Identifying Mixtures of Mixtures Using Bayesian Estimation

Gertraud Malsiner-Walli
Department of Applied Statistics, Johannes Kepler University Linz
and
Sylvia Frühwirth-Schnatter
Institute of Statistics and Mathematics, Wirtschaftsuniversität Wien
and
Bettina Grün*
Department of Applied Statistics, Johannes Kepler University Linz

February 24, 2015

Abstract

The use of a finite mixture of normal mixtures model in model-based clustering allows to capture non-Gaussian data clusters. However, identifying the clusters from the normal components is challenging and in general either achieved by imposing constraints on the model or by using post-processing procedures.

Within the Bayesian framework we propose a different approach to achieve identifiability. We specify hierarchical priors where the hyperparameters are carefully selected such that they are reflective of the cluster structure aimed at. Our approach allows to simultaneously (1) determine the number of clusters, (2) flexibly approximate the cluster distributions in a semi-parametric way using mixture of normals and (3) identify component-specific parameters and classify observations based on straight-forward MCMC sampling and post-processing steps to resolve the label switching issue. The proposed approach is illustrated in two simulation studies and on benchmark data sets.

Keywords: Dirichlet prior; Finite mixture model; Model-based clustering, Multivariate normal distribution; Normal gamma prior; Number of components.

*The author gratefully acknowledges support by the Austrian Science Fund (FWF): V170-N18.
1 Introduction

In many areas of applied statistics like economics, finance or public health it is often desirable to find groups of similar objects in a data set through the use of clustering techniques. Popular heuristic clustering techniques such as $k$-means (MacQueen, 1967) are based on distance measures and do not easily allow to incorporate specific knowledge about the underlying data generating distribution.

A more flexible approach to clustering data is based on finite mixture models, whereby the data in each mixture component are assumed to follow a parametric distribution with cluster-specific parameters varying over the components (see McLachlan and Peel, 2000 and Frühwirth-Schnatter, 2006 for a comprehensive survey of mixture models and their various applications). This so-called model-based clustering approach (Fraley and Raftery, 2002) is based on the notion that the component densities can be regarded as the “prototype shape of clusters to look for” (Hennig, 2010) and each mixture component may be interpreted as a distinct data cluster.

Most commonly, a finite mixture model with Gaussian component densities is fitted to the data in order to identify homogeneous data clusters within a heterogeneous population. However, assuming such a simple parametric form for the component density implies a strong assumption about the shape of the clusters and may lead to overfitting the number of clusters as well as a poor classification, if not supported by the data. Hence, a major limitation of Gaussian mixtures in the context of model-based clustering results from the presence of non-Gaussian data clusters, as typically encountered in practical applications.

Recent research demonstrates the usefulness of mixtures of parametric non-Gaussian component distributions such as the skew normal or skew-$t$ distribution to capture non-Gaussian data clusters, see, among many others, Lee and McLachlan (2014) and Vrbik and McNicholas (2014) in a frequentist framework and Frühwirth-Schnatter and Pyne (2010) in a Bayesian framework. However, as stated in Li (2005), for many applications it is difficult to decide which parametric distribution is appropriate to characterize a data cluster, especially in higher dimensions. Furthermore, the shape of the cluster densities can be of a form which is difficult to describe accurately by a basic parametric distribution. To better accommodate such data, recent advances in model-based clustering focused on designing
mixture models with more flexible, not necessarily parametric cluster densities.

A rather appealing approach, known as finite mixture of mixtures model, models the non-Gaussian clusters distributions themselves by finite Gaussian mixtures, exploiting the ability of finite normal mixtures to accurately approximate a wide class of probability distributions (Ferguson, 1983; Marron and Wand, 1992). Compared to a standard mixture with Gaussian components, mixture of mixtures models pose a two-level hierarchical structure which is particularly appealing in a clustering context. On the higher level, Gaussian components are grouped together to form non-Gaussian cluster distributions which are used for clustering the data. The individual Gaussian component densities appearing on the lower level of the model influence the clustering procedure only indirectly by accommodating possibly non-Gaussian, but otherwise homogeneous cluster distributions in a semi-parametric way.

This powerful and very flexible approach has been employed in various ways. Several researchers found it useful to work directly with the representation of the data generating distribution as a hierarchical mixture of mixtures model. Such Gaussian mixture of mixtures models have been investigated by Bartolucci (2005) for univariate data and Di Zio et al. (2007) for multivariate data using the EM algorithm. Both identify the clusters by imposing strong constraints on the locations and the covariance matrices of the Gaussian components. A different strand of literature in both the frequentist and Bayesian framework pursues the idea of creating meaningful clusters by combining components after having fitted a standard Gaussian mixture distribution to the data. The clusters are typically determined by successively merging subcomponents according to some criterion, e.g. the closeness of the means (Li, 2005), the modality of the obtained mixture density (Chan et al., 2008, 2010; Hennig, 2010; Cron et al., 2013; Lin et al., 2013) or the entropy of the resulting partition (Baudry et al., 2010).

Compared to the hierarchical mixture of mixtures approach proposed in this paper, the merging approach is prone to yield poor classifications, since the resulting clusters can only emerge as the union of clusters that have been identified under the misspecified Gaussian assumption. For illustration, the AIS data (see Appendix D) are first clustered by the function clustCombi (Baudry et al., 2010) of the R package mclust (Fraley et al., 2013).
Figure 1: AIS data set, variables “X.Bfat” and “LBM”. Scatter plots of the observations with different estimated classifications based on \texttt{Mclust} (left-hand side), \texttt{combiClust} (middle), and the sparse hierarchical mixture of mixtures approach developed in this paper ($K = 10, L = 4$) (right-hand side).

The results are shown in Figure 1. The first step identifies three clusters from fitting a standard Gaussian mixture (left-hand plot), and subsequently all data in the smallest cluster are merged with one of the bigger clusters to form two clusters (middle plot) which are not satisfactorily separated from each other due to the misspecification of the standard Gaussian mixture in the first step. In contrast, the sparse hierarchical mixture of mixtures approach we develop in the present paper identifies two well-separated clusters on the upper level of the hierarchy (right-hand plot).

Our novel sparse hierarchical mixture of mixtures model is derived within a Bayesian framework as an extension of the sparse finite mixture model introduced in Malsiner-Walli et al. (2015). Rather than using a single multivariate Gaussian distribution, we model the component densities in a sparse finite mixture model in a semi-parametric way through a Gaussian mixture distribution. Despite staying within the framework of finite mixtures on both levels of the hierarchy, the choice of a sparse prior on the cluster weights in the spirit of Rousseau and Mengersen (2011) allows for automatic selection of a suitable number of clusters from the data.

Our approach bears some relation to the emerging literature on Bayesian nonparametric (BNP) Gaussian mixtures such as Dirichlet process mixtures (DPM) (Quintana and Iglesias, 2003), truncated DPM (Suchard et al. 2010), and hierarchical DP mixtures (Cron et al. 2013). Further closely related approaches are infinite mixtures of infinite Gaussian
densities such as nested DPM \cite{Rodriguez2014} and species mixture models \cite{Argiento2014}. However, what distinguishes our approach from BNP methods, is the possibility to estimate not only the number of data clusters, but also cluster-specific functionals such as the parameters of the cluster distributions as well as cluster membership of the observations.

Statistical inference for finite mixtures is generally not easy due to problems such as label switching, spurious modes and unboundedness of the mixture likelihood \cite[see e.g.]{Fruhwirth-Schnatter2006}, but estimation of a mixture of mixtures model is particularly challenging due to additional identifiability issues. From a purely frequentist viewpoint, it cannot be inferred from the mixture likelihood of such a model which sub-components constitute which cluster. Since exchanging components between clusters on the lower level leads to different cluster distributions, while the density of the higher level mixture distribution remains the same, a mixture of mixtures model is not identifiable in the absence of additional information \cite{DiZio2007}.

As pointed out by \cite{Hennig2010}, “it rather has to be decided by the statistician under which conditions different Gaussian mixture components should be regarded as a common cluster”. In a Bayesian setting, a mixture of mixtures model becomes identifiable through the introduction of hierarchical priors on these model parameters. In the present paper, we specify a new hierarchical prior which favors highly overlapping subcomponent densities within each cluster by pulling the subcomponent means on the lower level toward the cluster center, making the cluster distributions themselves dense and connected. On the higher level, the prior is based on the notion that the cluster centers are quite distinct from each other compared to the spread of the clusters. Not surprisingly, the choice of the hyperparameters of this hierarchical prior turns out to be crucial in achieving identification and will be guided by a variance decomposition of the data.

We finally note that the implementational effort to estimate our model is moderate. Since we are staying within the framework of finite mixtures of normals and conditionally conjugate priors, standard Markov chain Monte Carlo (MCMC) methods based on data augmentation and Gibbs sampling \cite[see]{Diebolt1994,Fruhwirth-Schnatter2006} can be used. The number of data clusters is estimated directly from the MCMC out-
put avoiding additional, computationally intensive calculations of the marginal likelihood. Finally, identification of the cluster mixture distributions is obtained in a straightforward manner by clustering the draws of the cluster centers in the point process representation, as suggested by Frühwirth-Schnatter (2006, 2011).

The rest of the article is organized as follows. Section 2 describes the proposed strategy for estimating the cluster distributions and selecting the number of clusters by discussing in detail the prior specifications. Model estimation and identification strategies are discussed in Section 3. In Section 4 the performance of the proposed strategy is evaluated in two simulation studies. Application of the proposed method is illustrated on benchmark data sets in Section 5. Section 6 summarizes the results and limitations of the proposed approach.

2 Sparse hierarchical mixture of mixtures model

Following previous work on hierarchical mixtures of mixtures, we assume that $N$ observations $y_i$, $i = 1, \ldots, N$, $\dim(y_i) = r$, follow a mixture of $K$ cluster distributions,

$$p(y_i|\Theta, \eta) = \sum_{k=1}^{K} \eta_k p_k(y_i|\theta_k), \quad \Theta = (\theta_1, \ldots, \theta_K),$$

with each cluster distribution $p_k(y_i|\theta_k)$ being a mixture of $L$ normal subcomponents:

$$p_k(y_i|\theta_k) = \sum_{l=1}^{L} w_{kl} f_N(y_i|\mu_{kl}, \Sigma_{kl}).$$

The cluster weight distribution $\eta = (\eta_1, \ldots, \eta_K)$ and the cluster densities $p_k(y_i|\theta_k)$ appearing in the upper level (1) are relevant for clustering the observations based on Bayes’ rule. On the lower level (2), $L$ is chosen sufficiently large in order to allow in all clusters for an accurate semi-parametric approximation of the cluster distributions, by mixing in each cluster $k$ the multivariate Gaussian subcomponent densities $f_N(y_i|\mu_{kl}, \Sigma_{kl})$, $l = 1, \ldots, L$, according to the subcomponent weight vector $w_k = (w_{k1}, \ldots, w_{kL})$, where $w_{kl} \geq 0$ and $\sum_{l=1}^{L} w_{kl} = 1$. The cluster-specific parameter vector $\theta_k = (w_k, \mu_{k1}, \ldots, \mu_{kL}, \Sigma_{k1}, \ldots, \Sigma_{kL})$ consists of the weight vector $w_k$ as well as the means $\mu_{kl}$ and covariance matrices $\Sigma_{kl}$ of the Gaussian subcomponent densities.

In order to be able to identify the mixture distribution given in Equation (1) it is necessary to define what kind of shapes and forms are aimed at for a cluster distribution.
We give the following (vague) characterization of a data cluster: a data cluster is a very “dense” region of data points, with possibly no “gaps” within the cluster, whereas different clusters should be located well-separated from each other, i.e., here large “gaps” between the clusters are desired. We confine ourselves to the investigation of clusters with approximately convex cluster shapes, where the cluster center can be seen as a suitable representative for the whole cluster. This fact will also be used in Section 3 when the cluster distributions are identified based on clustering the draws of the cluster centers in the point process representation. Regarding volume, orientation or asymmetry of the data clusters we are looking for, no constraints on the cluster shapes and forms are imposed.

Based on this cluster concept, our aim is to model a dense and connected cluster distribution by a mixture with normal subcomponents. Different strategies regarding the modeling of the subcomponent means and covariance matrices could be employed. We decided to allow for flexible shapes for the single subcomponents but to ensure that they are overlapping. An alternative approach would be to use constrained simple shaped subcomponents, e.g., subcomponents with isotropic covariance matrices. However, in this case a large number of subcomponents might be needed to cover the whole cluster region and shrinkage of the subcomponent means toward the common cluster center may not be possible. Since some of the subcomponents have to be located far away from the cluster center in order to fit also boundary points, considerable distances have to be allowed between subcomponent means. This induces the risk of gaps within the cluster distribution and a connected cluster distribution may not result. Therefore, in our approach the cluster distributions are estimated as mixtures of only a few but unconstrained, highly dispersed and heavily overlapping subcomponents where the means are strongly pulled toward the cluster center. In this way, a connected cluster distribution is ensured.

In the Bayesian framework, our aim is to translate these modeling purposes into appropriate prior specifications for the mixture of mixtures model parameters. We basically employ standard conditionally conjugate priors for the mixture parameters. Thus, we assume that the cluster and the subcomponent weights are following symmetric Dirichlet distributions with hyperparameters $e_0$ and $d_0$ respectively, i.e. $\eta \sim Dir_K(e_0)$ and $w_k \sim Dir_L(d_0), \ k = 1,\ldots,K$. For each cluster $k, \ k = 1,\ldots,K$, a hierarchical prior is used
on the subcomponent covariance matrices $\Sigma_{kl}$ and means $\mu_{kl}$ for all $l = 1, \ldots, L$: $\Sigma_{kl}^{-1} \sim \mathcal{W}_r(c_0, C_{0k})$ with $C_{0k} \sim \mathcal{W}_r(g_0, G_0)$, whereas we assume that $\mu_{kl} \sim \mathcal{N}_r(b_{0k}, \sqrt{\Lambda_k}B_0\sqrt{\Lambda_k})$ with $b_{0k} \sim \mathcal{N}_r(m_0, M_0)$ and $\Lambda_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kr})$ with $\lambda_{kj} \sim \mathcal{G}(\nu, \nu)$, $j = 1, \ldots, r$. $\mathcal{N}_r(\cdot)$ and $\mathcal{W}_r(\cdot)$ denote the $r$-multivariate normal and Wishart distribution respectively and $\mathcal{G}(\cdot)$ the Gamma distribution, with parametrizations as used in Frühwirth-Schnatter (2006).

2.1 Priors on the subcomponent means and covariances

To model connected, dense, and well-separated cluster distributions, the prior subcomponent hyperparameters $m_0, M_0, G_0, c_0, g_0, \nu$ and $B_0$ have to be chosen carefully.

On the upper level, the covariance matrix $M_0$ controls the amount of prior shrinkage of the cluster centers $b_{0k}$ toward the overall data center $m_0 = \text{midpoint}(y)$, which we specify as the midpoint of the data. The matrix $M_0$ is defined to obtain an uninformative prior for $b_{0k}$, where the cluster locations are allowed to be widely spread apart and almost no shrinkage towards the data center takes place. Thus we choose $M_0 \gg S_y$, where $S_y$ is the empirical covariance matrix of the data, e.g. $M_0 = 10S_y$.

The hyperparameters $c_0, g_0$, and $G_0$ of the prior on the subcomponent covariance matrix $\Sigma_{kl}$ influence the volume and the flexibility of the subcomponent densities. The scale matrix $B_0$ of the prior on the subcomponent means $\mu_{kl}$ controls how dispersed from the cluster center $b_{0k}$ the single subcomponent means are allowed to be. Pulling the subcomponent means to the cluster center is important because the more the subcomponent means are allowed to differ from $b_{0k}$, the higher is the risk of gaps within a cluster. On the other hand, the diagonal matrix $\Lambda_k$ should act as a local adjustment factor which allows to correct the covariance matrix $B_0$ for each cluster and each dimension at a small scale.

Our strategy for appropriately specifying these hyperparameters is based on the variance decomposition of the mixture of mixtures model. The variance decomposition of a mixture model splits the total heterogeneity $\text{Cov}(Y)$ into the different sources contributing to the total heterogeneity.

For a standard Gaussian mixture model with $K$ clusters, the total heterogeneity $\text{Cov}(Y)$ consists of the variability of the cluster means $\mu_k$ around the mixture mean $\mu = \sum_k \eta_k \mu_k$ and the average variability within a cluster, defined by covariance matrices $\Sigma_k$ (Frühwirth-
Cov(Y) = \sum_{k=1}^{K} \eta_k \Sigma_k + \sum_{k=1}^{K} \eta_k \mu_k \mu_k' - \mu \mu' = (1 - \phi_B) Cov(Y) + \phi_B Cov(Y), \quad (3)

where \(\phi_B\) is the proportion of the total heterogeneity explained by the variability of the cluster centers \(\mu_k\) and \((1 - \phi_B)\) is the proportion of the total heterogeneity explained by the average variability within the clusters. The larger \(\phi_B\), the more the clusters are separated.

To illustrate the cluster separation induced by a given amount of explained heterogeneity \(\phi_B\), samples from a three-component standard Gaussian mixture with varying values of \(\phi_B\) are shown in Figure 2.

For a mixture of mixtures model, where the cluster distributions are mixture distributions themselves, a similar variance decomposition as in (3) holds, however, the heterogeneity \((1 - \phi_B) Cov(Y)\) explained within a cluster can be split further into two sources of variability. A proportion \(\phi_W\) of the cluster heterogeneity is explained by the variability of the subcomponent means \(\mu_{kl}\) around the cluster center \(\mu_k\), and the remaining proportion \((1 - \phi_W)\) is explained by the average variability within the subcomponents. This yields the following variance decomposition:

\[
\text{Cov}(Y) = \sum_{k=1}^{K} \eta_k \Sigma_k + \sum_{k=1}^{K} \eta_k \mu_k \mu_k' - \mu \mu' \\
= \sum_{k=1}^{K} \eta_k \sum_{l=1}^{L} w_{kl} \Sigma_{kl} + \sum_{k=1}^{K} \eta_k \left( \sum_{l=1}^{L} w_{kl} \mu_{kl} \mu_{kl}' - \mu_k \mu'_k \right) + \sum_{k=1}^{K} \eta_k \mu_k \mu_k' - \mu \mu' \quad (4)
\]

= \(1 - \phi_W\)(1 - \(\phi_B\)) Cov(Y) + \phi_W(1 - \(\phi_B\)) Cov(Y) + \phi_B Cov(Y).

Based on this variance decomposition we select the proportions \(\phi_B\) and \(\phi_W\) and incorporate them into the specification of the hyperparameters of our hierarchical prior.

\(\phi_B\) defines the proportion of variability explained by the different cluster centers. We suggest to specify \(\phi_B\) not too large, e.g., to use \(\phi_B = 0.5\). This specification may seem to be counterintuitive as in order to model well-separated clusters it would seem appropriate to select \(\phi_B\) large. However, if \(\phi_B\) is large, the major part of the total heterogeneity of the data is already explained by the variation (and separation) of the cluster means, and, as a consequence, only a small amount of heterogeneity is left for the within-cluster variability. This within-cluster variability in turn will get even more diminished by the variability
explained by the subcomponent means leading to a small amount of variability left for the subcomponents. Thus for large values of $\phi_B$, estimation of tight subcomponent densities would result, undermining our modeling aims.

$\phi_W$ defines the proportion of within-cluster variability explained by the subcomponent means and controls how strongly the subcomponent means are pulled together. Therefore, $\phi_W$ indirectly determines the overlap of the subcomponent densities. In order to achieve strong shrinkage of the subcomponent means toward the cluster center we suggest to select $\phi_W$ small, e.g. $\phi_W = 0.1$. Larger values of $\phi_W$ may allow for gaps within a cluster, which we want to avoid.

Given $\phi_B$ and $\phi_W$, we specify the scale matrix $G_0$ of the prior on $C_{0k}$ such that the a-priori expectation of the first term in the variance decomposition (4), given by

$$E\left(\sum_{k=1}^K \eta_k \sum_{l=1}^L w_{kl} \Sigma_{kl}\right) = \sum_{k=1}^K E(\eta_k) \sum_{l=1}^L E(w_{kl}) E(E(\Sigma_{kl}|C_{0k})) = g_0/(c_0 - (r + 1)/2)G_0^{-1},$$

matches the desired amount of heterogeneity explained by a subcomponent:

$$g_0/(c_0 - (r + 1)/2)G_0^{-1} = (1 - \phi_W)(1 - \phi_B)Cov(Y). \quad (5)$$

Replacing $Cov(Y)$ in (5) by the sample covariance $S_y$, and using only the main diagonal of the resulting matrix in order to take the scaling of the data into account (Stephens, 1997; Frühwirth-Schnatter, 2006), we get the following specification for $G_0$:

$$G_0^{-1} = (1 - \phi_W)(1 - \phi_B)(c_0 - (r + 1)/2)/g_0 \cdot \text{diag}(S_y). \quad (6)$$

Modeling the prior of the subcomponent covariance matrices is completed by defining the scalar prior hyperparameters $c_0$ and $g_0$. For specifying $c_0$, Frühwirth-Schnatter (2006, Section 6.3.2, p. 192) suggests to set $c_0 > 2 + (r - 1)/2$. In this way the eigenvalues of $\Sigma_{kl}^{-1} \Sigma_{km}^{-1}$ are bounded away from 0 avoiding singular matrices. We set $c_0 = 2.5 + (r - 1)/2$ to allow for a large variability of $\Sigma_{kl}$. For defining $g_0$, the Wishart density is regular if $g_0 > (r - 1)/2$. In the following we set $g_0 = 0.5 + (r - 1)/2$.

Regarding the prior specification of the subcomponent means $\mu_{kl}$, we select the scale matrix $B_0$ in order to concentrate a lot of mass near the cluster center $b_{0k}$, pulling subcomponent means towards the cluster center. Matching the a-priori expectation of the second
Figure 2: Variance decomposition of a mixture distribution. Scatter plots of samples from a standard normal mixture distribution with three components and equal weights, with a varying amount of heterogeneity explained by the variation of the component means, $\phi_B = 0.1$, $\phi_B = 0.5$ and $\phi_B = 0.9$ (from left to right).

term in the variance decomposition (4), given by

$$E \left( \sum_{k=1}^{K} \eta_k \left( \sum_{l=1}^{L} w_{kl} \mu_{kl} \mu'_{kl} - \mu_k \mu'_k \right) \right) = \sum_{k=1}^{K} E(\eta_k) \sum_{l=1}^{L} E(w_{kl}) \left( \mu_{kl} \mu'_{kl} - \mu_k \mu'_k \right) = B_0$$

where $B_0$ is the desired amount of heterogeneity explained by a subcomponent and, using again, only the main diagonal of $S_y$ we obtain $B_0 = \phi_W (1 - \phi_B) \text{diag}(S_y)$, which incorporates our idea that only a small proportion $\phi_W$ of the within-cluster variability should be explained by the variability of the subcomponent means.

After having chosen $\phi_B$ and $\phi_W$, basically the cluster structure and shape is a priori determined. However, in order to allow for a little bit more flexibility in capturing the unknown cluster shapes in the sense that within a cluster the amount of shrinkage of the subcomponent means toward the cluster center $b_{0k}$ does not necessarily need to be the same for all dimensions, for each cluster $k$ and each dimension $j$ additionally a random adaptation factor $\lambda_{kj}$ is introduced which adjusts the variance defined in $B_0$. $\lambda_{kj}$ follows a-priori a Gamma distribution with both hyperparameters equal to $\nu$:

$$\mu_{kl} \sim N_r(b_{0k}, \sqrt{\Lambda_k} B_0 \sqrt{\Lambda_k}), \quad \Lambda_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kr}), \quad k = 1, \ldots, K; \ l = 1, \ldots, L,$$

$$\lambda_{kj} \sim \mathcal{G}(\nu, \nu), \quad k = 1, \ldots, K; \ j = 1, \ldots, r.$$

This hierarchical prior specification for $\mu_{kl}$ corresponds to the normal gamma prior, introduced by Griffin and Brown (2010) in the context of variable selection in linear regression.
and applied by Frühwirth-Schnatter (2011) and Malsiner-Walli et al. (2015) in the context of finite mixture models. Selecting the hyperparameters of the Gamma distribution to be the same implies that the a-priori expectation of the covariance matrix of $\mu_{kl}$ equals the specified value of $B_0$. However, $\lambda_{kj}$ acts as a local adjustment factor which allows to shrink (or inflate) the variance of subcomponent means $\mu_{klj}$ in dimension $j$ in order to fit a more (or less) dense cluster distribution as specified by $B_0$. In order to allow only for small adjustments of the specified $B_0$, we choose $\nu = 10$, in this way almost 95% of the a-priori values of $\lambda_{kj}$ are between 0.5 and 1.5.

### 2.2 Priors on the cluster and subcomponent weights

We assume that the number of clusters $K$ is unknown and has to be estimated from the data. For estimating the number of clusters we use the concept of “sparse finite mixture models”, see Malsiner-Walli et al. (2015), and generalize it to the “sparse hierarchical mixture of mixtures model”.

Roughly speaking, the concept of “sparse finite mixture models” consists in defining an overfitting mixture model where the number of specified components $K$ clearly overfits the true number of components $K_{\text{true}}$. However, the simultaneous specification of a sparse prior on the mixture weights has the effect that a sparse estimation of the number of components results. This approach is inspired by results of Rousseau and Mengersen (2011) who investigate the asymptotic behavior of the posterior distribution of an overfitting mixture model.

Rousseau and Mengersen (2011) show that the way how the posterior distribution handles superfluous components in an overfitting mixture model, i.e. by either leaving superfluous components empty or splitting “true” components into two or more identical components, is asymptotically determined by the size of the hyperparameter $e_0$ of the Dirichlet prior on the weights. If $e_0 < d/2$, where $d$ is the dimension of the component-specific parameter $\theta_k$, then the posterior expectation of the weights asymptotically converges to zero for superfluous components. On the other hand, if $e_0 > d/2$, then the posterior density handles overfitting by defining at least two identical components, with non-negligible weight each.
We use these results regarding the influence of the weight prior on the asymptotic behavior of the posterior of an overfitting mixture distribution for specifying both the cluster and subcomponent weights priors $\text{Dir}_K(e_0)$ and $\text{Dir}_L(d_0)$. Our strategy consists in specifying overfitting mixture models on both levels of the hierarchy, i.e. both the specified number of clusters $K$ exceeds the true number of data clusters and the number of subcomponents $L$ forming each cluster is redundant for approximating the cluster distributions. However, the overfitting should be handled in a different way on the two levels.

On the cluster level, we want to estimate the number of clusters by leaving all superfluous clusters empty. Hence, we specify a sparse prior on the cluster weights $\eta = (\eta_1, \ldots, \eta_K)$ by choosing $e_0 \ll d/2$ so that superfluous clusters are emptied during MCMC sampling and the number of non-empty clusters is an estimator for the true number of clusters. In this way, specification of a sparse cluster weight prior in an overfitting mixture of mixtures model provides an “automatic tool” to select the number of clusters $K^{\text{true}}$, avoiding the expensive computation of model choice criteria or marginal likelihoods as, e.g., in Frühwirth-Schnatter (2004). Following Malsiner-Walli et al. (2015) we choose $e_0$ very small, e.g. $e_0 = 0.001$, to actually empty all superfluous clusters.

For specifying the prior on the subcomponent weights $\mathbf{w}_k = (w_{k1}, \ldots, w_{kL}) \sim \text{Dir}_L(d_0)$ within a cluster $k$, the modeling aim is different. Within a cluster, we are not interested in estimating the “true” number of subcomponents $L$ forming the cluster. We rather use the normal mixture distribution with $L$ components to fit a (possibly) non-Gaussian cluster distribution in a semi-parametric way. We have no clustering task at this level, thus rendering identification of the single subcomponents unnecessary. Therefore, we specify the same redundant number of normal components $L$ for each cluster. In the following we will use $L = 4$ as a default assuming that four subcomponents will be sufficient to capture the non-Gaussian shapes on the cluster level sufficiently well. Furthermore, we specify a non-sparse prior for the subcomponent weights by choosing $d_0$ large, $d_0 > d/2$, so that during MCMC sampling observations are assigned to all subcomponents within a cluster and empty subcomponents are avoided. In this way, a good density approximation of the cluster distribution is achieved. For our simulation studies and applications we use the specification $d_0 = d/2 + 2$. 

13
3 Model estimation and identification

Bayesian estimation of the sparse hierarchical mixture of mixtures model is performed using MCMC methods based on data augmentation and Gibbs sampling (Diebolt and Robert, 1994; Frühwirth-Schnatter, 2006, p. 75). We only need standard Gibbs sampling steps. The detailed MCMC sampling scheme is given in Appendix A.

When estimating a mixture model in the Bayesian framework, the label switching problem has to be addressed. In the mixture of mixtures model label switching occurs on two levels: on the cluster level, if the labeling of clusters switches between clusters, and on the subcomponent level, if within a cluster the labeling of the subcomponents changes.

On the cluster level, the label switching problem is caused by invariance of the mixture likelihood given in Equation (1) with respect to reordering of the clusters. On this level label switching has to be resolved, since the single cluster distributions need to be identified. To obtain a unique labeling of the clusters it is necessary to post-process the MCMC output. We follow Frühwirth-Schnatter (2006) who suggests to cluster the cluster-specific draws \{θ₁, ..., θₖ\} in the point process representation. The point process representation of the MCMC draws allows to study the posterior distribution of the cluster-specific parameters regardless of potential label switching, which makes it very useful for cluster identification. If the number \( K \) of fitted clusters matches the true number of clusters \( K_{\text{true}} \), it can be expected that the posterior draws cluster around the “true” points \{θ₁, ..., θₖ\} (Frühwirth-Schnatter, 2006, p. 96). However, in our case of an overfitting mixture where draws are sampled from empty components, the clustering procedure has to be adapted as suggested in Frühwirth-Schnatter (2011) and described in Malsiner-Walli et al. (2015).

First, we estimate the number of non-empty clusters \( \hat{K}_0 \). For this purpose, during MCMC sampling for each iteration \( m \) the number of non-empty clusters \( K_0^{(m)} \) is determined, i.e. the number of clusters to which observations have been assigned for this particular sweep of the sampler. Then, following Nobile (2004) we obtain the posterior distribution of the number \( K_0 \) of non-empty clusters \( P(K_0 = h|y_1, \ldots, y_N), h = 1, \ldots, K \), from the MCMC output. An estimator of the true number of clusters \( \hat{K}_0 \) is then given by the value visited most often by the MCMC procedure, i.e. the mode of the (estimated) posterior distribution \( P(K_0 = h|y_1, \ldots, y_N) \), see Frühwirth-Schnatter (2011).
After having estimated the number of non-empty clusters $\hat{K}_0$, we remove all draws generated in iterations where the number of non-empty clusters does not correspond to $\hat{K}_0$. Furthermore, we remove all cluster-specific draws sampled from empty components. The remaining cluster-specific draws are clustered into $\hat{K}_0$ groups using the $K$-means cluster algorithm (Hartigan and Wong, 1979). The clustering algorithm results in a classification sequence for each iteration indicating to which cluster a single cluster-specific draw belongs. These classification sequences are used to reorder the draws, and a unique labeling of the cluster-specific draws is achieved. Details on the clustering of the MCMC draws in the point process representation can be found in the Appendix of Malsiner-Walli et al. (2015).

To reduce the dimensionality of the clustering, Frühwirth-Schnatter (2006, 2011) suggests to cluster only a functional of the cluster-specific parameters $\theta_k$. Following this suggestion, we cluster the cluster centers determined by $\mu_k^{(m)} = \sum_{l=1}^L w_{kl}^{(m)} \mu_{kl}^{(m)}$ in the point process representation. In Appendix B, an example of the clustering in the point representation is given.

On the subcomponent level, label switching happens due to the invariance of Equation (2) with respect to reordering of the subcomponent. As we are only interested in estimating the whole cluster distribution, it is not necessary to identify the single subcomponents. Therefore, we can ignore the label switching problem on this level.

4 Simulation studies

In the following two simulation studies, the performance of the proposed strategy for selecting the unknown number of clusters and identifying the cluster distributions is illustrated. In the first simulation study we investigate whether we are able to capture dense non-Gaussian data clusters and estimate the true number of data clusters. Furthermore, the influence of the specified number of clusters $K$ and subcomponents $L$ on the clustering results is studied. In the second simulation study the sensitivity of the a-priori defined proportions $\phi_B$ and $\phi_W$ on the clustering result is investigated.

For each simulation setting, 10 data sets are generated and a sparse hierarchical mixture of mixtures model is estimated. Prior distributions and hyperparameters are specified as described in Sections 2.1 and 2.2. MCMC sampling is run for $M = 4000$ iterations after
a burn-in of 4000 draws. For the sampling, the starting classification of the observations is obtained by first clustering the observations into \( K \) groups using \( K \)-means clustering and by then allocating the observations within each group to the \( L \) subcomponents by using \( K \)-means clustering again. The estimated number of clusters is reported in Tables 1 and 2 where in parentheses the number of the data sets for which this number is estimated is given.

4.1 Simulation setup I

The simulation setup I consists of drawing samples with 800 observations grouped in four clusters. Each cluster is generated by a normal mixture with a different number of subcomponents. In Figure 3 the scatter plot of a data set and the 90% probability contour lines of the generating subcomponent distributions are shown. The number of generating distributions for each cluster (clockwise from top left) is 1, 2, 2, and 3. This simulation setup is inspired by Baudry et al. (2010) who use clusters similar to the elliptical and cross-shaped clusters on the top of the scatter plot in Figure 3. However, our simulation setup is expanded by the two clusters at the bottom which have a triangle and an \( L \) shape. Our aim is to recover the four clusters. The detailed model specification is given in Appendix C.

If we estimate a sparse finite mixture model (see Malsiner-Walli et al., 2015), which can be seen as a special case of the sparse hierarchical mixture of mixtures model with number of subcomponents \( L = 1 \), the estimated number of components is seven, as can be seen in the classification results shown in Figure 3 in the middle plot. This is expected, as by specifying a standard normal mixture the number of generating normal distributions is estimated rather than the number of data clusters. In contrast, if a sparse hierarchical mixture of mixtures model with \( K = 10 \) clusters and \( L = 4 \) subcomponents is fitted to the data, all but four clusters become empty during MCMC sampling and the four data clusters are captured rather well, as can be seen in the classification plot in Figure 3 on the right-hand side.

In order to study the effect of changing the specified number of clusters \( K \) and subcomponents \( L \) on the estimation result a simulation study consisting of 10 data sets with the simulation setup as explained above and varying numbers of clusters \( K = 4, 10, 15 \) and
Table 1: Simulation setup I: Results for the estimated number of non-empty clusters $\hat{K}_0$. The number of data sets estimating the reported number of non-empty clusters is given in parentheses.

| $K$ | 1   | 3   | 4   | 5   |
|-----|-----|-----|-----|-----|
| 4   | 4(10)| 4(10)| 4(10)| 4(10)|
| 10  | 7(9)| 4(10)| 4(10)| 4(10)|
| 15  | 7(9)| 4(10)| 4(9)| 4(10)|

First we study the effect of the number of specified subcomponents $L = 1, 3, 4, 5$ is performed. For each combination of $K$ and $L$ the estimated number of clusters is reported in Table 1.

For the data set shown in Figure 3, for an increasing number of specified subcomponents the estimated cluster distributions are plotted using the maximum a posteriori (MAP) weights, means and covariance matrices estimates of the subcomponents. The estimated cluster distributions look quite similar, regardless of the size of $L$. This may be due to the smoothing effect of the specified hyperpriors.
Figure 3: Simulation setup I. Scatter plot of a data set with the generating component densities shown with 90%-probability contour lines (left-hand side), and clustering results by estimating a sparse hierarchical mixture of mixtures model with $K = 10$, $L = 1$ (middle) and $K = 10$, $L = 4$ (right-hand side).

Figure 4: Simulation setup I, $K = 10$, different values of $L$. For the data set in Figure 3, the estimated cluster distributions (MAP posterior means, weights, and covariance matrices of the subcomponents) are plotted for an increasing number of specified subcomponents $L = 3, 4, 5$ (from left to right).

4.2 Simulation setup II

In Section 2.1 it is suggested to specify the between-cluster variability by $\phi_B = 0.5$ and the between-subcomponent variability by $\phi_W = 0.1$. As can be seen in the previous simulation study in Section 4.1 this a-priori specification gives promising results if the data clusters are well-separated. However, in contrast to the simulation setup I, in certain applications data clusters might be close or even overlapping. In this case, the clustering result might be sensitive in regard to the specification of $\phi_B$ and $\phi_W$. Therefore, in the following simulation setup it is investigated how the specification of $\phi_B$ and $\phi_W$ affects the identification of data clusters if data clusters are not well-separated. We want to study how robust the clustering
Figure 5: Simulation setup II: Scatter plot of one data set (left-hand side), classification according to the generating distributions (middle) and to the clusters (right-hand side).

results are against misspecification of the two proportions.

In order to mimic close data clusters, samples with 300 observations are generated from a three-component normal mixture, where, however, only two data clusters can be clearly distinguished. In Figure 5 the scatter plot of one data set is displayed, the generating distributions are described in Appendix C. For different values of $\phi_B$ (between 0.1 and 0.9) and $\phi_W$ (between 0.01 and 0.4) a sparse mixture of mixtures model with $K = 10$ clusters and $L = 4$ subcomponents is fitted and the number of clusters is estimated. For each combination of $\phi_B$ and $\phi_W$ the results are reported in Table 2.

Table 2 indicates that if $\phi_B$ increases, also $\phi_W$ has to increase in order to identify exactly two clusters. This makes sense since by increasing $\phi_B$ the a-priori within-cluster variability becomes smaller yielding tight subcomponent densities. Tight subcomponents in turn require a large proportion $\phi_W$ of variability explained by the subcomponent means to capture the whole cluster. Thus $\phi_W$ has to be increased too. However, $\phi_W$ has to be selected carefully. If $\phi_W$ is larger than actually needed, some subcomponents are likely to “emigrate” to other neighboring clusters. This leads finally to only one cluster being estimated for some data sets. This is basically the case for some of the combinations of $\phi_B$ and $\phi_W$ displayed in the upper triangle of the table. In contrast, if $\phi_W$ is smaller than needed, due to the induced shrinkage of the subcomponent means toward the cluster center, the specified cluster mixture distribution is not able to fit the whole data cluster and two cluster distributions are needed to fit a single data cluster. This can be seen for some of the combinations of $\phi_B$ and $\phi_W$ displayed in the lower triangle of the table.
Table 2: Simulation setup II, number of data sets = 10, $K = 10$, $L = 4$. Simulation results for estimating the number of clusters ($\hat{K}_0$) for different amounts of $\phi_B$ and $\phi_W$. The number of data sets estimating the reported $\hat{K}_0$ is given in parentheses.

5 Applications

To evaluate our approach on real data, we fit the sparse hierarchical mixture of mixtures models to various multivariate data sets which were previously used to benchmark algorithms in cluster analysis. Additionally, we apply the “merging strategy” proposed by Baudry et al. (2010) in order to compare the results to those of our approach.

The detailed description of the six investigated data sets can be found in Appendix D. For each investigated data set, our aim is to recover the known class labels of the observations. Thus the performance of the fitted model is measured by the misclassification rate and the adjusted Rand index of the resulting clustering and the class memberships. To calculate the misclassification rate, the “optimal” matching between the estimated cluster labels and the true known class labels is determined as the one minimizing the misclassification rate over all possible matches for each of the scenarios. The adjusted Rand index (Hubert and Arabie, 1985) is used to assess the similarity between the true and the estimated partition of the data. It is a corrected form of the Rand index (Rand, 1971) which
is adjusted for chance agreement. An adjusted Rand index of 1 corresponds to perfect agreement of two partitions whereas an adjusted Rand index of 0 corresponds to results no better than would be expected by randomly drawing two partitions, each with a fixed number of clusters and a fixed number of elements in each cluster. For all estimated models the number of estimated clusters $\hat{K}_0$, the adjusted Rand index $(adj)$, and misclassification rate $(er)$ are reported in Table 3.

In the first columns of Table 3 the name of the data set, the number of observations $N$, the number of variables $r$ and the number of classes $K^{true}$ (if known) are reported. To compare our approach to the merging approach proposed by Baudry et al. (2010), we use the function Mclust of the R package mclust (Fraley and Raftery, 2002; Fraley et al., 2012) to first fit a standard normal mixture distribution with the maximum number of components $K = 10$. The number of estimated normal components is reported in the column Mclust. Then the selected components are combined hierarchically to clusters by calling function clustCombi from the same package. The number of clusters is chosen by visual detection of the change point in the plot of the rescaled differences between successive entropy values, as suggested by Baudry et al. (2010). Furthermore, to compare our results to those obtained if a cluster distribution is modeled by a single normal distribution only, a sparse finite mixture model (Malsiner-Walli et al., 2015) is fitted to the data sets. These results are reported in the column SparseMix. A sparse hierarchical mixture of mixtures model with $L = 4$ and $L = 5$ is fitted to the data. The results for different values of $L$ are compared to investigate the robustness of the recommended default choice of $L = 4$. For each estimation, MCMC sampling is run for 4000 iterations after a burn-in of 4000 iterations.

As can be seen in Table 3 for all data sets we are able to capture the data clusters quite well both in terms of the estimated number of clusters and the clustering quality. This holds especially for the Yeast data set where clustCombi completely fails and estimates six clusters where the clustering quality has a negative adjusted Rand index $(adj = -0.02)$, i.e. is “worse than would be expected by guessing”, as noted by Franczak et al. (2012). Also, it can be noted that if the number of subcomponents $L$ is increased to 5, the estimation results remain quite stable, see last column in Table 3.
Table 3: Results for the estimated number of data clusters $\hat{K}_0$ for various benchmark data sets, using the functions \texttt{Mclust} to fit a standard mixture model and \texttt{clustCombi} to estimate a mixture with combined components (Fraley et al., 2012) (column “\texttt{Mclust}”), using a sparse finite mixture model (Malsiner-Walli et al., 2015) (column “\texttt{SparseMix}”), and estimating a sparse hierarchical mixture of mixtures model (column “\texttt{SparseMixMix}”) with $K = 10$ clusters and number of subcomponents $L = 4, 5$. Hyperparameter specifications are selected as described in Sections 2.1 and 2.2. \textit{adj} reports the adjusted Rand index, (“1” corresponds to perfect classification), and $er$ gives the proportion of misclassified observations.

6 Discussion

We propose suitable priors for fitting an identified mixture of normal mixtures model within the Bayesian framework of model-based clustering. This approach allows for (1) automatic determination of the number of clusters and (2) semi-parametric approximation of non-Gaussian cluster distributions by mixtures of normals. We only require the assumption that the cluster distributions are dense and connected. Our approach consists in the specification of highly structured informative priors on all model parameters. Through the prior specification, a rigid hierarchical structure on the normal subcomponents is imposed which allows for simultaneous estimation of clusters and their approximating distributions. This is in contrast to the two-step merging approaches, where in the first step the data
distribution is approximated by a suitable normal mixture model. However, because this approximation is made without making allowance for the data clusters which will be reconstructed only in the second step of the procedure, the general cluster structure might be missed by these approaches.

As we noted in our simulation studies, the way in which the cluster mixture distributions are modeled is crucial for the clustering result. Enforcing overlapping subcomponent densities is essential in order to avoid that a single subcomponent becomes too concentrated on a single observation as this would have the effect that the probability of this observation to belong to “its” cluster mixture distribution eventually decreases. Also, enforcing that to all subcomponents observations are assigned during MCMC sampling is important as the estimation of empty subcomponents would bias the cluster mixture distribution because of the “prior” subcomponents. For modeling large, overlapping subcomponent densities, crucial model parameters are the a-priori specified covariance matrix of the subcomponent means and the scale matrix of the Inverted Wishart prior for the subcomponent covariance matrices. We select both crucial hyperparameters based on considerations resulting from the variance decomposition of a mixture of mixtures model.

We found a prior setting which is able to capture dense and connected data clusters in a range of benchmark data sets. However, if interest lies in detection of different cluster shapes, also a different tuning of the prior parameters may be required. Therefore, it would be interesting to investigate in more detail how we can use certain prior settings to estimate certain kinds of data clusters. Then it would be possible to give recommendations which prior settings have to be used in order to capture certain types of data clusters. For instance, mixtures of shifted asymmetric Laplace (SAL) distributions, introduced by Franczak et al. (2012), have cluster distributions which are non-dense and have a strongly asymmetric shape with comet-like tails. In this case the prior specifications given in Sections 2.1 and 2.2 are not able to capture the clusters. However, they can be tuned in such a way to capture also this special kind of data clusters, as can be seen in the example given in Appendix E.

Our approach to estimate the number of clusters works well if the number of observations and the number of variables are not too large. For larger data sets we noticed that the
standard Gibbs sampler tends to get stuck in a local maximum of the posterior distribution and superfluous clusters do not become empty during sampling. However, we leave the development a modified sampling procedure for big data situations for future research.

A MCMC sampling scheme

Estimation of a sparse hierarchical mixture of mixtures model is performed through MCMC sampling based on data augmentation and Gibbs sampling. To indicate the cluster to which each observation belongs, latent allocation variables \( \mathbf{S} = (S_1, \ldots, S_N) \) taking values in \( \{1, \ldots, K\}^N \) are introduced such that

\[
p(y_i|\theta_1, \ldots, \theta_K, S_i = k) = p_k(y_i|\theta_k), \quad \text{and} \quad Pr(S_i = k|\eta) = \eta_k.
\]

Additionally, to indicate the subcomponent to which an observation within a cluster is assigned to, latent allocation variables \( \mathbf{I} = (I_1, \ldots, I_N) \) taking values in \( \{1, \ldots, L\}^N \) are introduced such that

\[
p_k(y_i|\theta_k, S_i = k, I_i = l) = f_N(y_i|\mu_{kl}, \Sigma_{kl}) \quad \text{and} \quad Pr(I_i = l|S_i = k, w_k) = w_{kl}.
\]

Based on the priors specified in Section 2, with fixed hyperparameters \( e_0, d_0, c_0, g_0, G_0, B_0, m_0, M_0, \nu, \)

the latent variables and parameters \( (\mathbf{S}, \mathbf{I}, \eta, \mathbf{w}_k, \mu_{kl}, \Sigma_{kl}, C_{0k}, b_{0k}, \lambda_{kj})\), \( k = 1, \ldots, K, \)
\( l = 1, \ldots, L, j = 1, \ldots, r \), are sampled from the posterior distribution using the following Gibbs sampling scheme. Note that conditional distributions given do not indicate that conditioning is also on the fixed hyperparameters.

(1) Sampling steps on the level of the cluster distribution:

(a) Parameter simulation step conditional on the classifications \( \mathbf{S} \). Sample \( \eta|\mathbf{S} \) from \( \text{Dir}(e_1, \ldots, e_K) \), \( e_k = e_0 + N_k \), where \( N_k = \# \{S_i|S_i = k\} \) is the number of observations allocated to cluster \( k \).

(b) Classification step for each observation \( y_i \) conditional on cluster-specific parameters: For each \( i = 1, \ldots, N \) sample the cluster assignment \( S_i \) from

\[
P(S_i = k|y_i, \theta_k, \eta_k) \propto \eta_kp_k(y_i|\theta_k), \quad k = 1, \ldots, K,
\]
where \( p_k(y_i|\theta_k) \) is the semi-parametric mixture approximation of the cluster density:

\[
p_k(y_i|\theta_k) = \sum_{l=1}^{L} w_{kl} f_N(y_i|\mu_{kl}, \Sigma_{kl}).
\]

Note that clustering of the observations is performed on the upper level of the model, using a collapsed Gibbs step, where the latent, within-cluster allocation variables \( I \) are integrated out.

(2) Within each cluster \( k, k = 1, \ldots, K \):

(a) **Classification step** for all observations \( y_i \), assigned to cluster \( k \) (i.e. \( S_i = k \)), conditional on the subcomponent weights and the subcomponent-specific parameters:

For each \( i = 1, \ldots, N \) sample \( I_i \) from

\[
P(I_i = l|y_i, \theta_k, S_i = k) \propto w_{kl} f_N(y_i|\mu_{kl}, \Sigma_{kl}), \ l = 1, \ldots, L.
\]

(b) **Parameter simulation step** conditional on the classifications \( I \) and \( S \):

i. Sample \( w_k|I, S \) from \( \text{Dir}(d_k1, \ldots, d_kL) \), \( d_{kl} = d_0 + N_{kl} \), where \( N_{kl} = \#\{I_i = l|S_i = k\} \) is the number of observations allocated to subcomponent \( l \) in cluster \( k \).

ii. For \( l = 1, \ldots, L \): Sample \( \Sigma^{-1}_{kl}|S, I, \mu_{kl}, C_{0k}, y \sim \mathcal{W}(c_{kl}, C_{kl}) \), where

\[
c_{kl} = c_0 + N_{kl}/2,
\]

\[
C_{kl} = C_{0k} + \frac{1}{2} \sum_{i: I_i=l, S_i=k} (y_i - \mu_{kl})(y_i - \mu_{kl})'.
\]

iii. For \( l = 1, \ldots, L \): Sample \( \mu_{kl}|S, I, b_{0k}, \Sigma_{kl}, \Lambda_k, y \sim \mathcal{N}(b_{kl}, B_{kl}) \), where

\[
B_{kl} = (\tilde{B}_{0k}^{-1} + N_{kl}\Sigma_{kl}^{-1})^{-1},
\]

\[
b_{kl} = B_{kl}(\tilde{B}_{0k}^{-1}b_{0k} + \Sigma_{kl}^{-1}N_{kl}y_{kl}),
\]

where \( \tilde{B}_{0k} = \sqrt{\Lambda_k}B_0\sqrt{\Lambda_k} \), \( \Lambda_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kr}) \), and \( y_{kl} \) is the subcomponent mean \( 1/N_{kl} \sum_{i: I_i=l, S_i=k} y_i \).
(3) For each cluster $k$, $k = 1, \ldots, K$: Sample hyperparameters $\lambda_{kj}$, $C_{0k}$, $b_{0k}$ from their full conditionals:

(a) For $j = 1, \ldots, r$: Sample $\lambda_{kj}|b_{0k}, \mu_{k1}, \ldots, \mu_{kL} \sim \mathcal{GIG}(p_{kL}, a_{kj}, b_{kj})$, where $\mathcal{GIG}$ is the generalized inverted Gaussian distribution and

\[
p_{kL} = -L/2 + \nu, \quad a_{kj} = 2\nu, \quad b_{kj} = \sum_{l=1}^{L} (\mu_{kl,j} - b_{0k,j})^2 / B_{0,jj}.
\]

(b) Sample $C_{0k}|\Sigma_{k1}, \ldots, \Sigma_{kL} \sim \mathcal{W}_r(g_0 + Lc_0, G_0 + \sum_{l=1}^{L} \Sigma_{kl}^{-1})$.

(c) Sample $b_{0k}|\tilde{B}_{0k}, \mu_{k1}, \ldots, \mu_{kL} \sim \mathcal{N}_r(\tilde{m}_k, \tilde{M}_k)$, where

\[
\tilde{M}_k = (M_0^{-1} + L\tilde{B}_{0k}^{-1})^{-1}, \quad \tilde{m}_k = \tilde{M}_k \left( M_0^{-1}m_0 + \tilde{B}_{0k}^{-1} \sum_{l=1}^{L} \mu_{kl} \right).
\]

B Clustering the draws in the point process representation

To illustrate the clustering of the draws in the point process representation, a sparse hierarchical mixture of mixtures model with $K = 10$ clusters and $L = 4$ subcomponents is fitted to the AIS data set (see Figure 1 and Section 5). The point process representation of the weighted cluster mean draws $\mu_k^{(m)} = \sum_{l=1}^{L} w_{kl}^{(m)} \mu_{kl}^{(m)}$ of all clusters, including empty clusters, is shown in Figure B.6 on the left-hand side. Since a lot of draws are sampled from empty clusters, i.e. from the prior distribution, the plot shows a cloud of overlapping posterior distributions where no cluster structure can be distinguished. However, since during MCMC sampling in almost all iterations only two clusters were non-empty, the estimated number of clusters is $\hat{K}_0 = 2$. Thus all draws generated in iterations where the number of non-empty clusters is different from two and all draws from empty clusters are removed. The point process representation of the remaining cluster-specific draws is shown in the
Figure B.6: AIS data set, $K = 10$, $L = 4$, $\phi_B = 0.5$, $\phi_W = 0.1$: Point process representation of the cluster centers $\mu_k$ of all 10 clusters (left-hand side) and only from those where $\hat{K}_0 = 2$ (middle). Right-hand side: Point process representation of the means of all subcomponents forming the cluster in the bottom right in the plot in the middle.

scatter plot in the middle of Figure B.6. Now the draws cluster around two well-separated points, and the two clusters can be easily identified.

To illustrate the subcomponent distributions which are used to approximate the cluster distribution, the point process representation of the subcomponent means are shown in Figure B.6 on the right-hand side for the cluster discernible at the bottom right in Figure B.6 in the middle. The plot clearly indicates that all subcomponent means are shrunken toward the cluster mean as the variation of the subcomponent means is about the same as the variation of the cluster means.

C Simulation setups

C.1 Simulation setup I

The four clusters are generated by sampling 800 observations from an eight-component normal mixture with component means

$$
(\mu_1 \; \mu_2 \; \ldots \; \mu_8) = 
\begin{pmatrix}
6 & 4 & 8 & 22.5 & 20 & 22 & 22 & 6.5 \\
1.5 & 6 & 6 & 1.5 & 8 & 31 & 31 & 29
\end{pmatrix},
$$
variance-covariance matrices

\[
\Sigma_1 = \begin{pmatrix} 4.84 & 0 \\ 0 & 2.89 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 3.61 & 5.05 \\ 5.05 & 14.44 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 3.61 & -5.05 \\ -5.05 & 14.44 \end{pmatrix},
\]

\[
\Sigma_4 = \begin{pmatrix} 12.25 & 0 \\ 0 & 3.24 \end{pmatrix}, \quad \Sigma_5 = \begin{pmatrix} 3.24 & 0 \\ 0 & 12.25 \end{pmatrix}, \quad \Sigma_6 = \begin{pmatrix} 14.44 & 0 \\ 0 & 2.25 \end{pmatrix},
\]

\[
\Sigma_7 = \begin{pmatrix} 2.25 & 0 \\ 0 & 17.64 \end{pmatrix}, \quad \Sigma_8 = \begin{pmatrix} 2.25 & 4.2 \\ 4.2 & 16.0 \end{pmatrix},
\]

and weight vector \( \eta = 1/4(1/3, 1/3, 1/3, 1/2, 1/2, 1/2, 1/2, 1) \).

The first three normal distributions generate the triangle-shaped cluster, the next two the L-shaped cluster, and the last three distributions the cross-shaped and the elliptical cluster.

### C.2 Simulation setup II

For each data set 300 observations are sampled from a normal mixture with component means

\[
(\mu_1 \quad \mu_2 \quad \mu_3) = \begin{pmatrix} 2 & 4.2 & 7.8 \\ 2 & 4.2 & 7.8 \end{pmatrix},
\]

variance-covariance matrices \( \Sigma_1 = \Sigma_2 = \Sigma_3 = I_2 \) and equal weights \( \eta = (1/3, 1/3, 1/3) \).

### D Description of the data sets

The following data sets are investigated. The Yeast data set \cite{nakai1991} aims at predicting the cellular localization sites of proteins and can be downloaded from the UCI machine learning repository \cite{bache2013}. As in Franczak et al. \cite{franczak2012}, we aim at distinguishing between the two localization sites CYT (cytosolic or cytoskeletal) and ME3 (membrane protein, no N-terminal signal) by considering a subset of three variables, namely McGeoch’s method for signal sequence (mcg), the score of the ALOM membrane spanning region prediction program (alm) and the score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins (vac).
The Flea beetles data set (Lubischew 1962) considers 6 physical measurements of 74 male flea beetles belonging to three different species. It is available in the R package DPpackage (Jara et al. 2011).

The Australian Institute of Sport (AIS) data set (Cook and Weisberg 1994) consists of 11 physical measurements on 202 athletes (100 female and 102 male). It has previously been analyzed by Cook and Weisberg (1994). As in Lee and McLachlan (2013), we only consider three variables, namely body mass index (BMI), lean body mass (LBM) and the percentage of body fat (Bfat). The data set is contained in the R package locfit (Loader 2013).

The Breast Cancer Wisconsin (Diagnostic) data set (Mangasarian et al. 1995) describes characteristics of the cell nuclei present in images. The clustering aim is to distinguish between benign and malignant tumors. It can be downloaded from the UCI machine learning repository. Following Fraley and Raftery (2002) and Viroli (2010) we use a subset of three attributes: extreme area, extreme smoothness, and mean texture. Additionally, we scaled the data.

### E  Fitting a mixture of two SAL distributions

Although it is not the purpose of our approach to capture non-dense data clusters, we apply it to the challenging cluster shapes generated by SAL distributions, which are introduced by Franczak et al. (2012) in order to capture asymmetric data clusters with outliers. We sampled data from a mixture of two SAL distributions according to Section 4.2 in Franczak et al. (2012). The data set is shown in Figure E.7 on the left-hand side.

If we fit a sparse hierarchical mixture of mixtures model with \( K = 10 \) clusters and \( L = 4 \) subcomponents and priors and hyperpriors specified as in Sections 2.1 and 2.2, four clusters are estimated, as can be seen in the middle plot of Figure E.7. Evidently, the standard prior setting, tuned to capture dense homogeneous data clusters, performs badly for this kind of clusters. Thus, in order to take the specific data cluster shapes into account, we adjust the prior specifications accordingly. A data cluster generated by a SAL distribution is not homogeneously dense, it rather consists of a relatively dense kernel on one side of the cluster and a non-dense, light and comet-like tail with possibly extreme observations on
the other side. Therefore within a cluster, subcomponents with very different covariance matrices are required in order to fit the whole cluster distribution. Since specification of hyperpriors on $\lambda_{kj}$ and $C_{0k}$ has a smoothing and balancing effect on the subcomponent densities, we omit these hyperprior specifications, and choose fixed values for $k = 1, \ldots, K$, i.e. $C_{0k} = g_0 \cdot G_0^{-1}$ and $\lambda_{kj} \equiv 1, j = 1, \ldots, r$.

Additionally, in order to reach also extreme points, we increase both the number of subcomponents to $L = 5$ and the a-priori variability explained by the subcomponent means to $\phi_W = 0.2$. At the same time we adjust the proportion of heterogeneity explained by the cluster means by decreasing $\phi_B$ to 0.4, thus keeping the subcomponent covariance matrices large. If we estimate again a sparse hierarchical mixture of mixtures model with these modified prior settings, the two clusters can be identified, see Figure E.7 on the right-hand side.

![Figure E.7](image)

Figure E.7: Samples from a mixture of two SAL distributions (left-hand side), the estimated clusters for $K = 10, L = 4, \phi_B = 0.5, \phi_W = 0.1, \nu_1 = \nu_2 = 10$ (middle), and for $K = 10, L = 5, \phi_B = 0.4, \phi_W = 0.2$, with fixed hyperparameters $C_{0k}$ and $\lambda_{kl}$ (right-hand side).

References

Argiento, R., A. Cremaschi, and A. Guglielmi (2014). A ”density-based” algorithm for cluster analysis using species sampling Gaussian mixture models. *Journal of Computational and Graphical Statistics* 23(4), 1126–1142.
Bache, K. and M. Lichman (2013). UCI machine learning repository. URL http://archive.ics.uci.edu/ml

Bartolucci, F. (2005). Clustering univariate observations via mixtures of unimodal normal mixtures. *Journal of Classification* 22(2), 203–219.

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

Chan, C., F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler (2008). Statistical mixture modelling for cell subtype identification in flow cytometry. *Cytometry, A* 73, 693–701.

Chan, C., L. Lin, J. Frelinger, V. Hebert, D. Gagnon, C. Landry, R. P. Sékaly, J. Enzor, J. Staats, K. J. Weinhold, M. Jaimes, and M. West (2010). Optimization of a highly standardized carboxyfluorescein succinimidyl ester flow cytometry panel and gating strategy design with discriminative information measure evaluation. *Cytometry A* 77, 1126–1136.

Cook, R. D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. Wiley.

Cron, A. J., C. Gouttefangeas, J. Frelinger, L. Lin, S. K. Singh, C. M. Britten, M. J. P. Welters, S. H. van der Burg, M. West, and C. Chan (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLOS Computational Biology* 9(7), 1–14.

Di Zio, M., U. Guarnera, and R. Rocci (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data Analysis* 51(5), 2573–2585.

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

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi and J. S. Rustagi (Eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernov on His Sixtieth Birthday*, pp. 287–302. New York: Academic Press.
Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (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.

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

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

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

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

Frühwirth-Schnatter, S. and S. Pyne (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* 11(2), 317–336.

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

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS136: A k-means clustering algorithm. *Applied Statistics* 28(1), 100–108.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* 4(1), 3–34.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2(1), 193–218.
Jara, A., T. E. Hanson, F. A. Quintana, P. Müller, and G. L. Rosner (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software* 40(5), 1.

Lee, S. and G. J. McLachlan (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods and Applications* 22(4), 427–454.

Lee, S. and G. J. McLachlan (2014). Finite mixtures of multivariate skew $t$-distributions: Some recent and new results. *Statistics and Computing* 24(2), 181–202.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 3(14), 547–568.

Lin, L., C. Chan, S. R. Hadrup, T. M. Froesig, Q. Wang, and M. West (2013). Hierarchical Bayesian mixture modelling for antigen-specific T-cell subtyping in combinatorially encoded flow cytometry studies. *Statistical Applications in Genetics and Molecular Biology* 12, 309–331.

Loader, C. (2013). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-9.1.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics* 18(4), 455–477.

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, pp. 281–297.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2015). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 1–22.

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research* 43(4), 570–577.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *The Annals of Statistics* 20, 712–736.
McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley.

Nakai, K. and M. Kanehisa (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Bioinformatics* 11(2), 95–110.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics* 32(5), 2044–2073.

Quintana, F. A. and P. L. Iglesias (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Ser. B* 65, 557–574.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336), 846–850.

Rodriguez, A. and R. Vuppala (2014). Supervised and semi-supervised classification using nested nonparametric mixture models. *Unpublished manuscript*.

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

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. Ph. D. thesis, University of Oxford.

Suchard, M. A., Q. Wang, C. Chan, J. Frelinger, A. J. Cron, and M. West (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics* 19, 419–438.

Viroli, C. (2010). Dimensionally reduced model-based clustering through mixtures of factor mixtures analyzers. *Journal of Classification* 27(3), 363–388.

Vrbik, I. and P. D. McNicholas (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis* 71, 196–210.