMCMC under "no prior"

Uniform prior on K, $\boldsymbol{\eta} \sim \mathcal{D}_K(1)$



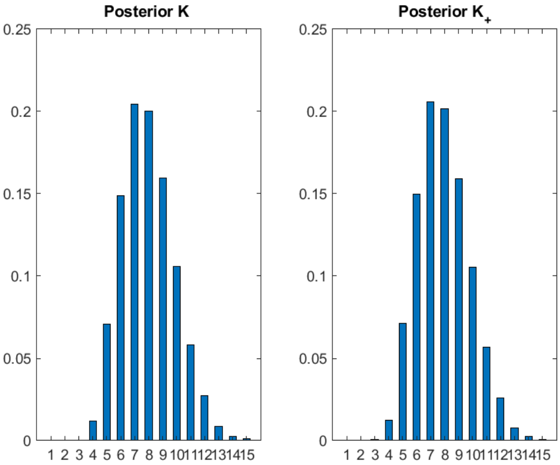
Eye Tracking Data, RJMCMC with informative priors

$K - 1 \sim \mathcal{P}(4)$, $\boldsymbol{\eta} \sim \mathcal{D}_K(4)$



Eye Tracking Data, RJMCMC - density estimation

$K-1 \sim \mathcal{P}(4)$, $\boldsymbol{\eta} \sim \mathcal{D}_K(4)$



Information criteria

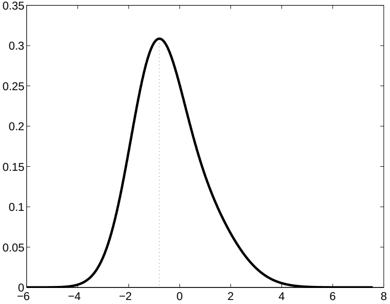
- Minimize BIC_K defined as

$$\mathrm{BIC}_K = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_K, \mathcal{M}_K) + \log(N) d_K,$$

  where d_K is the number of unknown parameters in the mixture distribution and $\hat{\boldsymbol{\vartheta}}_K$ is the ML estimator.
  - BIC_K is an asymptotic approximation to $-2 \log p(\mathbf{y}|\mathcal{M}_K)$ which ignores the prior $p(\boldsymbol{\vartheta}_K|\mathcal{M}_K)$;
  - BIC_K consistent for K, if component density correctly specified [Keribin, 2000]
- AIC_K criterion - penalty equals **2**d_K.

Components versus clusters

- For (large) data sets, BIC and the marginal likelihood tends to overfit the number of clusters, because the clustering kernel is likely to be misspecified.
- Several normal distributions may be necessary to capture skewness and kurtosis in a single skew cluster, e.g. a mixture of two Gaussians with $\mu_1 = -1$, $\mu_2 = 0.5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\eta_1 = 0.6$



Entropy-based criteria

- [Biernacki et al., 2000] introduce the integrated classification likelihood criterion which is approximately equal to [McLachlan and Peel, 2000]:

$$\text{ICL-BIC}_K = \text{BIC}_K + 2\text{EN}(\hat{\boldsymbol{\vartheta}}_K).$$

- The **entropy** $\text{EN}(\boldsymbol{\vartheta}_K)$ measures how well the finite mixture model defined by $\boldsymbol{\vartheta}_K$ classifies the data into K distinct clusters:

$$\text{EN}(\boldsymbol{\vartheta}_K) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(S_i = k|\mathbf{y}_i, \boldsymbol{\vartheta}_K) \log \Pr(S_i = k|\mathbf{y}_i, \boldsymbol{\vartheta}_K),$$

- The ICL-BIC_K criterion penalizes not only model complexity, but also the failure of the model to provide a classification into well-separated clusters.

Part II

Hidden Markov and Markov Switching Models

Aitkin, M. (1996).

A general maximum likelihood analysis of overdispersion in generalized linear models.

Statistics and Computing, 6:251–262.

Albert, J. H. and Chib, S. (1993).

Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts.

Journal of Business & Economic Statistics, 11:1–15.

Alspach, D. L. and Sorenson, H. W. (1972).

Nonlinear Bayesian estimation using Gaussian sum approximations.

IEEE Transactions on Automatic Control, 17:439–448.

Ang, A. and Bekaert, G. (2002).

Regime switches in interest rates.

Journal of Business & Economic Statistics, 20:163–182.

Banfield, J. D. and Raftery, A. E. (1993).

Model-based Gaussian and non-Gaussian clustering.

Biometrics, 49:803–821.

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

Bennett, D. A., Schneider, J. A., Buchman, A. S., de Leon, C. M., Bienias, J. L., and Wilson, R. S. (2005).
The Rush Memory and Aging Project: Study Design and Baseline Characteristics of the Study Cohort.
Neuroepidemiology, 25:163–175.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997).
Inference in model-based cluster analysis.
Statistics and Computing, 7:1–10.

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.

Binder, D. A. (1978).
Bayesian cluster analysis.
Biometrika, 65:31–38.

Bollerslev, T. (1986).
Generalized autoregressive conditional heteroskedasticity.
Journal of Econometrics, 31:307–327.

Cai, J. (1994).
A Markov model of switching-regime ARCH.
Journal of Business & Economic Statistics, 12:309–316.

Carter, C. K. and Kohn, R. (1997).
Semiparametric Bayesian inference for time series with mixed spectra.
Journal of the Royal Statistical Society, Ser. B, 59:255–268.

Cecchetti, S. G., Lam, P., and Mark, N. C. (1990).
Mean reversion in equilibrium asset prices.
The American Economic Review, 80:398–418.

Celeux, G. (1998).
Bayesian inference for mixture: The label switching problem.
In Green, P. J. and Rayne, R., editors, *COMPSTAT 98*, pages 227–232. Physica, Heidelberg.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006).

Deviance information criteria for missing data models.
Bayesian Analysis, 1:651–674.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019).
Model selection for mixture models – perspectives and strategies.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 7, pages 117–154. CRC Press, Boca Raton, FL.

Celeux, G., Hurn, M., and Robert, C. P. (2000).
Computational and inferential difficulties with mixture posterior distributions.
Journal of the American Statistical Association, 95:957–970.

Chib, S. (1996).
Calculating posterior distributions and modal estimates in Markov mixture models.
Journal of Econometrics, 75:79–97.

Chib, S., Nardari, F., and Shephard, N. (2002).
Markov chain Monte Carlo methods for stochastic volatility models.
Journal of Econometrics, 108:281–316.

Dasgupta, A. and Raftery, A. E. (1998).
Detecting features in spatial point processes with clutter via model-based clustering.

Journal of the American Statistical Association, 93:294–302.

Diebolt, J. and Robert, C. P. (1994).
Estimation of finite mixture distributions through Bayesian sampling.
Journal of the Royal Statistical Society, Ser. B, 56:363–375.

Engel, C. (1994).
Can the Markov switching model forecast exchange rates?
Journal of International Economics, 36:151–165.

Engel, C. and Hamilton, J. D. (1990).
Long swings in the Dollar: Are they in the data and do markets know it?
The American Economic Review, 80:689–713.

Engel, C. and Kim, C.-J. (1999).
The long-run U.S./U.K. real exchange rate.
Journal of Money, Credit, and Banking, 31:335–356.

Engle, R. F. (1982).
Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation.
Econometrica, 50:987–1007.

Escobar, M. D. and West, M. (1998).
Computing nonparametric hierarchical models.
In Dey, D., Müller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, number 133 in Lecture Notes in Statistics, pages 1–22. Springer, Berlin.

Everitt, B. S. (1979).
Unresolved problems in cluster analysis.
Biometrics, 35:169–181.

Fama, E. (1965).
The behavior of stock market prices.
Journal of Business, 38:34–105.

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., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012).
mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.
Technical Report 597, Department of Statistics, University of Washington.

Francq, C., Roussignol, M., and Zakoian, J. (2001).
Conditional heteroscedasticity driven by hidden Markov chains.
Journal of Time Series Analysis, 22:197–220.

Frühwirth-Schnatter, S. (1994).
Data augmentation and dynamic linear models.
Journal of Time Series Analysis, 15:183–202.

Frühwirth-Schnatter, S. (2001a).
Fully Bayesian analysis of switching Gaussian state space models.
Annals of the Institute of Statistical Mathematics, 53:31–49.

Frühwirth-Schnatter, S. (2001b).
Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models.
Journal of the American Statistical Association, 96:194–209.

Frühwirth-Schnatter, S. (2004).
Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques.
Econometrics Journal, 7:143–167.

Frühwirth-Schnatter, S. (2006).
Finite Mixture and Markov Switching Models.
Springer, New York.

Frühwirth-Schnatter, S. (2011a).
Dealing with label switching under model uncertainty.
In Mengersen, K., Robert, C. P., and Titterington, D., editors, *Mixture estimation and applications*, chapter 10, pages 213–239. Wiley, Chichester.

Frühwirth-Schnatter, S. (2011b).
Panel data analysis - A survey on model-based clustering of time series.
Advances in Data Analysis and Classification, 5:251–280.

Frühwirth-Schnatter, S. (2019).
Keeping the balance – Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and markov mixture models.
Brazilian Journal of Probability and Statistics, 33:706–733.

Frühwirth-Schnatter, S. and Frühwirth, R. (2007).
Auxiliary mixture sampling with applications to logistic models.
Computational Statistics and Data Analysis, 51:3509–3528.

Frühwirth-Schnatter, S. and Frühwirth, R. (2010).
Data augmentation and MCMC for binary and multinomial logit models.
In Kneib, T. and Tutz, G., editors, *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 111–132. Physica-Verlag, Heidelberg.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009).
Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data.
Statistics and Computing, 19:479–492.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019).
From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering.
Advances in Data Analysis and Classification, 13:33–64.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., and Winter-Ebmer, R. (2012).
Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering.
Journal of Applied Econometrics, 27:1116–1137.

Frühwirth-Schnatter, S., Pittner, S., Weber, A., and Winter-Ebmer, R. (2018).
Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering.
Annals of Applied Statistics, 12:1786–1830.

Frühwirth-Schnatter, S. and Pyne, S. (2010).
Bayesian inference for finite mixtures of univariate and multivariate skew normal and skew-t distributions.
Biostatistics, 11:317 – 336.

Frühwirth-Schnatter, S., Tüchler, R., and Otter, T. (2004).
Bayesian analysis of the heterogeneity model.
Journal of Business & Economic Statistics, 22:2–15.

Frühwirth-Schnatter, S. and Wagner, H. (2006).
Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling.
Biometrika, 93:827–841.

Garcia, R. and Perron, P. (1996).
An analysis of real interest rate under regime shift.
The Review of Economics and Statistics, 78:111–125.

Gormley, I. C. and Frühwirth-Schnatter, S. (2019).
Mixture of experts models.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 12, pages 271–307. CRC Press, Boca Raton, FL.

Gormley, I. C. and Murphy, T. B. (2008).
Exploring voting blocs within the Irish electorate: A mixture modeling approach.
Journal of the American Statistical Association, 103:1014–1027.

Granger, C. W. J. and Orr, D. (1972).
Infinite variance and research strategy in time series analysis.
Journal of the American Statistical Association, 67:275–285.

Gray, S. F. (1996).
Modeling the conditional distribution of interest rates as a regime switching process.
Journal of Financial Economics, 42:27–62.

Grilli, V. and Kaminsky, G. (1991).
Nominal exchange rate regimes and the real exchange rate: Evidence from the United States and Great Britain, 1885-1986.
Journal of Monetary Economics, 27:191–212.

Grün, B. (2019).
Model-based clustering.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 8, pages 157–192. CRC Press, Boca Raton, FL.

Hamilton, J. D. (1988).

Rational expectations econometric analysis of changes in regime: An investigation on the term structure of interest rates.

Journal of Economic Dynamics and Control, 12:385–423.

Hamilton, J. D. (1989).

A new approach to the economic analysis of nonstationary time series and the business cycle.

Econometrica, 57:357–384.

Hamilton, J. D. and Susmel, R. (1994).

Autoregressive conditional heteroskedasticity and changes in regime.

Journal of Econometrics, 64:307–333.

Hennig, C. (2010).

Methods for merging Gaussian mixture components.

Advances in Data Analysis and Classification, 4:3–34.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005).

Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling.

Statistical Science, 20:50–67.

Kaufmann, S. (2019).
Hidden Markov models in time series, with applications in economics.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 13, pages 308–341. CRC Press, Boca Raton, FL.

Kaufmann, S. and Frühwirth-Schnatter, S. (2002).
Bayesian analysis of switching ARCH models.
Journal of Time Series Analysis, 23:425–458.

Keribin, C. (2000).
Consistent estimation of the order of mixture models.
Sankhyā A, 62:49–66.

Kiefer, N. M. and Wolfowitz, J. (1956).
Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters.
Annals of Mathematical Statistics, 27:887–906.

Kim, S., Shephard, N., and Chib, S. (1998).
Stochastic volatility: Likelihood inference and comparison with ARCH models.
Review of Economic Studies, 65:361–393.

Kon, S. J. (1984).
Models of stock returns – A comparison.
The Journal of Finance, 39:147–165.

Lamoureux, C. G. and Lastrapes, W. D. (1990).
Persistence in variance, structural change and the GARCH model.
Journal of Business & Economic Statistics, 8:225–234.

Lee, S. X. and McLachlan, G. J. (2013).
EMMIXuskew: An R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm.
Journal of Statistical Software, 55(12):1–22.

Leisch, F. (2004).
Exploring the structure of mixture model components.
In Antoch, J., editor, *COMPSTAT 2004. Proceedings in Computational Statistics*, pages 1405–1412. Physica-Verlag/Springer, Heidelberg.

Li, J. (2005).
Clustering based on a multi-layer mixture model.
Journal of Computational and Graphical Statistics, 14:547–568.

MacQueen, J. (1967).
Some methods for classification and analysis of multivariate observations.
In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 281–297.

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

Malsiner Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017).
Identifying mixtures of mixtures using Bayesian estimation.
Journal of Computational and Graphical Statistics, 26:285–295.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005).
Bayesian modelling and inference on mixtures of distributions.
In Dey, D. and Rao, C., editors, *Bayesian Thinking, Modelling and Computation*, volume 25 of *Handbook of Statistics*, chapter 16, page ADD. North-Holland, Amsterdam.

McCulloch, R. E. and Tsay, R. S. (1994).
Statistical analysis of economic time series via Markov switching models.
Journal of Time Series Analysis, 15:523–539.

McLachlan, G. J. and Peel, D. (2000).
Finite Mixture Models.
Wiley Series in Probability and Statistics. Wiley, New York.

McQueen, G. and Thorely, S. (1991).
Are stock returns predictable? A test using Markov chains.
The Journal of Finance, 46:239–263.

Melnykov, V. (2016).
Merging mixture components for clustering through pairwise overlap.
Journal of Computational and Graphical Statistics, 25:66–90.

Meng, X.-L. and Schilling, S. (1996).
Fitting full-information item factor models and an empirical investigation of bridge sampling.
Journal of the American Statistical Association, 91:1254–1267.

Meng, X.-L. and Wong, W. H. (1996).
Simulating ratios of normalizing constants via a simple identity: A theoretical exploration.
Statistica Sinica, 6:831–860.

Miller, J. W. and Harrison, M. T. (2018).

Mixture models with a prior on the number of components.
Journal of the American Statistical Association, 113:340–356.

Neftçi, S. N. (1984).
Are economic time series asymmetric over the business cycle?
Journal of Political Economy, 92:307–328.

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

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

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007).
Stochastic volatility with leverage: Fast and efficient likelihood inference.
Journal of Econometrics, 140:425–449.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996).
Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition.

Journal of the American Statistical Association, 91:953–960.

Polson, N. G., Scott, J. G., and Windle, J. (2013).
Bayesian inference for logistic models using Pólya-Gamma latent variables.
Journal of the American Statistical Association, 108:1339–49.

Poskitt, D. S. and Chung, S.-H. (1996).
Markov chain models, time series analysis and extreme value theory.
Advances in Applied Probability, 28:405–425.

Ray, S. and Lindsay, B. (2005).
The topography of multivariate normal mixtures.
The Annals of Statistics, 33:2042–2065.

Richardson, S. and Green, P. J. (1997).
On Bayesian analysis of mixtures with an unknown number of components.
Journal of the Royal Statistical Society, Ser. B, 59:731–792.

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

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998).
Stylized facts of daily return series and the hidden Markov model.
Journal of Applied Econometrics, 13:217–244.

Sclove, S. L. (1983).
Time series segmentation: A model and a method.
Information Science, 29:7–25.

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

Shephard, N. (1994).
Partial non-Gaussian state space.
Biometrika, 81:115–131.

So, M. K. P., Lam, K., and Li, W. K. (1998).
A stochastic volatility model with Markov switching.
Journal of Business & Economic Statistics, 16:244–253.

Sorenson, H. W. and Alspach, D. L. (1971).

Recursive Bayesian estimation using Gaussian sums.

Automatica, 7:465–479.

Sperrin, M., Jaki, T., and Wit, E. (2010).

Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models.

Statistics and Computing, 20:357–366.

Spezia, L. (2009).

Reversible jump and the label switching problem in hidden Markov models.

Journal of Statistical Planning and Inference, 139:2305–2315.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002).

Bayesian measures of model complexity and fit.

Journal of the Royal Statistical Society, Ser. B, 64:583–639.

Stephens, M. (2000a).

Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods.

The Annals of Statistics, 28:40–74.

Stephens, M. (2000b).

Dealing with label switching in mixture models.

Journal of the Royal Statistical Society, Ser. B, 62:795–809.

Symons, M. J. (1981).

Clustering criteria and multivariate normal mixtures.

Biometrics, 37:35–43.

Timmermann, A. (2000).

Moments of Markov switching models.

Journal of Econometrics, 96:75–111.

Tucker, A. (1992).

A reexamination of finite- and infinite-variance distributions as models of daily stock returns.

Journal of Business & Economic Statistics, 10:73–81.

Turner, C. M., Startz, R., and Nelson, C. R. (1989).

A Markov model of heteroscedasticity, risk, and learning in the stock market.

Journal of Financial Economics, 25:3–22.

CHECK.

Wilson, R., Bienias, J., Evans, D., and Bennett, D. (2004).

The Religious Orders Study: Overview and Change in Cognitive and Motor Speed.
Aging, Neuropsychol, Cogn., 11:280–303.

Wolfe, J. H. (1970).
Pattern clustering by multivariate mixture analysis.
Multivariate Behavioral Research, 5:329–350.
CHECK.

Yerebakan, H. Z., Rajwa, B., and Dundar, M. (2014).
The infinite mixture of infinite Gaussian mixtures.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27 of *Proceedings from the Neural Information Processing Systems Conference*.