LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Bettina Grün & Friedrich Leisch

Dealing with label switching in mixture models under genuine multimodality

Technical Report Number 039, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Dealing with label switching in mixture models under genuine multimodality

Bettina Grün [a,*], Friedrich Leisch [b]

[a] *Department für Statistik und Mathematik, Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Wien, Austria*

[b] *Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, D-80539 München, Germany*

This is a preprint of an article which has been accepted for publication in the *Journal of Multivariate Analysis*. The original publication is available at www.elsevier.com/locate/jmva.

DOI: 10.1016/j.jmva.2008.09.006

Abstract

The fitting of finite mixture models is an ill-defined estimation problem as completely different parameterizations can induce similar mixture distributions. This leads to multiple modes in the likelihood which is a problem for frequentist maximum likelihood estimation, and complicates statistical inference of Markov chain Monte Carlo draws in Bayesian estimation. For the analysis of the posterior density of these draws a suitable separation into different modes is desirable. In addition, a unique labelling of the component specific estimates is necessary to solve the label switching problem. This paper presents and compares two approaches to achieve these goals: relabelling under multimodality and constrained clustering. The algorithmic details are discussed and their application is demonstrated on artificial and real-world data.

Key words: constrained clustering, finite mixture models, label switching, multimodality

[*] Corresponding author. tel. +43 1 31336 5032; fax +43 1 31336 774
Email addresses: Bettina.Gruen@wu-wien.ac.at (Bettina Grün), Friedrich.Leisch@stat.uni-muenchen.de (Friedrich Leisch).

1 Introduction

The use of finite mixtures in applications has increased in popularity in the last decades because maximum likelihood estimation has been facilitated by the expectation-maximization (EM) algorithm [1] and Bayesian estimation of finite mixture models has become feasible with the advent of Markov chain Monte Carlo (MCMC) methods. Gibbs sampling is the most commonly used approach and it is done by augmenting the data with the unobservable variable of class membership similar to the EM algorithm [2]. For symmetric priors and components from the same distributional family label switching [3] makes it impossible to make component specific inference directly from the MCMC draws. Different approaches have been proposed to determine suitable estimates. These recent developments have led to several overviews on sampling schemes for mixture models and handling label switching problems [4–6]. The methods proposed include specification of an (artificial) ordering constraint [2,7], fixing the membership of some observations [8], applying label-invariant loss functions, cluster and relabelling algorithms [9–11] and relabelling with respect to the maximum a posteriori (MAP) estimate [4].

All the proposed approaches and their illustrations focus on the case where "no genuine multimodality" (see Section 2) of the posterior density is given, i.e., it is assumed that the modes of the posterior density are identical up to a permutation of the components. However, multiple genuine modes can occur due to the fact that fitting finite mixtures to data is an ill-conditioned problem and similar mixture distributions may result from different parameterizations, e.g. due to unidentifiability of the mixture distribution. Finite mixture models do in fact not only suffer from trivial identifiability problems due to label switching and empty or identical components, but also from generic identifiability problems [6]. A generic identifiability problem for mixtures of regressions is for example intra-component label switching which occurs if different combinations of the components between the covariate points determine the same mixture distribution due to the violation of the coverage condition on the covariate matrix [12]. Up to our knowledge, only Stephens [11] outlines an approach where the possibility of multiple genuine modes of the posteriors is taken into account, and component specific estimates for each of the modes are determined.

In this paper a new method for determining a suitable labelling of the components under genuine multimodality is proposed. It allows to make component specific inference for each mode separately using constrained clustering [13,14]. The constraints ensure that the observations from the same MCMC draw are assigned to different clusters. Several genuine modes in the posterior are modeled by including more clusters than there are segments in the mixture. The new approach is compared to the previously suggested one using the galaxy dataset (which has previously been used to illustrate the problem of genuine multimodality [11]), as well as a mixture of linear regression models.

2 Genuine multimodality

In the following we consider finite mixture models of form

$$h(\boldsymbol{y}_i|\boldsymbol{x}_i, \Theta) = \sum_{s=1}^{S} \pi_s f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_s), \tag{1}$$

where h is the mixture density, $\boldsymbol{y}_i$ is the vector of responses, and $\boldsymbol{x}_i$ an optional vector of covariates for observation i, $i = 1, \dots, N$. S is the number of components, f the component density function (which is assumed to be from the same distributional family for all components), $\boldsymbol{\theta}_s$ the component specific parameters of density f and π_s the component weights. The component specific parameters are denoted by $\boldsymbol{\vartheta}_s = (\pi_s, \boldsymbol{\theta}_s)$ and $\Theta = (\boldsymbol{\vartheta}_s)_{s=1,\dots,S}$ is the vector of all parameters. It is assumed that $\Theta \in \Omega$, where Ω is the set of admissible parameter vectors with

- $0 < \pi_s \leq 1$, $\forall s = 1, \dots, S$,
- $\sum_{s=1}^{S} \pi_s = 1$ and
- $\boldsymbol{\theta}_s \neq \boldsymbol{\theta}_t$ $\forall s \neq t$ with $s, t \in \{1, \dots, S\}$.

Please note that only the conditional density of $\boldsymbol{y}_i$ given $\boldsymbol{x}_i$ is investigated. For the distribution of $\boldsymbol{x}_i$ we assume that it is component independent. In general a dependency between variables $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ is assumed because otherwise variable $\boldsymbol{x}_i$ could be omitted and $h(\boldsymbol{y}_i|\Theta)$ could be analyzed.

The posterior density is given by

$$p(\Theta|\boldsymbol{y}_i, \boldsymbol{x}_i) \propto h(\boldsymbol{y}_i|\boldsymbol{x}_i, \Theta)p(\Theta),$$

where $p(\Theta)$ denotes the prior density. In the following the prior is assumed to be symmetric with respect to the components. As improper priors can lead to improper posteriors during Gibbs sampling due to empty components only proper priors are considered.

The a-posteriori probabilities for each observation which can be used to either classify the data or examine the overlap of the components are given by

$$\tau_{is}(\Theta) = \tau_s(\Theta|\boldsymbol{y}_i, \boldsymbol{x}_i) = \frac{\pi_s f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_s)}{\sum_{t=1}^{S} \pi_t f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_t)}, \qquad s = 1, \dots, S.$$

Let $\mathscr{A}_S = \mathscr{A}_S(f, \Omega)$ be the set of all finite mixture models with S components and mixture densities given by Equation (1), i.e. the component density function is f and $\Theta \in \Omega$. Due to label switching $\mathscr{A}_S$ induces a system of equivalence classes Ξ on Ω where two elements of Ω are in the same equivalence class if there exists a permutation such that they are equal

$$\Theta_1, \Theta_2 \in \Xi \Leftrightarrow \exists \nu \in \mathrm{Perm}(S) : \Theta_1 = \nu(\Theta_2).$$

$\mathrm{Perm}(S)$ denotes the set of all possible permutations of S objects. Let $\tilde{\Omega} = \mathrm{ident}(\Omega) \subset \Omega$ be the subset of parameterizations which contain only one permutation of each possible set of component parameters (see also [12]). In the following focus is given to competing parameterizations for the same model which are not equivalent in the parameter space $\tilde{\Omega}$ of the equivalence classes induced by label permutation. The presence of these genuine competing parameterizations is referred to as *genuine multimodality*:

Definition 1. The posterior density p of the parameters $\Theta \in \Omega$ is called *genuinely multimodal* if it holds for the set of modes $\mathcal{M}$ of p that

$$\exists \Theta_1, \Theta_2 \in \mathcal{M} : \Theta_1 \neq \nu(\Theta_2) \qquad \forall \nu \in \mathrm{Perm}(S).$$

A *mode* is defined as a local maximum in the probability density function. In the case of a density function with constant values at a peak, all of the points on this peak shall be considered a single mode (cf. [15, p.1646]).

The posterior density p is called *not genuinely multimodal* if the set of modes $\mathcal{M}$ contains only parameterizations which are identical up to a suitable permutation of the components. An equivalent definition where the admissible parameter space Ω has been suitably restricted to $\tilde{\Omega}$ is given by

Definition 2. The posterior density p of the parameters $\Theta \in \tilde{\Omega}$ is called *not genuinely multimodal* if the set of modes $\mathcal{M}$ of p is a singleton.

3 Relabelling under genuine multimodality

Label switching complicates the detection of genuine multimodality of the posterior. A straight forward procedure would be to first restrict the admissible parameters space Ω to $\tilde{\Omega}$ where for each equivalence class induced by label switching a single representative parameterization is selected. Then one could analyze the resulting parameterizations in the parameter space $\tilde{\Omega}$. However, if genuine multimodality is present, relabelling algorithms may fail completely in selecting a suitable subspace $\tilde{\Omega}$, because they assume that the different modes result from permutations of the same parameterization. This implies that these stepwise procedures will then fail to find both an allocation to the different modes as well as a unique labelling of the components. It is therefore preferable to pursue an approach where the mode allocations and the relabelling of components are simultaneously determined.

An extension of the relabelling algorithm to genuine multimodality was proposed by Stephens [11] using a decision theoretic approach. Assume we are given B parameter vectors (MCMC draws, bootstrap replica, …) which we want to assign to M different genuine modes. With each mode we associate a mode-specific action $a_m, m = 1, \ldots, M$. We measure the loss (for a definition see for example [16]) for taking action a_m when the true parameter is Θ by $\mathcal{L}_0(a_m; \Theta)$, see Section 5 for possible actions and loss functions. The label invariant mode specific loss which takes all possible permutations of the true parameter into account is given by

$$\mathcal{L}^M(a_m; \Theta) = \min_{\nu} \mathcal{L}_0(a_m; \nu(\Theta)).$$

Assume we undertake each mode-specific action a_m with prior probability ξ_m, where $\xi_m > 0 \; \forall m$ and $\sum_{m=1}^{M} \xi_m = 1$. Let $\boldsymbol{a} = (a_m)_{m=1,\ldots,M}$ and $\boldsymbol{\xi} = (\xi_m)_{m=1,\ldots,M}$ be the vectors of all mode-specific actions and action probabilities, respectively. Then the pair $(\boldsymbol{\xi}, \boldsymbol{a})$ describes the overall action pattern given all modes. Using a loss-minimizing strategy, the loss for selecting action $(\boldsymbol{\xi}, \boldsymbol{a})$ given the true parameter vector Θ is given by

$$\mathcal{L}\Big((\boldsymbol{\xi}, \boldsymbol{a}); \Theta\Big) = \min_{m} \Big\{ -\log \xi_m + \mathcal{L}^M(a_m; \Theta) \Big\}.$$

The following outlines an algorithm for estimating $(\boldsymbol{\xi}, \boldsymbol{a})$ for B MCMC draws where for each draw b the parameter vector is given by Θ_b.

Algorithm 1. Starting with some initial values for the permutations $\nu_{b,m}$ of the components for MCMC draw b and mode m (setting them all to the identity permutation for example) and the mode assignments m_b, $b = 1, \ldots, B$ (using a random partition of the draws for example), iterate the following steps until a fixed point is reached holding all other parameters fixed in each step:

Step 1: Determine $\boldsymbol{\xi}$ by

$$\xi_m = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{\{m_b = m\}},$$

where $\mathbb{I}$ is the indicator function.

Step 2: For $m = 1, \ldots, M$ choose action a_m such that

$$a_m = \arg\min_{a} \sum_{b=1}^{B} \mathbb{I}_{\{m_b = m\}} \mathcal{L}_0(a; \nu_{b,m}(\Theta_b)).$$

Step 3: For $b = 1, \dots, B$ and $m = 1, \dots, M$ choose $\nu_{b,m}$ such that

$$\nu_{b,m} = \arg\min_{\nu} \mathcal{L}_0(a_m; \nu(\Theta_b)).$$

Step 4: For $b = 1, \dots, B$ choose m_b such that

$$m_b = \arg\min_{m} \left[-\log \xi_m + \mathcal{L}_0(a_m; \nu_{b,m}(\Theta_b))\right].$$

If in any step the minimum is not unique, a solution is randomly chosen, except if for unchanged parameters the minimum is also attained. In this case the unchanged parameters are retained.

The algorithm is guaranteed to converge as the objective function is decreased in each step before a fixed point is reached and the number of possible combinations of the component permutations and assignments to the different modes is finite.

Corollary 1. The objective function is decreased in each iteration until a fixed point is reached.

Proof. The comparison of the values of the objective function from iteration $(n-1)$ to iteration (n) is for all $n \geq 2$ given by

$$\begin{aligned}
&\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{I}_{\{m_b^{(n-1)}=m\}} \left[-\log(\xi_m^{(n-1)}) + \mathcal{L}_0(a_m^{(n-1)}; \nu_{b,m}^{(n-1)}(\Theta_b))\right] \overset{\text{Step 1}}{\geq} \\
&\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{I}_{\{m_b^{(n-1)}=m\}} \left[-\log(\xi_m^{(n)}) + \mathcal{L}_0(a_m^{(n-1)}; \nu_{b,m}^{(n-1)}(\Theta_b))\right] \overset{\text{Step 2}}{\geq} \\
&\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{I}_{\{m_b^{(n-1)}=m\}} \left[-\log(\xi_m^{(n)}) + \mathcal{L}_0(a_m^{(n)}; \nu_{b,m}^{(n-1)}(\Theta_b))\right] \overset{\text{Step 3}}{\geq} \\
&\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{I}_{\{m_b^{(n-1)}=m\}} \left[-\log(\xi_m^{(n)}) + \mathcal{L}_0(a_m^{(n)}; \nu_{b,m}^{(n)}(\Theta_b))\right] \overset{\text{Step 4}}{\geq} \\
&\sum_{b=1}^{B}\sum_{m=1}^{M} \mathbb{I}_{\{m_b^{(n)}=m\}} \left[-\log(\xi_m^{(n)}) + \mathcal{L}_0(a_m^{(n)}; \nu_{b,m}^{(n)}(\Theta_b))\right].
\end{aligned}$$

The superscripts $(n-1)$ and (n) denote in which iteration the parameters were determined. For Step 1 the inequality condition applies due to the Gibbs inequality and because $\sum_{b=1}^{B} \mathbb{I}_{\{m_b^{(n-1)}=m\}} = B\xi_m^{(n)}$. In the other steps the inequalities follow directly from the minimizations made. If for all four steps equality holds for the inequalities the parameters are the same for step (n) and $(n-1)$ and a fixed point is reached. Otherwise, the objective function has been decreased from step $(n-1)$ to step (n). □

However, the optimum found may only be a local minimum. To increase the chance of detecting the global optimum the algorithm is in general run repeatedly with different random initializations. Please note that a mode can become empty during the iterations of the algorithm and the solution returned will then have less modes than a priori specified.

The computational burden of Step 2 depends on the loss function. For some loss functions the solution can be determined in closed form, while for others a general purpose optimizer has to be used. The optimal ordering in Step 3 can be determined quickly by solving a linear sum assignment problem (LSAP) if the loss can be divided into a sum of component specific losses, i.e. the loss $\mathcal{L}_0$ is of form

$$\mathcal{L}_0(a; \Theta) = \sum_{s=1}^{S} \mathcal{L}_0^s(a(s); \boldsymbol{\vartheta}_s),$$

where $a(s)$ is the component specific action taken.

The LSAP aims at finding a minimum cost assignment of S objects to K persons given a cost matrix of dimension $S \times K$ under the constraint that not more than one object is assigned to each person. This problem can be solved, e.g., using a primal-dual algorithm such as the so-called Hungarian method [17] which finds the optimum in time $\mathcal{O}(K^3)$. For the relabelling algorithm a special case of the LSAP has to be solved where $S = K$. The remaining two steps (Step 1 and 4) are easily computed.

A heuristic method for determining the optimal number of modes is to investigate the improvement of the objective function for an increasing number of modes. As long as the number of modes in the data is larger than the number of modes fitted with the relabelling algorithm a large improvement of the objective function can be achieved by adding a mode. If the number of modes fitted exceeds the number of modes in the data a true mode is randomly split and only a slight improvement in the objective function can be observed. A subjective decision on the number of modes can hence be based on a barplot of the values of the objective function for different number of modes. Other strategies for determining the number of clusters for clustering approaches have been suggested and might also be useful in this context (see for example [18] and [19]).

4 Constrained clustering

By extending the relabelling approach to multiple genuine modes a separate action is determined for each mode. This has the disadvantage that components which are identical in several different modes are not identified and hence not relabelled using the combined information. An alternative approach which overcomes this drawback is to use constrained clustering [13,14], where the component specific estimates are used as input data. This first pools all information from all modes, and then derives an appropriate assignment to different genuine modes (if present).

Constrained clustering is the same as ordinary clustering except that the derived partition of the data has to fulfill certain additional restrictions or constraints. Different possible constraints are for example *must-link* or *cannot-link* constraints. Must-link constraints are imposed if it has to be ensured that certain data vectors are assigned to the same cluster. This is a valuable constraint for example if individuals are clustered and repeated observations are available for them. Cannot-link constraints ensure that certain observations are assigned to different clusters. These constraints are used in the case of determining suitable labels for the components of each MCMC draw because it has to be ensured that the component specific estimates of each draw are assigned to different clusters. By combining the cluster assignments with the information which estimates belong to the same MCMC draw, an assignment to different modes is derived.

The input data $X_{B,S}$ is given by $\{x_{b,s} : b = 1, \ldots, B; s = 1, \ldots, S\}$, where $x_{b,s}$ is either equal to $\boldsymbol{\vartheta}^b_s$ after suitable data pre-processing or the posterior probabilities $\boldsymbol{\tau}_s(\Theta_b) = (\tau_{is}(\Theta_b))_{i=1,\ldots,N}$. If the parameters are used as input data data pre-processing aims at determining a suitable weighting of the variables which is especially important in this case because the scales in general differ. Standardization as a form of weighting assigns equal weights to each of the variables. Often not all variables contribute equally to identifying the cluster structure in the data and true clusters are masked by the presence of irrelevant variables. Feature selection aims at determining the optimal subset of variables for identifying the cluster structure in the data. Different methods for this data pre-processing step have been proposed (see for example [20] and [21]).

The proposed constrained K-centroids clustering approach determines K centroids $C_K = \{c_1, \ldots, c_k\}$ by minimizing

$$\sum_{b=1}^{B}\sum_{s=1}^{S} d(c(x_{b,s}), x_{b,s})$$

under the condition

$$c(x_{b,s}) \neq c(x_{b,t}) \quad \forall s \neq t; s, t \in \{1, \ldots, S\}, \qquad \forall b = 1, \ldots, B,$$

where $c(x) \in C_K$ is the cluster centroid closest to x with respect to dissimilarity $d(\cdot, \cdot)$.

A solution to this optimization problem can be found using the following algorithm:

Algorithm 2. Start with a random set of initial centroids $C_K = \{c_1, \ldots, c_K\}$, e.g. by randomly choosing K unique vectors from the data. Then iterate the following steps until a fixed point is reached:

Step A: Assign each vector of component specific estimates $x_{b,s}$ to the cluster of the closest centroid:

$$c(x_{b,s}) := \arg\min_{c \in C_K} d(c, x_{b,s}).$$

Step B: If the constraint is violated for the estimates of one draw, i.e.

$$\exists b, s, t : \quad (s \neq t) \wedge (c(x_{b,s}) = c(x_{b,t}))$$

then find the best assignment to the clusters under the constraint. This can again be made by solving an LSAP.

Step C: Update the set of centroids by minimizing the following functions $\forall k$:

$$c_k := \arg\min_{c} \sum_{x_{b,s} \in A_k} d(c, x_{b,s}),$$

where A_k is the set of points in cluster k, i.e., $A_k := \{x_{b,s} \in X_{B,S} | c(x_{b,s}) = c_k\}$.

This algorithm has been implemented in R package **flexclust** [22], see [14] for details on the LSAP in Step B.

The optimal solution of the constrained clustering approach if the number of clusters K is equal to the number of segments S is equivalent to the solution of the relabelling algorithm proposed by Stephens [11] if the loss used is given by

$$\mathcal{L}_0(a; \Theta) = \sum_{s=1}^{S} d(a(s), \boldsymbol{\vartheta}_s). \tag{2}$$

Similar to the determination of the number of modes in the relabelling algorithm the optimal number of clusters can be determined by examining a barplot of the within cluster dissimilarities for different number of clusters. This simple heuristic suggests to choose the number of clusters where an elbow in the curve can be observed.

5 Loss functions and dissimilarity measures

For the relabelling algorithm a suitable label-invariant loss function has to be selected, while for the constrained clustering approach a dissimilarity measure has to be chosen. Given a dissimilarity measure a corresponding loss function is induced (see Equation (2)). Different loss functions for the relabelling algorithm have been proposed [10,11]. In this section the Kullback-Leibler divergence (KL) [23] which has been previously proposed for the relabelling algorithm as loss function is analyzed for suitability as dissimilarity measure in the constrained clustering approach. For the KL divergence the a-posteriori probabilities are used as input data. This is a sensible approach if cluster inference shall be made and it has the advantage that it is independent of the component specific model and can be used for arbitrary finite mixture models.

The Kullback-Leibler divergence measures the difference between a given "true" probability measure p to an arbitrary probability measure q and is given by

$$d_{\mathrm{KL}}(q,p) = \int p(x) \log \frac{p(x)}{q(x)} dx = \int p(x)(\log p(x) - \log q(x)) dx.$$

The KL divergence is a dissimilarity measure for probability measures or more general for a set of objects of equal length with nonnegative elements. It is not symmetric, but it can be interpreted as measuring the error made by replacing a given probability measure p with q.

Numerical instabilities can occur if the probability measure has very small values for certain points as the logarithm is converging to minus infinity for values converging to zero. To avoid these problems slightly modified input values are used in the following for the KL divergence. Values smaller than a threshold ϵ are replaced with ϵ. This ensures that the logarithm is bounded. In order to guarantee nonnegativity the resulting input values are rescaled to be of equal length.

For the relabelling algorithm the mode-specific loss function can be taken as the sum of the KL divergences between the a-posteriori probabilities of the observations and the action a_m. This is given by

$$\mathcal{L}_0^M(Q^m; \Theta) = \sum_{i=1}^{N} \sum_{s=1}^{S} \tau_{is}(\Theta) \log \left(\frac{\tau_{is}(\Theta)}{q_{is}^m} \right).$$

The action a^m is given by $Q^m = (q_{is}^m)_{i=1,\ldots,N;s=1,\ldots,S}$, where q_{is}^m represents the probability that observation i is assigned to group s for mode m. If it can be assumed that the empirical distribution of the observations approximates the unconditional mixture distribution induced by Θ, it holds that

$$\frac{1}{N}\mathcal{L}_0^M(Q^m;\Theta) \approx d_{\text{KL}}(\Pi(Q^m),\Pi(\Theta)) + \sum_{s=1}^{S} \pi_s(\Theta) d_{\text{KL}}(f_s(Q^m), f_s(\Theta)) \\ - d_{\text{KL}}(h(Q^m), h(\Theta)), \quad (3)$$

where f_s is the unconditional component specific density function $f(\cdot,\cdot|\theta_s)$, h is the unconditional mixture density and $\Pi = (\pi_s)_{s=1,\ldots,S}$. The densities f_s and h and the parameter vector Π are all either induced by Θ or the action Q^m. The mode-specific actions (Step 2) are given in closed form by determining the means of the correctly labelled a-posteriori probabilities over the B replica.

For the use of the KL divergence as dissimilarity measure for the constrained clustering approach a partition into a sum of component-specific losses is easily possible. Due to its asymmetry the order of the input arguments has to be decided. The centroids are intuitively inserted as the first argument in the objective function because the KL divergence then measures the loss for using the centroid instead of the observation. This order of the input values also has the advantage that the centroids can be determined in closed form by averaging over the observations assigned to the respective clusters. To ensure nonnegativity the component specific posteriors have to be rescaled to be of equal length. This implies that the relative weight or size of the components is neglected in measuring the dissimilarity, i.e. the component distributions are more or less directly compared instead of the with π_s weighted component distributions. This gives the following dissimilarity measure for the component specific posteriors for the constrained clustering algorithm:

$$d(\boldsymbol{q}_k, \boldsymbol{\tau}_s(\Theta)) = \pi_s(\Theta) \sum_{i=1}^{N} \frac{\tau_{is}(\Theta)}{\pi_s(\Theta)} \log\left(\frac{\tau_{is}(\Theta)}{q_{ik}\pi_s(\Theta)}\right),$$

where $\boldsymbol{q}_k = (q_{ik})_{i=1,\ldots,N}$ is the k^{th} centroid with $\sum_{i=1}^{N} q_{ik} = 1 \ \forall k$. Please note that the transformation exploits the equality $\pi_s(\Theta) = \sum_{j=1}^{N} \tau_{js}(\Theta)$. This dissimilarity measure is referred to as weighted KL divergence where the rescaled posteriors are used as input.

Under the assumption that the empirical distribution approximates the unconditional mixture distribution it holds for the sum over all components that

$$\frac{1}{N}\sum_{s=1}^{S} d(\boldsymbol{q}_{k(s)}, \boldsymbol{\tau}_s(\Theta)) \approx \sum_{s=1}^{S} \pi_s d_{\text{KL}}(f_{k(s)}(Q), f_s(\Theta)) \\ - d_{\text{KL}}(h(Q\{(k(s))_s\}), h(\Theta)).$$

$Q = \{\boldsymbol{q}_k\}_k$ and $h(Q\{(k(s))_s\})$ is the unconditional mixture density with S components induced by selecting the S components given by $k(s)$, $s = 1,\ldots,S$, from Q. This signifies that the sum over the component specific dissimilarities is the same as the loss using the KL divergence between the a-posteriori probabilities (see Equation (3)) except that the KL divergence between the vector of the component weights is not taken into account. This comparison indicates that the constrained clustering approach and the relabelling algorithm will in general lead to similar results especially if the components are of similar size.

6 Illustration

Two examples are given for the application of the two proposed approaches. First a mixture of three t-distributions is fitted to the galaxy data set. The need to account for genuine multimodality for this example was already previously indicated [11]. The second example uses a finite mixture of Gaussian regression models where the true underlying mixture distribution is not identifiable due to intra-component label switching. The knowledge of the true underlying data generating process allows to check if the algorithms are able to detect the two a-priori known modes.

The KL divergence with the truncated posterior probabilities is used for the relabelling algorithm and the weighted KL divergence with the truncated and rescaled posterior probabilities is used for the constrained clustering method. Truncation means that values smaller than $\epsilon = 1.5\text{e-}154$ are replaced by ϵ. The data analysis is made with the statistical computing environment R [24]. jags (Just Another Gibbs Sampler) [25] is used as sampling engine.

6.1 Mixture of t-distributions using the Galaxy data set

The data set consists of velocities (in 10^3 km/s) of 82 galaxies from six well-separated conic sections of an survey of the Corona Borealis area. The data is assumed to come from a mixture of three t-distributions with 4 degrees of freedom. Details of the priors and the corresponding Gibbs sampling steps are given in [10]. The Gibbs sampler is run for 10000 iterations where the first 5000 draws are discarded as burn-in. Traces of the remaining draws for the component specific means are given in Figure 1. It can be seen that the MCMC draws cluster around a different mode for the iterations between 1000 and 2000 with a label switching between Component 1 and 3 for the iterations before 1000 and after 2000. Imposing an ordering constraint would eliminate the label switching within a mode, but would not allow to automatically differentiate between the two modes.

For the constrained clustering method the algorithm is started with 10 differ-

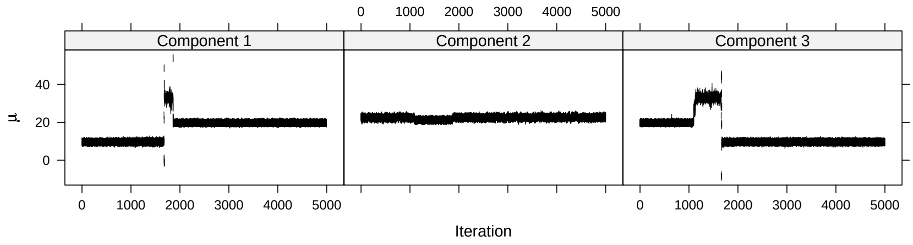

Fig. 1. Raw output of the Gibbs sampler.

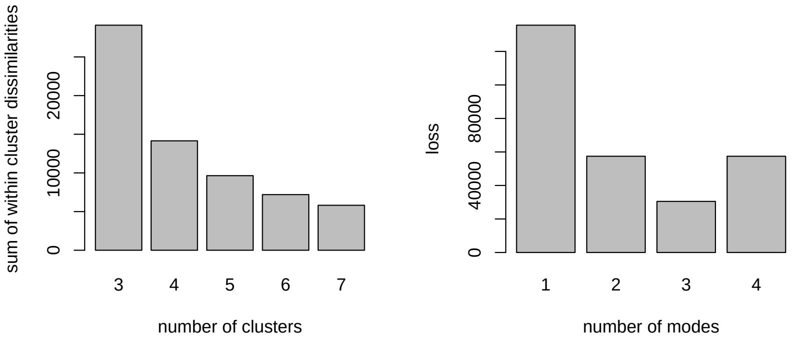

Fig. 2. Diagnostic plot for the number of clusters for the constrained clustering approach left and for the number of modes for the relabelling algorithm right.

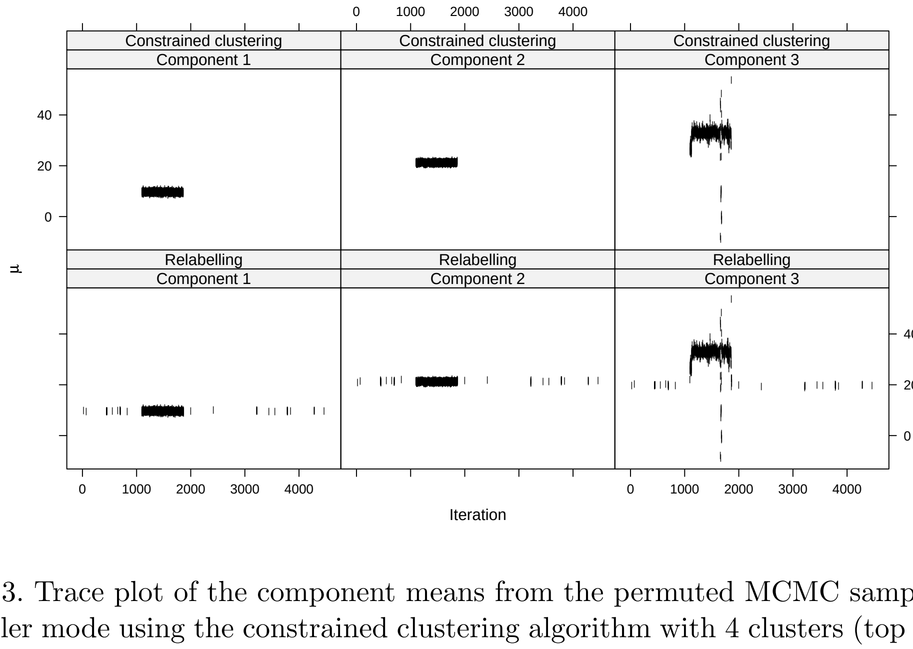

Fig. 3. Trace plot of the component means from the permuted MCMC sample and smaller mode using the constrained clustering algorithm with 4 clusters (top panel) and the relabelling algorithm with 2 modes (bottom panel).

ent random initializations (i.e. K distinct data points are randomly selected as centroids) and the number of clusters is varied from 3 to 7. To select a suitable number of clusters the sum of within cluster dissimilarities are plotted against the number of clusters (see Figure 2 on the left). An elbow can be observed where the turning point is for 4 clusters which suggests that this is the optimal number of clusters. The constrained clustering approach was hence able to detect that there are multiple genuine modes present, because for unimodality the optimal number of clusters is equal to the number of components of the mixture. In addition it can be concluded that the two modes only differ with respect to one component while two components are stable. If 4 clusters are selected the resulting cluster assignments can be combined with the information which observations are from the same MCMC draw in order to determine a mode assignment. This gives 2 modes which contain 84.8% and 15.2% of the MCMC draws respectively. The traces for each cluster and the smaller mode are given in Figure 3 in the top panel labelled "Constrained clustering".

For the relabelling approach under genuine multimodality the algorithm is again started with 10 different random initializations (i.e. the labels of the components of each draw are randomly permuted and the draws are randomly partitioned). The number of modes is varied from 1 to 4. In order to determine the suitable number of modes the total loss is plotted against the number of modes (see Figure 2 on the right). Please note that even though the algorithm was initialized with 1 to 4 different modes the best solutions detected over 10 random initializations have only 1, 2, 3 and 2 different modes as modes which become empty during the run of the algorithm are discarded. The plot indicates that the suitable number of modes is 2. If 2 modes are selected they contain 16.2% and 83.8% of the MCMC draws respectively. The traces for each component and the smaller mode are given in Figure 3 in the bottom panel labelled "Relabelling".

The congruence between the cluster and mode assignments of the constrained clustering and relabelling approaches is determined using the Rand index corrected for chance [26] as an objective measure to assess the similarity between the labellings. This gives a value of 0.95 for the mode assignments and the cluster assignments are identical where the mode assignments correspond. A further investigation of the draws which are assigned to different modes indicates that the different mode assignments only occur because they are assigned to the smaller mode for the constrained clustering approach and to the larger for the relabelling approach. An investigation of the means indicates that the solution of the constrained clustering approach is better in achieving a unique labelling. However, to evaluate the performance of the algorithms by only comparing the mean values might be inappropriate because the clustering basis were the posterior probabilities which do not only depend on the means but are also highly influenced by the variances which are allowed to vary between the components. The draws where the mode assignments differ are in fact those which are hard to classify because they have a similar dissimilarity to both modes due to the differences in variance.

6.2 Mixture of Gaussian regressions

The mixture regression example is given by

$$H(y|\mathbf{x}, \Theta) = \sum_{s=1}^{3} \frac{1}{3} \phi(y; \mu_s(\boldsymbol{x}), 0.1)$$

where $\mu_s(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_s$ and $\phi(\cdot; \mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2. The regressors are assumed to consist of an intercept, a continuous variable $x_1 \in [0, 1]$ and an interaction term between a binary variable x_2 and x_1. For simplicity of presentation no main effect of the binary variable x_2 is included, i.e., the coefficient is equal to 0 for all components. As Gaussian mixture distributions are generically identifiable the means, variances and component sizes are uniquely determined in each covariate point [27]. Due to the specific structure of the covariate matrix, only the following three covariate points are necessary to uniquely determine the marginal distributions in each possible covariate point. Let the means $\boldsymbol{\mu}_s$ for component s given the covariate matrix $\boldsymbol{X}$ of the three points be given by

$$\boldsymbol{X} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \boldsymbol{\mu}_1 = \begin{pmatrix} 4 \\ 4 \\ 2 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 2 \\ 4 \end{pmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}.$$

As the ordering of the components in each point is not unique due to the violation of the coverage condition [12], the two possible solutions for $\boldsymbol{\beta} := (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ are:

Solution 1: $\boldsymbol{\beta}_1^{(1)} = (4, 0, -2)'$, $\boldsymbol{\beta}_2^{(1)} = (4, -2, 2)'$, $\boldsymbol{\beta}_3^{(1)} = (2, 0, 0)'$,
Solution 2: $\boldsymbol{\beta}_1^{(2)} = (4, 0, \ \ 0)'$, $\boldsymbol{\beta}_2^{(2)} = (4, -2, 0)'$, $\boldsymbol{\beta}_3^{(2)} = (2, 0, 0)'$.

The omission of x_2 in the regression clearly simplifies the example, because the mixture with the same marginal distributions where the binary variable x_2 is also included in the regression and allowed to vary between the components, has 6 different parameterizations.

In the following we use a sample with 100 observations from this mixture distribution, where the x_1 values are equidistantly given in the interval $[0, 1]$

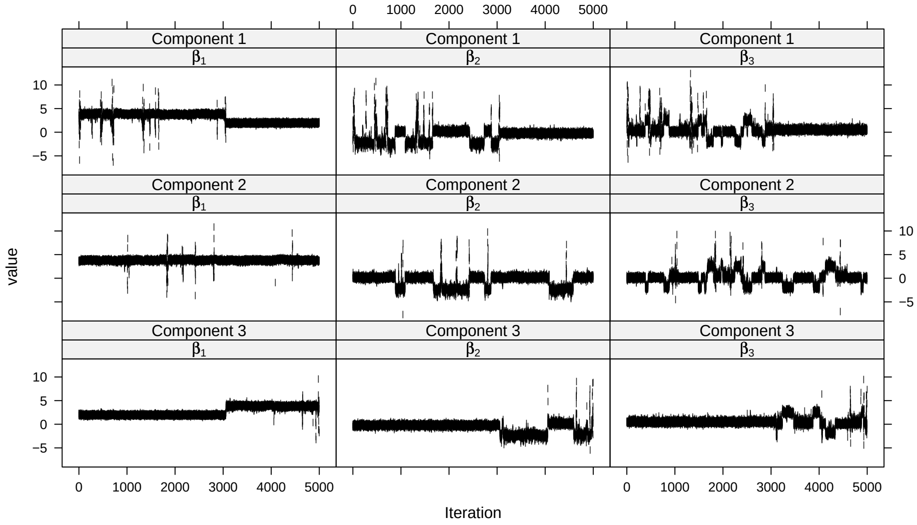

Fig. 4. Raw output of the Gibbs sampler.

and both x_2 values are observed for each x_1 value. We fit a finite mixture model with 3 components to the sample using a Gibbs sampler with similar priors and initial values as for the galaxy data. 55000 draws are simulated where the first 5000 draws are discarded as burn-in and for the remaining 50000 draws a thinning interval of 10 is used and only 5000 draws are recorded. The raw output using the recorded draws of the Gibbs sampler is given in Figure 4. The jumps in the traces clearly indicate that different modes of the posterior are visited even though it might be hard to assess at a first glance if genuinely different modes are visited.

For the constrained clustering approach 10 different random initializations are performed and the number of clusters is varied from 3 to 7. The diagnostic plot of the sum of within cluster dissimilarities against the number of clusters suggests 5 clusters (see Figure 5 on the left). The combination of the information which observations are from the same MCMC draw and the cluster labels gives 4 different modes where the two largest modes contain 62.2% and 26.6% of the MCMC draws and all other modes less than 6%. To illustrate the results the traces of the parameter β_2 are given in Figure 6 on the top panel separately for each cluster and only for the largest mode.

For the relabelling approach the algorithm is randomly initialized 10 times. The number of modes is varied from 1 to 4 and the plot of the number of modes versus the total loss suggests that the suitable number of modes is 2 (see Figure 5 on the right). If 2 modes are selected they contain 72.1% and 27.9% of the MCMC draws respectively. The traces of the parameter β_2 are given in Figure 6 separately for each component and only for the larger mode. The figure indicates that because the relabelling algorithm had to assign each draw to one of the two modes spurious draws are also included in the

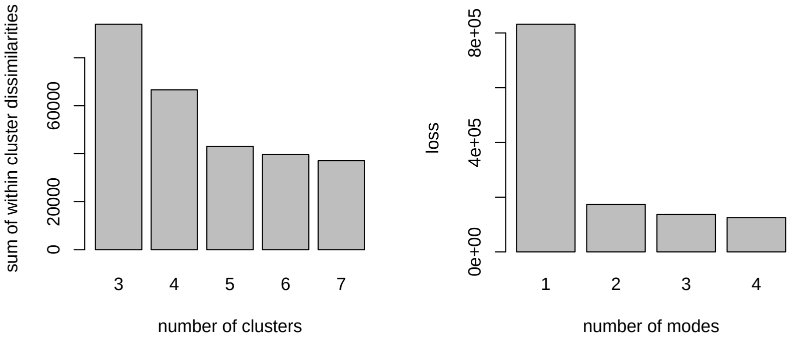

Fig. 5. Diagnostic plot for the number of clusters for the constrained clustering approach left and for the number of modes for the relabelling algorithm right.

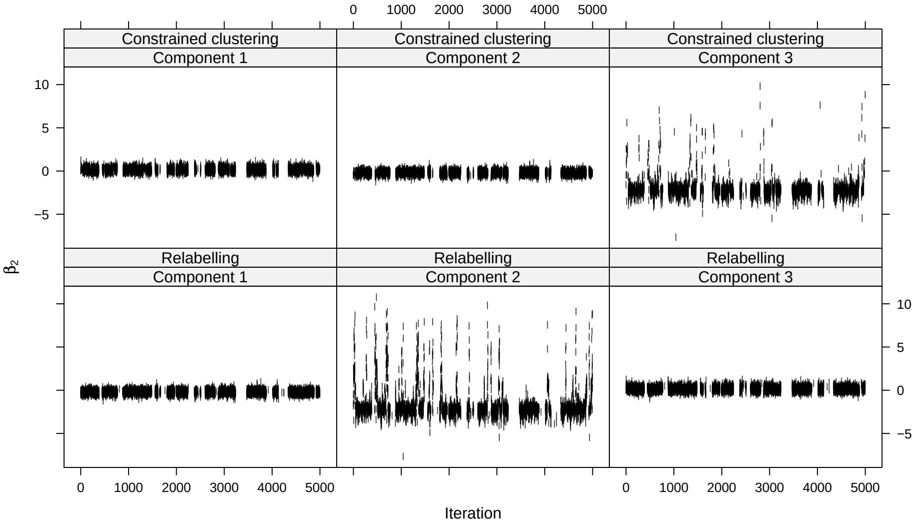

Fig. 6. Trace plot of the parameter estimates for β_2 from the permuted MCMC sample of the largest mode usint the constrained clustering algorithm with 5 clusters (top panel) and the relabelling algorithm with 2 modes (bottom panel).

larger mode. This signifies that the parameter values for Component 2 are for example not symmetrically clustering around a mean value. The constrained clustering approach is performing better in this case because it was able to eliminate spurious draws by assigning them to additional modes.

The congruence between the cluster and mode assignments of the constrained clustering and relabelling approaches can also be determined using the Rand index corrected for chance which are equal to 0.73 for the mode assignments and 1 for the cluster assignments where the mode assignments correspond. In this case the congruence between the mode assignments is relatively low because the constrained clustering approach made a classification into more different modes.

Table 1
Mean estimates (standard deviations) for each component for the overall data after relabelling with respect to an ordering constraint on β_1 ("Overall"), for the two largest modes for the constrained clustering algorithm with 5 clusters ("Clustering") and for each mode of the relabelling algorithm and 2 modes ("Relabelling").

Method	Mode	Size	Comp.	β_1	β_2	β_3
Overall		1.00	1	1.94 (1.08)	-0.06 (1.49)	0.64 (1.32)
			2	3.65 (0.46)	-0.41 (1.36)	0.00 (1.35)
			3	4.03 (0.96)	-1.28 (1.11)	0.79 (0.92)
Clustering	11010	0.62	1	3.85 (0.15)	0.19 (0.27)	0.17 (0.21)
			2	2.00 (0.14)	-0.16 (0.24)	0.58 (0.28)
			3	3.73 (0.80)	-2.05 (1.12)	0.60 (0.98)
	01101	0.27	1	1.99 (0.14)	-0.17 (0.25)	0.51 (0.28)
			2	3.78 (0.20)	0.32 (0.31)	-1.76 (0.42)
			3	4.02 (0.15)	-2.43 (0.42)	2.60 (0.38)
Relabelling	1	0.72	1	2.01 (0.14)	-0.17 (0.23)	0.57 (0.27)
			2	3.72 (0.98)	-1.59 (1.90)	0.72 (1.47)
			3	3.83 (0.16)	0.19 (0.27)	0.18 (0.22)
	2	0.28	1	1.99 (0.14)	-0.18 (0.25)	0.51 (0.28)
			2	3.77 (0.20)	0.33 (0.31)	-1.71 (0.50)
			3	4.02 (0.15)	-2.39 (0.49)	2.57 (0.45)

A comparison of the results derived using (1) an ordering constraint on β_1 for the overall dataset without accounting for the presence of different modes, (2) using the constrained clustering approach with 5 clusters and (3) the relabelling approach with 2 modes is given in Table 1. The mean values for each component are given separately for each mode together with the standard deviations in round parentheses. Please note that for the constrained clustering approach the component with the smallest intercept is estimated separately for each mode even though the same cluster is contained in both modes.

Ignoring the presence of genuine multimodality and imposing an ordering constraint is not successful in revealing any of the true underlying parameterizations. The ordering constraint approach also suffers from the fact that each parameter has theoretically the same value for at least two components. The other two approaches are able to identify the two modes in the likelihood which correspond to the two different parameterizations of the true underlying mixture distribution. The constrained clustering approach classifies only

88.8% of the draws to these two modes while the remaining draws are classified as spurious. The relabelling approach assigns each draw to one of the modes and this leads to larger estimates for the standard deviations.

7 Conclusions

The difficulty in the estimation of finite mixture models often stems from the fact that the likelihood or posterior densities are genuinely multimodal. For MCMC sampling the label switching problem also has to be addressed before component specific inference can be made for the posterior distribution. Most methods proposed to solve the label switching problem do not account for the possibility of genuine multimodality of the posterior density and are likely to fail under genuine multimodality. It is therefore necessary to have tools available which work under these conditions.

In this paper two approaches to determine simultaneously a mode assignment as well as a unique labelling of the components for each mode are presented. The two methods are equivalent under the assumption of no genuine multimodality and only differ in the way they extend the model to account for the presence of different genuine modes. In the exemplary applications both methods are shown to succeed in determining a suitable labelling. Both methods are only exploratory tools for the analysis of the MCMC draws, because they require the data analyst to determine the suitable number of clusters or modes. Diagnostic tools can assist in this decision, but the final decision can be ambiguous, especially if one mode occurs only rarely.

The advantage of the constrained clustering approach if compared to the relabelling approach under genuine multimodality is that (1) it allows to eliminate spurious draws and determine suitable labellings and mode assignments for the remaining draws and (2) enables easy identification of components which are part of several different modes. For illustrating the methods the KL divergence was selected as loss and dissimilarity measure because it can be applied for different kinds of mixture models such as model-based clustering or mixtures of generalized regression models. In the future we want to investigate the performance of other measures which are not based on the a-posteriori probabilities but directly use the parameter estimates. In addition it would be interesting to also analyze the performance of the proposed methods for applications where the genuinely different modes have different number of components.

This paper focused on Bayesian estimation problems. However, similar problems arise in a frequentist setting if bootstrap methods are applied for model diagnostics [28]. If the EM algorithm is randomly initialized for determining the maximum likelihood estimates of the bootstrap samples the solutions will correspond to different modes of the likelihood which exist either due to label switching or are due to genuine multimodality. Both proposed methods can also be used in this setting to determine a suitable labeling as well as a separation into different modes (if needed) for the parameter estimates. Additional methods for checking for the presence of genuine multimodality have been proposed in this context, which allow to determine if it is necessary to account for genuine multimodality [29].

Acknowledgements

This research was supported by the Austrian Science Foundation (FWF) under grants P17382-N12 and T351.

References

[1] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM-algorithm, Journal of the Royal Statistical Society B 39 (1977) 1–38.

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

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

[4] J.-M. Marin, K. Mengersen, C. P. Robert, Bayesian modelling and inference on mixtures of distributions, in: D. Dey, C. Rao (Eds.), Bayesian Thinking, Modeling and Computation, Vol. 25 of Handbook of Statistics, North–Holland, Amsterdam, 2005, Ch. 16, pp. 459–507.

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

[6] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, Springer Series in Statistics, Springer, New York, 2006.

[7] S. Richardson, P. J. Green, On Bayesian analysis of mixtures with an unknown number of components, Journal of the Royal Statistical Society B 59 (4) (1997) 731–92.

[8] H. Chung, E. Loken, J. L. Schafer, Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution, The American Statistician 58 (2) (2004) 152–158.

[9] G. Celeux, Bayesian inference for mixture: The label-switching problem, in: R. Payne, P. Green (Eds.), Compstat 1998 — Proceedings in Computational Statistics, Physica Verlag, Heidelberg, 1998, pp. 227–232.

[10] M. Stephens, Bayesian methods for mixtures of normal distributions, Ph.D. thesis, University of Oxford (1997).

[11] M. Stephens, Dealing with label switching in mixture models, Journal of the Royal Statistical Society B 62 (4) (2000) 795–809.

[12] C. Hennig, Identifiability of models for clusterwise linear regression, Journal of Classification 17 (2) (2000) 273–296.

[13] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl, Constrained k-means clustering with background knowledge, in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 577–584.

[14] F. Leisch, B. Grün, Extending standard cluster algorithms to allow for group constraints, in: A. Rizzi, M. Vichi (Eds.), Compstat 2006—Proceedings in Computational Statistics, Physica Verlag, Heidelberg, Germany, 2006, pp. 885–892.

[15] M. C. Minnotte, Nonparameteric testing of the existence of modes, The Annals of Statistics 25 (4) (1997) 1646–1660.

[16] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data, John Wiley & Sons, Inc., New York, USA, 1990.

[17] C. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Prentice Hall, Englewood Cliffs, USA, 1982.

[18] A. Hardy, On the number of clusters, Computational Statistics & Data Analysis 23 (1) (1996) 83–96.

[19] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society B 63 (2) (2001) 411–423.

[20] M. J. Brusco, J. D. Cradit, A variable-selection heuristic for k-means clustering, Psychometrika 66 (2) (2001) 249–270.

[21] J. H. Friedman, J. J. Meulman, Clustering objects on subsets of attributes, Journal of the Royal Statistical Society B 66 (4) (2004) 815–849.

[22] F. Leisch, A toolbox for k-centroids cluster analysis, Computational Statistics & Data Analysis 51 (2) (2006) 526–544.

[23] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86.

[24] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2008).
URL `http://www.R-project.org`

[25] M. Plummer, JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, in: K. Hornik, F. Leisch, A. Zeileis (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Technische Universität Wien, Vienna, Austria, 2003, ISSN 1609-395X. URL `http://www.ci.tuwien.ac.at/Conferences/DSC.html`

[26] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1) (1985) 193–218.

[27] H. Teicher, Identifiability of finite mixtures, The Annals of Mathematical Statistics 34 (4) (1963) 1265–1269.

[28] B. Grün, F. Leisch, Bootstrapping finite mixture models, in: J. Antoch (Ed.), Compstat 2004 — Proceedings in Computational Statistics, Physica Verlag, Heidelberg, 2004, pp. 1115–1122.

[29] B. Grün, F. Leisch, Testing for genuine multimodality in finite mixture models: Application to linear regression models, in: R. Decker, H.-J. Lenz (Eds.), Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation, Vol. 33 of Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2007, pp. 209–216.