WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Finite Mixture and Markov Switching Models

Sylvia Frühwirth-Schnatter

Western Swiss Doctoral School in Statistics and Probability, February 2020

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

Finite mixture distributions

Density of a finite mixture distribution

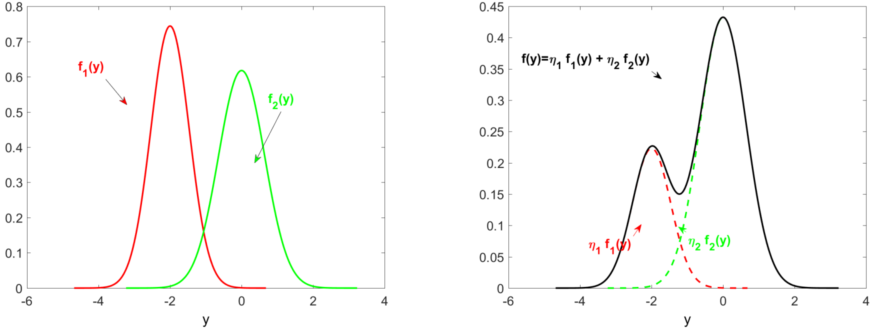
The density of a finite mixture distribution is defined by

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

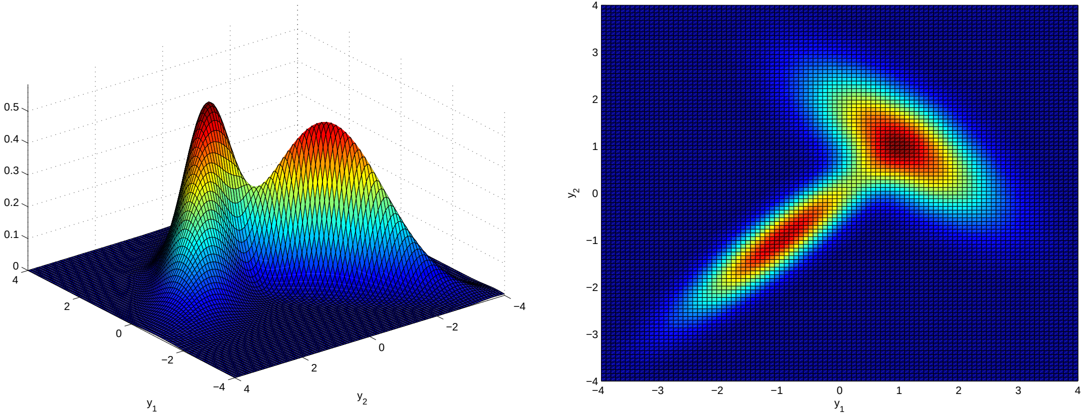
- K is the number of components;
- $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ is the weight distribution with $\eta_k \geq 0$, $\sum_{k=1}^{K} \eta_k = 1$;
- the component densities $f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k)$ arise from the same distribution family $\mathcal{T}(\boldsymbol{\theta})$;
- $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ vary over the components;
- $\mathbf{y}$ can be univariate or multivariate, continuous, discrete-valued, mixed-type, time series data, outcomes of a regression model, ...

Illustration

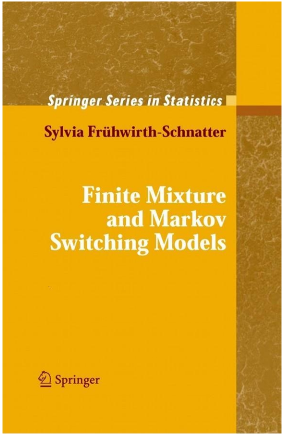
- Define a mixture of $K = 2$ distributions with Gaussian components densities
- $f_1(y) = f_{\mathcal{N}}(y; -2, 1)$ and $f_2(y) = f_{\mathcal{N}}(y; 0, 2)$,
- and weights $\eta_1 = 0.3$ and $\eta_2 = 0.7$.



Mixture of two bivariate normal distributions



For more details see . . .



2006

2019

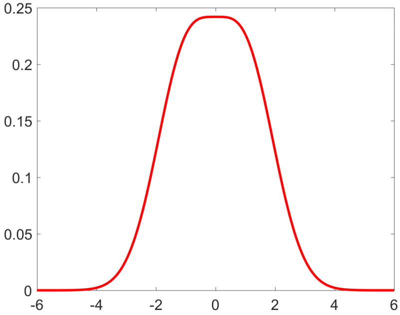
Practical relevance of finite mixture models

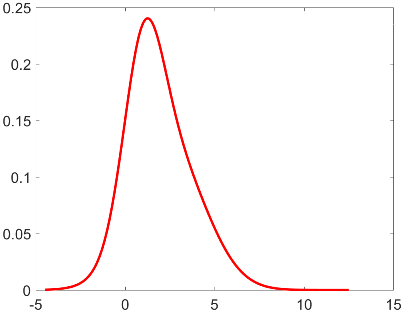
Finite mixture distributions are useful for

- ***Density estimation:*** capture many specific properties of real data such as multimodality, skewness, and kurtosis
- ***Flexible modelling:*** deal in a natural way with special issues such as non-normality and unobserved heterogeneity
- ***Model-based clustering:*** arise as marginal distribution of models for unsupervised clustering

Density approximation based on finite mixtures

Finite mixture of normal distributions are very useful for flexible modelling of non-Gaussian densities





$$0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(1,1) \quad 0.4\mathcal{N}(1,1.2) + 0.6\mathcal{N}(2.5,4)$$

Approximation Property

- Let $g(y)$ be an arbitrary probability density function.
- Let $q_K(y)$ be a mixture of normals:

$$q_K(y) = \sum_{r=1}^{K} w_r f_N(y; m_r, s_r^2).$$

- For increasing K, the distance between $g(y)$ and $q_K(y)$, e.g. the Kullback-Leibler distance

$$\int_{\Re} g(y) \log \frac{g(y)}{q_K(y)} dy$$

can be made arbitrarily small.

Approximation Property

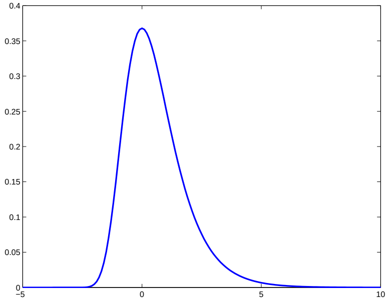
- To approximate $g(y)$ for a fixed K, select
  - the weights $w_1, \dots, w_K$,
  - the means $m_1, \dots, m_K$,
  - and the variances $s_1^2, \dots, s_K^2$,

  such that the distance between $g(y)$ and $q_K(y)$ is minimized.
- This is not a parameter estimation problem.
- This is a problem of numerical optimization.

Example

- Consider the density the type I extreme value distribution:

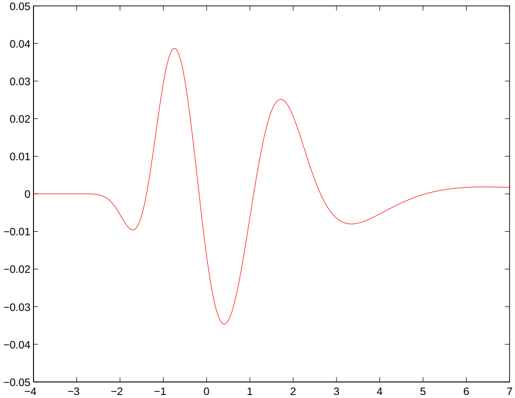
$$g(y) = \exp(-y - e^{-y}).$$



- This is also the density of the random variable $-\log Y$, where $Y \sim \mathcal{E}(1)$ follows the standard exponential distribution.

Approximation for $K = 2$

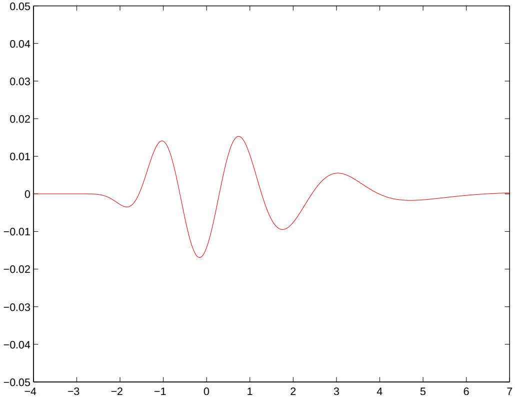
Optimal 2 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 3$

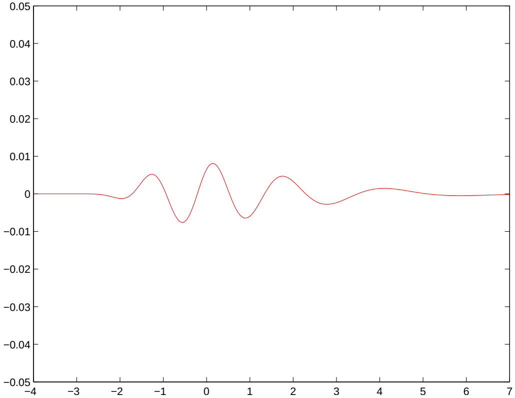
Optimal 3 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 4$

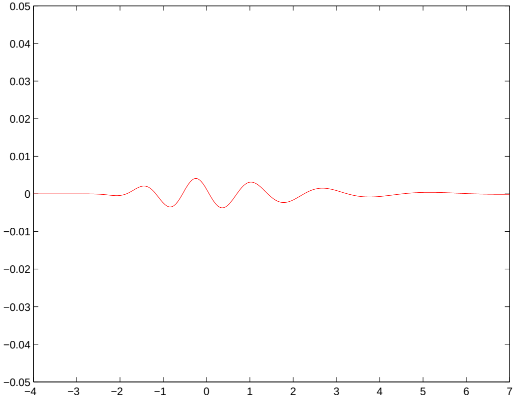
Optimal 4 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 5$

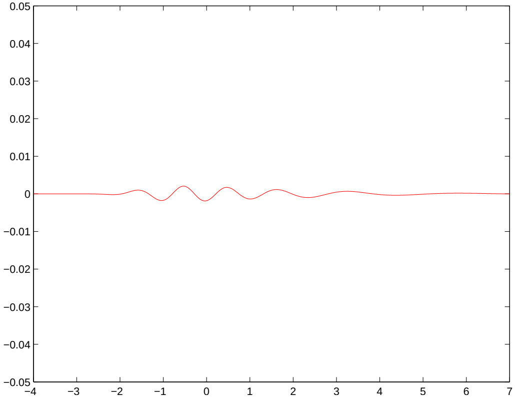
Optimal 5 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 6$

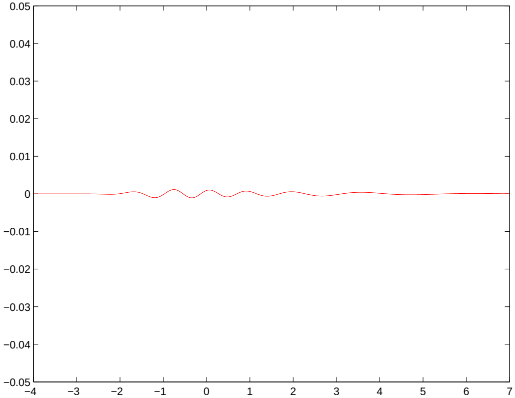
Optimal 6 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 7$

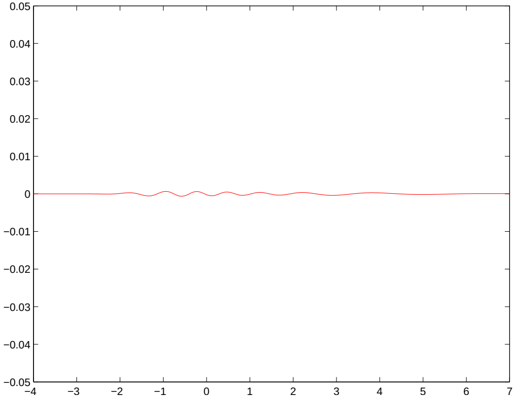
Optimal 7 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 8$

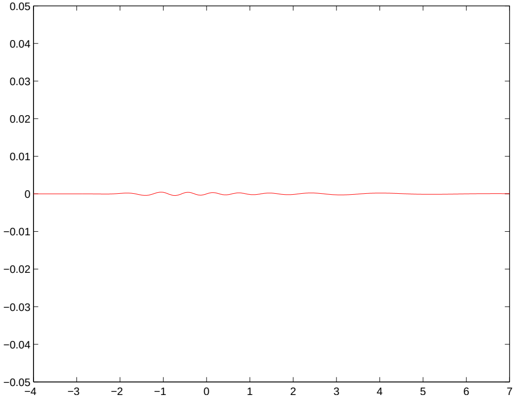
Optimal 8 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 9$

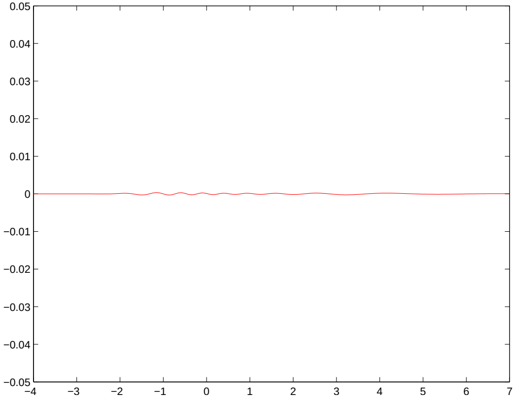
Optimal 9 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Approximation for $K = 10$

Optimal 10 component mixture approximation



Kernel of the KL distance $g(y) \log \frac{g(y)}{q_K(y)}$

Density Approximation for $K = 10$

Approximate the non-normal density $g(y)$ by a normal mixture of 10 components with parameters m_r and s_r and weight w_r for the rth component:

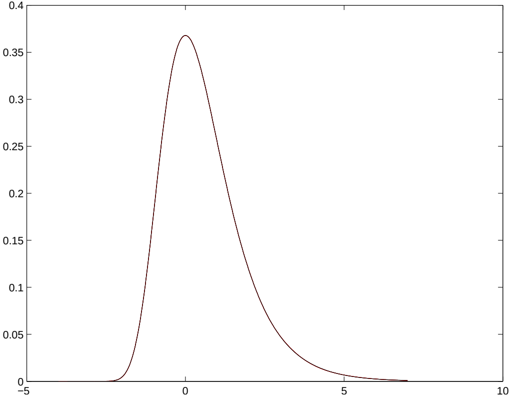
$$g(y) = \exp\{-y - e^{-y}\} \approx q_{10}(y) = \sum_{r=1}^{10} w_r f_N(y; m_r, s_r^2).$$

The mixture was estimated in [Frühwirth-Schnatter and Frühwirth, 2007] by minimizing the Kullback-Leibler distance of the estimated mixture from the exact density:

w_r	0.00397	0.0396	0.168	0.147	0.125	0.101	0.104	0.116	0.107	0.088
m_r	5.09	3.29	1.82	1.24	0.764	0.391	0.0431	-0.306	-0.673	-1.06
s_r^2	4.5	2.02	1.1	0.422	0.198	0.107	0.0778	0.0766	0.0947	0.146

Density Approximation for $K = 10$

The mixture approximation to the density of the type I extreme value distribution



Bayesian Computation Based on Finite Mixture Approximations

- Gaussian mixtures are useful for developing simple estimation procedures for non-normal models [Sorenson and Alspach, 1971, Alspach and Sorenson, 1972]
- Stochastic volatility modelling: [Shephard, 1994], [Kim et al., 1998] and [Chib et al., 2002] use a 7 component normal mixture approximation of the density of the log of a χ^2_1-distributed random variable, improved by [Omori et al., 2007]
- Spectral analysis: [Carter and Kohn, 1997] use a 5 component normal mixture approximation of the density of the log of an $\mathcal{E}(1)$-distributed random variable
- Non-Gaussian models: [Frühwirth-Schnatter and Wagner, 2006] and [Frühwirth-Schnatter and Frühwirth, 2007] use a 10 component normal mixture approximation of the density of minus log of an $\mathcal{E}(1)$-distributed random variable

Outline

Part I: Finite Mixture Models and Model-based Clustering

Unsupervised Clustering

- **Group previously unstructured data** into groups which contain observations that are similar in some sense
- The investigator expects that there exist meaningful subcategories of the data under investigation, however, there are no external criterion by which to define these groups
- The investigator relies on an **internal criterion** and is willing to **let the data speak** (suggest sensible clusters)
- Many clustering criteria have been developed over the past decades for cross sectional data, much less so for time series data

Why is unsupervised clustering difficult?

- Assume that N subjects should be grouped into K clusters.
- Find an **optimal** partition among all possible partitions $\mathbf{S} = (S_1, \dots, S_N)$, where $S_i \in \{1, \dots, K\}$.
- Search in the rather large space $\mathcal{I} = \bigotimes_{i=1}^{N} \{1, \dots, K\}$, increasing rapidly with the number of subjects N and the number of clusters K:
  - $N = 10$, $K = 3$: 59049 different allocations
  - $N = 100$, $K = 3$: roughly $5 \cdot 10^{47}$ different allocations
- Exploring this large space is challenging; **there are simply too many possibilities**.

Challenges in cluster analysis

[Everitt, 1979]:

- Selecting a suitable **clustering criterion**
- **Computational issues** (identifying a sensible search strategy for the latent allocations, choosing sensible starting values)
- Selecting the **number of clusters**
- Review: [Grün, 2019]

Common statistical cluster technique

- **Heuristic clustering techniques:**
  - based on distance measures, e.g. such as *k*-means [MacQueen, 1967]
  - difficult to extend to discrete data, time series and other complex data structures
- **Model based clustering:**
  - based on finite mixture models [Banfield and Raftery, 1993, Bensmail et al., 1997, Dasgupta and Raftery, 1998, Fraley and Raftery, 2002]
  - much easier to extend to discrete data, time series and complex data structures

Clustering based on Finite mixtures

- Consider a population involving two latent clusters:
  - ***Cluster 1*** ($S_i = 1$), $\mathrm{Pr}(S_i = 1) = \eta_1$ (cluster size):

$$p(\mathbf{y}_i|S_i = 1) = f_N(\mathbf{y}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

  - ***Cluster 2*** ($S_i = 2$), $\mathrm{Pr}(S_i = 2) = \eta_2 = 1 - \eta_1$ (cluster size):

$$p(\mathbf{y}_i|S_i = 2) = f_N(\mathbf{y}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

Marginal distribution

The marginal distribution of $\mathbf{y}_i$ is a mixture distribution:

$$p(\mathbf{y}_i) = \eta_1 f_N(\mathbf{y}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \eta_2 f_N(\mathbf{y}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

Cluster Analysis Based on Mixtures of Normals

Multivariate mixtures of normals distributions

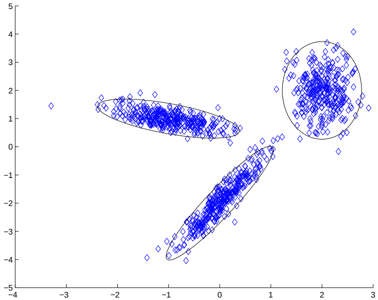
For a vector $\mathbf{y}_i$ with metric features $y_{ij}, j = 1, \ldots, r$, a particular useful models are multivariate mixture of normals distributions:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \eta_1 f_N(\mathbf{y}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \ldots + \eta_K f_N(\mathbf{y}_i; \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K),$$

- Clustering kernel $f_N(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the density of a multivariate normal distribution with cluster-specific mean $\boldsymbol{\mu}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}_k$.
- Seminal papers: [Wolfe, 1970], [Scott and Symons, 1971], [Symons, 1981], [Binder, 1978], [Banfield and Raftery, 1993]

Heterogeneous Mixtures of Normals

Different variance-covariance matrices in the different groups



500 observations from a three-component mixture of heterogeneous bivariate normal distributions

Bayes' classification

- In general, a finite mixture distribution is defined by

$$p(\mathbf{y}) = \eta_1 p(\mathbf{y}|\boldsymbol{\theta}_1) + \cdots + \eta_K p(\mathbf{y}|\boldsymbol{\theta}_K),$$

where $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is the pdf of the distribution in the kth component.

- The finite mixture distribution allows classification of each observation $\mathbf{y}_i$ conditional on knowing $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \eta_1, \dots, \eta_K)$:

Classification of $\mathbf{y}_i$ for fixed $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \eta_1, \dots, \eta_K)$

$$\mathrm{Pr}(S_i = k|\boldsymbol{\vartheta}, \mathbf{y}_i) \propto p(\mathbf{y}_i|\boldsymbol{\vartheta}, S_i = k)\mathrm{Pr}(S_i = k|\boldsymbol{\vartheta}) \propto p(\mathbf{y}_i|\boldsymbol{\theta}_k)\eta_k, \qquad \forall k = 1, \dots, K$$

- The component density $p(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is essential for classification.
- It is called **clustering kernel** in the context of model-based clustering.

Relation to Other Clustering Approaches

- [Scott and Symons, 1971] realized that Bayesian maximum aposteriori classification using certain types of multivariate mixtures of normal distributions is related to common clustering criteria:
  - isotropic mixtures with $\boldsymbol{\Sigma}_k \equiv \sigma^2 \mathbf{I}_r$ are equivalent to minimizing $\mathrm{tr}\,(\boldsymbol{W}(\mathbf{S}))$,
  - homogeneous mixture with $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ are equivalent to minimizing $|\boldsymbol{W}(\mathbf{S})|$,
- where

$$\boldsymbol{W}(\mathbf{S}) = \sum_{k=1}^{K} \boldsymbol{W}_k(\mathbf{S}),$$

$$\boldsymbol{W}_k(\mathbf{S}) = \sum_{i:S_i=k} (\mathbf{y}_i - \overline{\mathbf{y}}_k)(\mathbf{y}_i - \overline{\mathbf{y}}_k)', \quad \overline{\mathbf{y}}_k = \frac{1}{N_k} \sum_{i:S_i=k} \mathbf{y}_i.$$

Why is this relation important?

- Sensible clustering criteria are obtained by deriving the optimal classification for a mixture model from a certain distribution.
- This relation is helpful because:
  - it reduces the problem of choosing a certain clustering criteria to a model choice problem within a well-defined probabilistic framework.
  - it shows how to carry out clustering for more general data (discrete-valued data, times series, ...)
- It has been noted in several empirical studies, that
  - the $\mathrm{tr}\,(\boldsymbol{W}(\mathbf{S}))$ criterion imposes an spherical structure on the grouping even if the true groups are of different shape,
  - the $|\boldsymbol{W}(\mathbf{S})|$ allows for elliptical clusters.
- ***The clustering kernel has to capture salient feature of the observed data.***

More general mixtures

- The **idea of model-based clustering is very generic** - can be easily extended to more general clustering kernels
- **Finite mixture for discrete-valued data:**
  - Poisson and negative binomial mixture for count data;
  - latent class models for multivariate binary data
- **Finite mixtures of skew-N and skew-t distributions**: recent research demonstrates the usefulness of parametric non-Gaussian component distributions
- finite mixtures of **GLM regression models**
- clustering (discrete-valued) **time series**

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

The Bayesian Approach toward Estimation

- Many authors used a **Bayesian approach** to estimate finite mixtures
- Joint parameter estimation and classification is easily implemented using Markov chain Monte Carlo (MCMC) methods [Diebolt and Robert, 1994]
- Inference is possible for interesting, possibly non-linear functionals of the parameters
- The prior distribution regularizes the likelihood function
- see, e.g., [Celeux et al., 2000]

Problems with the likelihood function

- Consider a univariate normal mixture with two components:

$$p(y_i|\mu_2, \sigma_2^2) = \eta_1 f_N(y_i; \mu_1, \sigma_1^2) + (1 - \eta_1) f_N(y_i; \mu_2, \sigma_2^2),$$

  - μ_1, σ_1^2 and η_1 are known;
  - μ_2 and σ_2^2 are unknown.
- Whenever $\mu_2 = y_i$ (where y_i is any of the observed values):

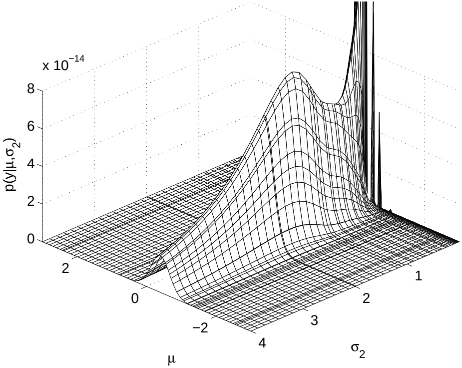
$$p(y_i|\mu_2 = y_i, \sigma_2^2) = c_{i1} + \frac{1 - \eta_1}{\sqrt{2\pi\sigma_2^2}}, \quad c_{i1} = \eta_1 f_N(y_i; \mu_1, \sigma_1^2),$$

$$\lim_{\sigma_2^2 \to 0} p(y_1, \ldots, y_N|\mu_2 = y_i, \sigma_2^2) = \infty.$$

- Hence, the likelihood function has many spurious modes close to 0 [Kiefer and Wolfowitz, 1956].

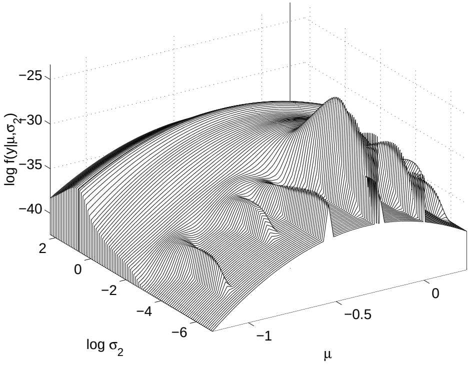
The observed-data likelihood function is unbounded

Surface plot of the observed-data likelihood function $\log p(y_1, \ldots, y_N | \mu_2, \sigma_2)$ ($\mu_2^{\text{true}} = 0$, $\sigma_2^{\text{true}} = 2$)



Zooming in . . .

Zooming into very small variances



Regularization of the observed-data likelihood

- Don't let the component specific variances σ_2^2 become too small.
- Add the "regularization" prior $1/\sigma_2^2 \sim \mathcal{G}(c_0, C_0)$ with $C_0 > 0$:

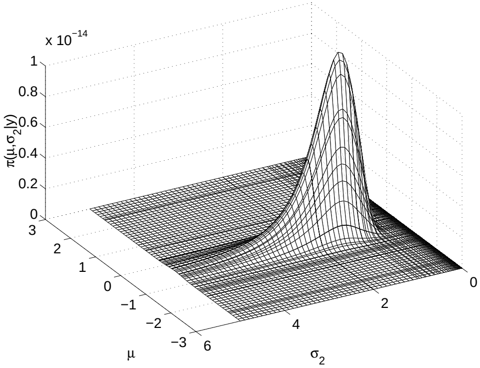
$$p(y_i|\mu_2 = y_i, \sigma_2^2)p(\sigma_2^2) \propto \left(c_{i1} + \frac{1-\eta_1}{\sqrt{2\pi\sigma_2^2}}\right)\left(\frac{1}{\sigma_2^2}\right)^{c_0+1} \exp(-\frac{C_0}{\sigma_2^2}).$$

- Penalizes the likelihood as $\sigma_2^2 \to 0$:

$$\lim_{\sigma_2^2 \to 0} p(y_1, \ldots, y_N|\mu_2 = y_i, \sigma_2^2) = 0.$$

Regularized likelihood function

Posterior density (regularized likelihood function) $p(\mu_2, \sigma_2 | y_1, \dots, y_N)$ under the prior $1/\sigma_2^2 \sim \mathcal{G}(1, 4)$



MCMC Estimation

Following [Diebolt and Robert, 1994], the most popular method for Bayesian estimation of finite mixtures is to apply Markov chain Monte Carlo methods:

- **Data augmentation –** introduce the sequence of hidden indicators $\mathbf{S} = (S_1, \ldots, S_N)$ as latent variables
- **Gibbs sampling –** repeat the following sampling steps:
  (a) "Estimation for a known grouping": sample the component specific parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and the weight distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ conditional on knowing $\mathbf{S}$ and the data.
  (b) "Classification for known parameters": sample the hidden indicators $\mathbf{S} = (S_1, \ldots, S_N)$ conditional on knowing $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and $\boldsymbol{\eta}$.

  See [Frühwirth-Schnatter, 2006], Section 3.5 for an extensive review.

Choosing priors

- Dirichlet distribution on the weight distribution $\boldsymbol{\eta} \sim \mathcal{D}\left(e_1, \ldots, e_K\right)$;
- Conditionally conjugate priors on $\boldsymbol{\theta}_k|\psi$: step [(a)] in one sweep
- Conditionally non-conjugate priors on $\boldsymbol{\theta}_k|\psi$: step [(a)] in two sweeps
- Hierarchical prior $\psi \sim p(\psi)$

The Label Switching problem

- A mixture distribution is invariant to reordering the components, e.g. for $K = 3$:

$$
\begin{aligned}
p(\mathbf{y}) &= \eta_1 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_1) + \eta_2 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_2) + \eta_3 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_3) \\
&= \eta_3 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_3) + \eta_1 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_1) + \eta_2 f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_2).
\end{aligned} \tag{1}
$$

- But so is an estimated mixture with component -specific parameters $(\hat{\eta}_k, \hat{\boldsymbol{\theta}}_k)$, e.g. for $K = 3$:

$$
\begin{aligned}
p(\mathbf{y}) &= \hat{\eta}_1 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1) + \hat{\eta}_2 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2) + \hat{\eta}_3 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_3) \\
&= \hat{\eta}_3 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_3) + \hat{\eta}_1 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1) + \hat{\eta}_2 f_{\mathcal{T}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2).
\end{aligned} \tag{2}
$$

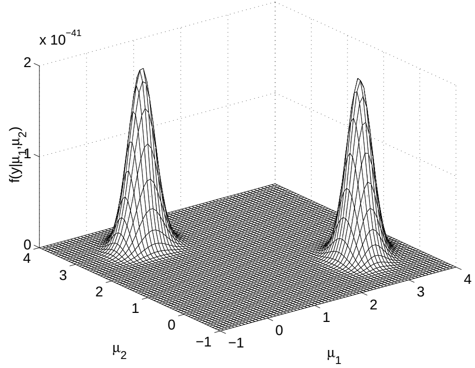
- There is no reason why the numbering in (1) and (2) should be the same.

The label switching problem

- Relabeling the states of the hidden indicator $\mathbf{S}$ leaves the observed-data likelihood function unchanged.
- This causes multi-modality; the observed-data likelihood function is multimodal with at most $K!$ modes.
- For a symmetric prior distribution, the posterior distribution is symmetric and multimodal.
- When sampling from the (unconstrained) posterior via MCMC methods you do not know which component of the sampled parameter correspond to which group and label switching might occur.

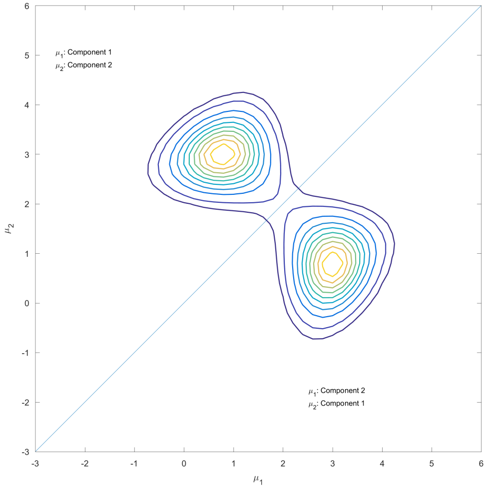
Invariance of the observed-data likelihood function

Observed-data likelihood function $p(\mathbf{y}|\mu_1, \mu_2)$ (simulated data with $\mu_1 = 0$ and $\mu_2 = 3$)



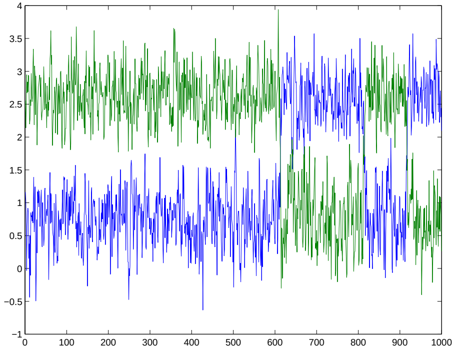
Invariance of the posterior

Contour plots of unconstrained posterior $p(\mu_1, \mu_2|\mathbf{y})$ for the simulated data



Label switching in the MCMC output

MCMC draws from $p(\mu_1, \mu_2|\mathbf{y})$ for the simulated data

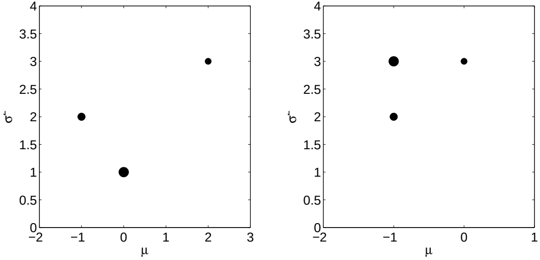


Dealing with the label switching problem

- Let the component specific parameter $\boldsymbol{\theta}_k$ take values in Θ.
- Relabel the draws $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ of a mixture with K components
- Most papers work in the *full parameter space* Θ^K to identify suitable permutations of the labels
  [Celeux, 1998, Celeux et al., 2000, Stephens, 2000b, Marin et al., 2005, Jasra et al., 2005, Nobile and Fearnside, 2007, Sperrin et al., 2010, Spezia, 2009]
- "Simple" relabeling [Frühwirth-Schnatter, 2001b]
  - operates in Θ *or even a subspace* $\tilde{\Theta} \subset \Theta$
  - Clustering in the point process representation

Point Process Representation of a Finite Mixture Model

- Any finite mixture distribution has a representation as marked point process and may be seen as a distribution of the points $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ over the parameter space Θ [Stephens, 2000a]
- Point process representation of univariate normal mixtures with 3 components

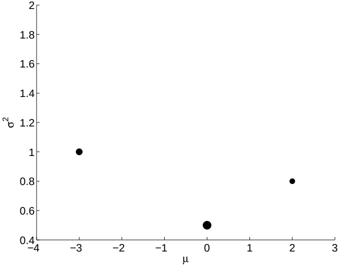


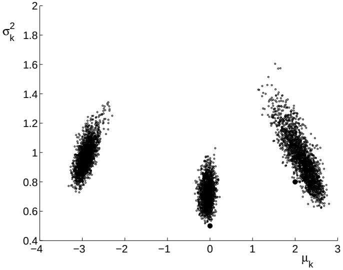
Labelling Based on the Point Process Representation

- [Frühwirth-Schnatter, 2001b] suggested to use the point process representation of the MCMC draws to identify a mixture model.
- The MCMC draws scatter around the points corresponding to the “true” point process representation
- A visual inspection of these plots allows to study the difference in the component specific parameters and to formulate an identifiability constraint. This works well in lower dimensions.
- In higher dimensional problems, heuristic cluster methods such as k-means are used.

Exploring the point process representation

- Example: mixture of three univariate normal distributions with $\eta_1 = 0.3$, $\eta_2 = 0.5$, $K = 3$, $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.8$





- The MCMC draws scatter around the points corresponding to the "true" point process representation
- The spread of the clouds representing the uncertainty of estimating the parameters of the mixture

Point process representations in higher dimensions

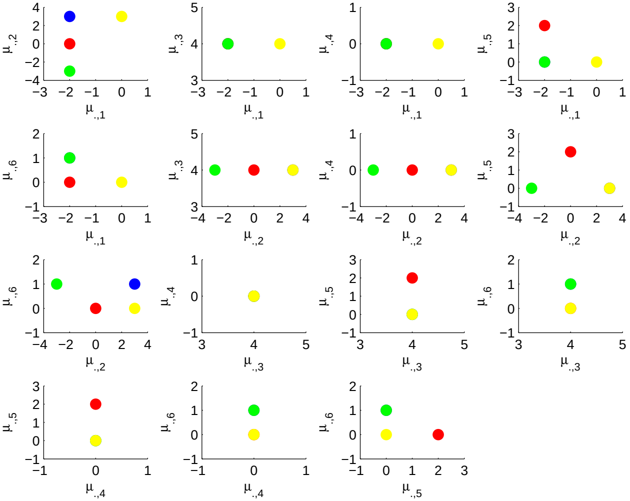
Consider following mixture of 4 multivariate normals of dimension $r = 6$ with

$$\begin{pmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \boldsymbol{\mu}_3 & \boldsymbol{\mu}_4 \end{pmatrix} = \begin{pmatrix} -2 & -2 & -2 & 0 \\ 3 & 0 & -3 & 3 \\ 4 & 4 & 4 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix},$$

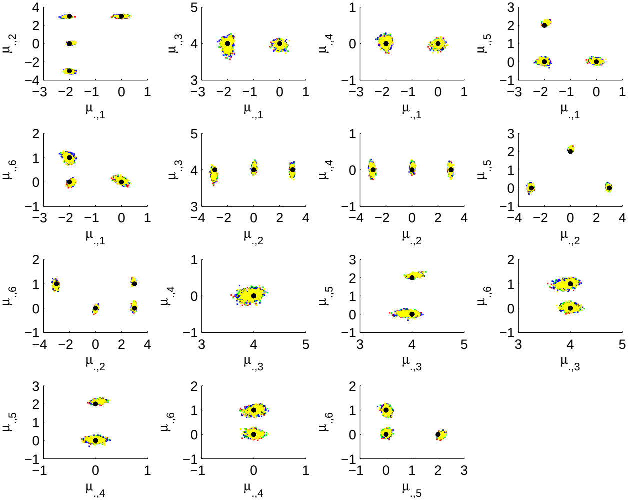
$$\boldsymbol{\Sigma}_1 = 0.5\mathbf{I}_r,\ \boldsymbol{\Sigma}_2 = 4\mathbf{I}_r + 0.2\mathbf{e}_r,\ \boldsymbol{\Sigma}_3 = 4\mathbf{I}_r - 0.2\mathbf{e}_r,\ \boldsymbol{\Sigma}_4 = \mathbf{I}_r.$$

$\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathrm{vec}(\boldsymbol{\Sigma}))$ contains $r + r(r+1)/2 = 27$ coefficients.

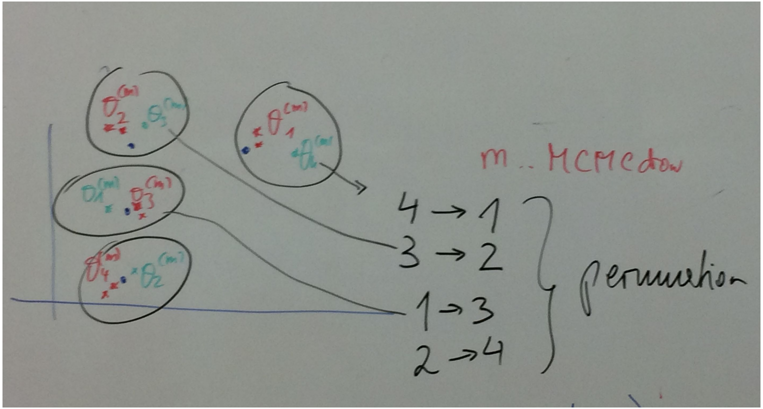
Two-dimensional projections of the point process representation



Point process representation of 5000 draws (1000 observations)



Clustering in the Point Process Representation



Clustering in the Point Process Representation

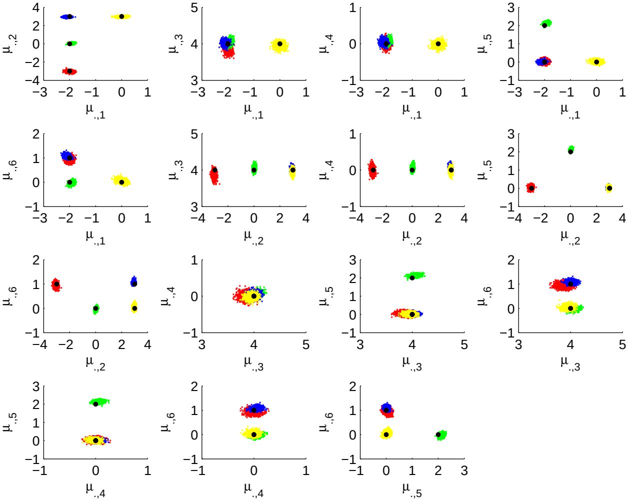
Labeling through k-means clustering in the point process representation of the MCMC draws

- Apply k-means clustering to all KM posterior draws of the parameter vector $\boldsymbol{\theta}_k^{(m)}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$.
- This delivers a classification index $I_k^{(m)} \in \{1, \ldots, K\}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$.
- Check, if $\rho_m = (I_1^{(m)}, \ldots, I_K^{(m)})$ is a permutation of $\{1, \ldots, K\}$.
- In this case, a unique labelling is achieved by reordering the draws through ρ_m:
  - (c1) $\eta_1^{(m)}, \ldots, \eta_K^{(m)}$ is substituted by $\eta_{\rho_m^{-1}(1)}^{(m)}, \ldots, \eta_{\rho_m^{-1}(K)}^{(m)}$;
  - (c2) $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ is substituted by $\boldsymbol{\theta}_{\rho_m^{-1}(1)}^{(m)}, \ldots, \boldsymbol{\theta}_{\rho_m^{-1}(K)}^{(m)}$;
  - (c3) $S_1^{(m)}, \ldots, S_N^{(m)}$ is substituted by $\rho_m(S_1^{(m)}), \ldots, \rho_m(S_N^{(m)})$.
- Remove draws, where ρ_m is not a permutation.

Application to the Example

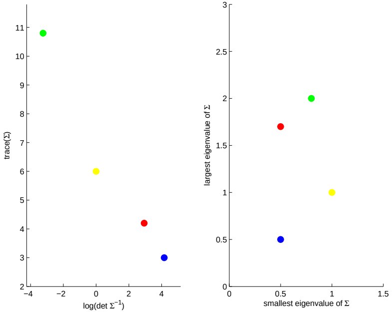
- Component specific parameter $\boldsymbol{\theta}_k$ contains $r + r(r+1)/2 = 27$ coefficients.
- Use only the component mean, i.e. $\boldsymbol{\theta}_k = (\mu_{k,1} \cdots \mu_{k,r})'$; $\boldsymbol{\theta}_k$ contains 6 elements.
- k-means clustering identifies 4 clusters in $MK = 20\,000$ realizations of the 6-dimensional variable $\boldsymbol{\theta}_k^{(m)}$.
- For each $\boldsymbol{\theta}_k^{(m)}$ a classification index $I_k^{(m)}$ results.
- All classification sequences $\rho_m = (I_1^{(m)}, \ldots, I_4^{(m)})$, $m = 1, \ldots, M$ turned out to be permutations of $\{1, \ldots, 4\}$.

Point process representation of 5000 *identified* MCMC draws



Application to the Example

- It is usually sufficient to consider a subset of the components-specific parameters to obtain those classification indices.
- One could add measures describing $\boldsymbol{\Sigma}_k$, e.g. $\mathrm{Diag}\left(\boldsymbol{\Sigma}_k\right)$, $|\boldsymbol{\Sigma}_k|$, or eigenvalues of $\boldsymbol{\Sigma}_k$.



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*.