Gertraud Malsiner-Walli and Sylvia Frühwirth-Schnatter and Bettina Grün

Identifying mixtures of mixtures using Bayesian estimation

Article (Accepted for Publication)
(Refereed)

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

Abstract

The use of a finite mixture of normal distributions 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 based on sparse finite mixtures to achieve identifiability. We specify a hierarchical prior where the hyperparameters are carefully selected such that they are reflective of the cluster structure aimed at. In addition, this prior allows to estimate the model using standard MCMC sampling methods. In combination with a post-processing approach which resolves the label switching issue and results in an identified model, our approach allows to simultaneously (1) determine the number of clusters, (2) flexibly approximate the cluster distributions in a semi-parametric way using finite mixtures of normals and (3) identify cluster-specific parameters and classify observations. The proposed approach is illustrated in two simulation studies and on benchmark data sets.

Keywords: Dirichlet prior; Finite mixture model; Model-based clustering; Bayesian nonparametric mixture model; 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. A flexible approach to clustering data is based on mixture models, whereby the data in each mixture component are assumed to follow a parametric distribution with component-specific parameters varying over the components. 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 to identify homogeneous data clusters within a heterogeneous population. However, assuming such a simple parametric form for the component densities 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 densities such as the skew normal or skew-*t* distribution to capture non-Gaussian data clusters, see Frühwirth-Schnatter and Pyne (2010), Lee and McLachlan (2014) and Vrbik and McNicholas (2014), among others. 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. In addition, the shape of the cluster densities can be of a form which is not easily captured by a 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 mixture of mixtures, models the non-Gaussian cluster distributions themselves by Gaussian mixtures, exploiting the ability of normal mixtures to accurately approximate a wide class of probability distributions. Compared to a mixture with Gaussian components, mixture of mixtures models impose a two-level hierarchical structure which is par-

ticularly 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, both within the framework of finite and infinite mixtures.

Statistical inference for finite mixtures is generally not easy due to problems such as label switching, spurious modes and unboundedness of the mixture likelihood (see e.g. Frühwirth-Schnatter, 2006, Chapter 2), but estimation of a mixture of mixtures model is particularly challenging due to additional identifiability issues. Since exchanging subcomponents 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 from the mixture likelihood in the absence of additional information. For example, strong identifiability constraints on the locations and the covariance matrices of the Gaussian components were imposed by Bartolucci (2005) for univariate data and by Di Zio et al. (2007) for multivariate data to estimate finite mixtures of Gaussian mixtures.

A different strand of literature pursues the idea of creating meaningful clusters after having fitted a standard Gaussian mixture model to the data. The clusters are determined by successively merging components according to some criterion, e.g. the closeness of the means (Li, 2005), the modality of the obtained mixture density (Chan et al., 2008; Hennig, 2010), the degree of overlapping measured by misclassification probabilities (Melnykov, 2016) or the entropy of the resulting partition (Baudry et al., 2010). However, such two-step approaches might miss the general cluster structure, see Appendix E for an example.

In the present paper, we identify the mixture of mixtures model within a Bayesian framework through a hierarchical prior construction and propose a method to simultaneously select a suitable number of clusters. In our approach both the identification of the model and the estimation of the number of clusters is achieved by employing a selectively informative prior parameter setting on the model parameters.

Our choice of prior parameters is driven by assumptions on the cluster shapes assumed to be present in the data, thus being in line with Hennig (2010) who emphasizes that, *"it rather has to be decided by the statistician under which conditions different Gaussian mixture components should be regarded as a common cluster"*. This prior specification introduces dependence among the 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. The choice of the hyperparameters of this hierarchical prior turns out to be crucial in achieving identification and is guided by a variance decomposition of the data.

Regarding the estimation of the number of clusters, a sparse hierarchical mixture of mixtures model is derived as an extension of the sparse finite mixture model introduced in Malsiner-Walli et al. (2016). There, based on theoretical results derived by Rousseau and Mengersen (2011), an overfitting Gaussian mixture with K components is specified where a sparse prior on the mixture weights has the effect of assigning the observations to fewer than K components. Thus, the number of clusters can be estimated by the most frequent number of non-empty components encountered during Markov chain Monte Carlo (MCMC) sampling. In this paper, rather than using a single multivariate Gaussian distribution, we model the component densities in a semi-parametric way through a Gaussian mixture distribution, and again use a sparse prior on the cluster weights to automatically select a suitable number of clusters on the upper level.

Specifying a sparse prior on the weights is closely related to Bayesian nonparametric (BNP) Gaussian mixture models such as Dirichlet process mixtures (DPMs; Ferguson, 1983; Escobar and West, 1995). The sparse prior on the cluster weights induces clustering of the observations, similar as for DPMs which have been applied in a clustering context by Quintana and Iglesias (2003), Medvedovic et al. (2004) and Dahl (2006), among others. The hierarchical mixture of mixtures model we introduce is similar to hierarchical BNP approaches such as the hierarchical DPM (Teh et al., 2006). Very closely related BNP approaches are infinite mixtures of infinite Gaussian densities such as the nested DPM (Rodriguez et al., 2008), the infinite mixture of infinite Gaussian mixtures (Yerebakan et al., 2014), and species mixture models (Argiento et al., 2014)

which directly work on the partition of the data. We discuss in Sections 2.4 and 3.1 similarities as well as differences between our approach and BNP models.

We finally note that the implementation effort to estimate our model is moderate and standard MCMC methods based on data augmentation and Gibbs sampling (see Frühwirth-Schnatter, 2006) can be used. Several approaches proposed in the literature can be used to post-process the MCMC draws in order to obtain a clustering of the data and also to allow for cluster-specific inference. For our simulation studies and applications we adapt and extend the method suggested by Frühwirth-Schnatter (2006, 2011) which determines a unique labeling for the MCMC draws by clustering the draws in the point process representation.

The rest of the article is organized as follows. Section 2 describes the proposed strategy, including detailed prior specifications, and relates our method to the two-layer BNP approaches in Rodriguez et al. (2008) and Yerebakan et al. (2014). Clustering and model estimation issues are discussed in Section 3. The performance of the proposed strategy is evaluated in Section 4 for various benchmark data sets. Section 5 concludes.

2 Sparse hierarchical mixture of mixtures model

2.1 Model definition

Following previous work on hierarchical mixtures of mixtures, we assume that N observations $\mathbf{y}_i$, $i = 1, \ldots, N$ of dimension $\dim(\mathbf{y}_i) = r$ are drawn independently from a finite mixture distribution with K components,

$$p(\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k p_k(\mathbf{y}_i|\boldsymbol{\theta}_k), \quad \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K), \tag{1}$$

with each component distribution $p_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ being a mixture of L normal subcomponents:

$$p_k(\mathbf{y}_i|\boldsymbol{\theta}_k) = \sum_{l=1}^{L} w_{kl} f_{\mathcal{N}}(\mathbf{y}_i|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}). \tag{2}$$

In order to distinguish the component distributions on the upper level from the Gaussian components on the lower level, we will refer to the former ones as “cluster distributions”. For clustering

the observations based on Bayes' rule, the cluster weights $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ and the cluster densities $p_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ on the upper level (1) are relevant.

Since the number of data clusters is unknown and needs to be inferred from the data, we assume that (1) is an overfitting mixture, i.e. the specified number of clusters K exceeds the number of clusters present in the data. Following the concept of sparse finite mixtures (Malsiner-Walli et al., 2016), we choose a symmetric Dirichlet distribution as prior for the weight distribution, i.e. $\boldsymbol{\eta}|e_0 \sim Dir_K(e_0)$, and base our choice of e_0 on the results of Rousseau and Mengersen (2011) concerning the asymptotic behavior of the posterior distribution of an overfitting mixture model. They show that this behavior is determined by the hyperparameter e_0 of the Dirichlet prior on the weights. In particular, they prove that, if $e_0 < d/2$, where d is the dimension of the cluster-specific parameters $\boldsymbol{\theta}_k$, then the posterior expectation of the weights associated with superfluous clusters asymptotically converges to zero.

Hence, we specify a sparse prior on the cluster weights $\boldsymbol{\eta}$ by choosing $e_0 \ll d/2$ so that superfluous clusters are emptied during MCMC sampling and the number of non-empty clusters on the cluster level is an estimator for the unknown number of data clusters. In this way, the specification of a sparse cluster weight prior in an overfitting mixture of mixtures model provides an "automatic tool" to select the number of clusters, avoiding the expensive computation of marginal likelihoods as, e.g., in Frühwirth-Schnatter (2004). Empirical results in Malsiner-Walli et al. (2016) indicate that e_0 needs to be chosen very small, e.g. $e_0 = 0.001$, to actually empty all superfluous clusters in the finite sample case.

On the lower level (2), in each cluster k, a semi-parametric approximation of the cluster distributions is achieved by mixing L multivariate Gaussian subcomponent densities $f_{\mathcal{N}}(\mathbf{y}_i|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl})$, $l = 1, \ldots, L$, according to the subcomponent weight vector $\mathbf{w}_k = (w_{k1}, \ldots, w_{kL})$. The cluster-specific parameter vector

$$\boldsymbol{\theta}_k = (\mathbf{w}_k, \boldsymbol{\mu}_{k1}, \ldots, \boldsymbol{\mu}_{kL}, \boldsymbol{\Sigma}_{k1}, \ldots, \boldsymbol{\Sigma}_{kL}) \tag{3}$$

consists of $\mathbf{w}_k$ as well as the means $\boldsymbol{\mu}_{kl}$ and covariance matrices $\boldsymbol{\Sigma}_{kl}$ of all Gaussian subcomponent densities. L is typically unknown, but as we are not interested in estimating the "true" number of subcomponents L forming the cluster, we only ensure that L is chosen sufficiently large to obtain an

accurate approximation of the cluster distributions. While the choice of L is not crucial to ensure a good model fit as long as L is sufficiently large, a too generous choice of L should be avoided for computational reasons as the computational complexity of the estimation increases with the number of subcomponents L.

By choosing the prior $\mathbf{w}_k \sim Dir_L(d_0)$ with $d_0 = d/2+2$, the approximation of the cluster density is obtained by filling all L subcomponents, thus avoiding empty subcomponents. This choice is motivated again by the results of Rousseau and Mengersen (2011) who show that, if $d_0 > d/2$, the posterior density asymptotically handles an overfitting mixture by splitting "true" components into two or more identical components.

2.2 Identification through hierarchical priors

When fitting the finite mixture model (1) with semi-parametric cluster densities given by (2), we face a special identifiability problem, since the likelihood is entirely agnostic about which subcomponents form a cluster. Indeed, the likelihood is completely ignorant concerning the issue which of the $K{\cdot}L$ components belong together, since (1) can be written as an expanded Gaussian mixture with $K{\cdot}L$ components with weights $\tilde{w}_{kl} = \eta_k w_{kl}$,

$$p(\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \tilde{w}_{kl} f_{\mathcal{N}}(\mathbf{y}_i|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}). \tag{4}$$

These $K{\cdot}L$ components can be permuted in $(K{\cdot}L)!$ different ways and the resulting ordering can be used to group them into K different cluster densities, without changing the mixture likelihood (4). Hence, the identification of (1), up to label switching on the upper level, hinges entirely on the prior distribution.

Subsequently, we suggest a hierarchical prior that addresses these issues explicitly. Conditional on a set of fixed hyperparameters $\phi_0 = (e_0, d_0, c_0, g_0, \mathbf{G}_0, \mathbf{B}_0, \mathbf{m}_0, \mathbf{M}_0, \nu)$, the weight distribution $\boldsymbol{\eta}|e_0 \sim Dir_K(e_0)$ and the K cluster-specific parameter vectors $\boldsymbol{\theta}_k|\phi_0 \overset{iid}{\sim} p(\boldsymbol{\theta}_k|\phi_0)$ are independent a priori, i.e.:

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\phi_0) = p(\boldsymbol{\eta}|e_0) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\phi_0). \tag{5}$$

This prior formulation ensures that the K non-Gaussian cluster distributions of the upper level mixture (1) are invariant to permutations. Within each cluster k, the prior distribution $p(\boldsymbol{\theta}_k|\phi_0)$ admits the following block independence structure:

$$p(\boldsymbol{\theta}_k|\phi_0) = p(\mathbf{w}_k|d_0)p(\boldsymbol{\mu}_{k1},\ldots,\boldsymbol{\mu}_{kL}|\mathbf{B}_0,\mathbf{m}_0,\mathbf{M}_0,\nu)p(\boldsymbol{\Sigma}_{k1},\ldots,\boldsymbol{\Sigma}_{kL}|c_0,g_0,\mathbf{G}_0), \tag{6}$$

where $\mathbf{w}_k|d_0 \overset{iid}{\sim} Dir_L(d_0)$. Conditional on ϕ_0, the subcomponent means $\boldsymbol{\mu}_{k1},\ldots,\boldsymbol{\mu}_{kL}$ are dependent a priori as are the subcomponent covariance matrices $\boldsymbol{\Sigma}_{k1},\ldots,\boldsymbol{\Sigma}_{kL}$. However, they are assumed to be exchangeable to guarantee that within each cluster k, the L Gaussian subcomponents in (2) can be permuted without changing the prior.

To create this dependence, a hierarchical "random effects" prior is formulated, where, on the upper level, conditional on the fixed upper level hyperparameters $(g_0,\mathbf{G}_0,\mathbf{m}_0,\mathbf{M}_0,\nu)$, cluster specific random hyperparameters $(\mathbf{C}_{0k},\mathbf{b}_{0k})$, and $\boldsymbol{\Lambda}_k = \text{diag}(\lambda_{k1},\ldots,\lambda_{kr})$, are drawn independently for each $k = 1,\ldots,K$ from a set of three independent base distributions:

$$\mathbf{C}_{0k}|g_0,\mathbf{G}_0 \overset{iid}{\sim} \mathcal{W}_r(g_0,\mathbf{G}_0), \qquad \mathbf{b}_{0k}|\mathbf{m}_0,\mathbf{M}_0 \overset{iid}{\sim} \mathcal{N}_r(\mathbf{m}_0,\mathbf{M}_0), \qquad (\lambda_{k1},\ldots,\lambda_{kr})|\nu \overset{iid}{\sim} \mathcal{G}(\nu,\nu), \tag{7}$$

where $\mathcal{N}_r()$ and $\mathcal{W}_r()$ denote the r-multivariate normal and Wishart distribution, respectively, and $\mathcal{G}()$ the gamma distribution, parametrized such that $E(\lambda_{kl}|\nu) = 1$.

On the lower level, conditional on the cluster specific random hyperparameters $(\mathbf{C}_{0k},\mathbf{b}_{0k},\boldsymbol{\Lambda}_k)$ and the fixed lower level hyperparameters $(\mathbf{B}_0,c_0)$, the L subcomponent means $\boldsymbol{\mu}_{kl}$ and covariance matrices $\boldsymbol{\Sigma}_{kl}$ are drawn independently for all $l = 1,\ldots,L$:

$$\boldsymbol{\mu}_{kl}|\mathbf{B}_0,\mathbf{b}_{0k},\boldsymbol{\Lambda}_k \overset{iid}{\sim} \mathcal{N}_r(\mathbf{b}_{0k},\sqrt{\boldsymbol{\Lambda}_k}\mathbf{B}_0\sqrt{\boldsymbol{\Lambda}_k}), \quad \boldsymbol{\Sigma}_{kl}^{-1}|c_0,\mathbf{C}_{0k} \overset{iid}{\sim} \mathcal{W}_r(c_0,\mathbf{C}_{0k}). \tag{8}$$

2.3 Tuning the hyperparameters

To identify the mixture of mixtures model given in (1) and (2) through the prior defined in Section 2.2, the fixed hyperparameters ϕ_0 have to be chosen carefully. In addition, we select them in a way to take the data scaling into account, avoiding the need to standardize the data prior to data analysis.

First, it is essential to clarify what kind of shapes and forms are aimed at as cluster distributions.

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 distribution, whereas different clusters should be located well-separated from each other, i.e. here large "gaps" between the cluster distributions 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 entire cluster. 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 of normal subcomponents. Various 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, ensuring that they strongly overlap at the same time. 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 then 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 a Bayesian framework, we need to translate these modeling purposes into appropriate choices of hyperparameters. On the upper level, the covariance matrix $\mathbf{M}_0$ controls the amount of prior shrinkage of the cluster centers $\mathbf{b}_{0k}$ toward the overall data center $\mathbf{m}_0$, which we specify as the midpoint of the data. To obtain a prior, where the cluster centers $\mathbf{b}_{0k}$ are allowed to be widely spread apart and almost no shrinkage toward $\mathbf{m}_0$ takes place, we choose $\mathbf{M}_0 \gg \mathbf{S}_y$, where $\mathbf{S}_y$ is the sample covariance matrix of all data, e.g. $\mathbf{M}_0 = 10\mathbf{S}_y$.

Our strategy for appropriately specifying the hyperparameters $\mathbf{G}_0$ and $\mathbf{B}_0$ is based on the variance decomposition of the mixture of mixtures model, which splits $Cov(\mathbf{Y})$ into the different

sources of variation. For a finite mixture model with K clusters, as given in (1), the total heterogeneity $Cov(\mathbf{Y})$ can be decomposed in the following way (Frühwirth-Schnatter, 2006, p. 170):

$$Cov(\mathbf{Y}) = \sum_{k=1}^{K} \eta_k \boldsymbol{\Sigma}_k + \sum_{k=1}^{K} \eta_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k' - \boldsymbol{\mu}\boldsymbol{\mu}' = (1 - \phi_B) Cov(\mathbf{Y}) + \phi_B Cov(\mathbf{Y}), \tag{9}$$

where the cluster means $\boldsymbol{\mu}_k$ and the cluster covariance matrices $\boldsymbol{\Sigma}_k$ are the first and second moments of the cluster distribution $p_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ and $\boldsymbol{\mu} = \sum_k \eta_k \boldsymbol{\mu}_k$ is the mixture mean. In this decomposition ϕ_B is the proportion of the total heterogeneity explained by the variability of the cluster means $\boldsymbol{\mu}_k$ and $(1 - \phi_B)$ is the proportion explained by the average variability within the clusters. The larger ϕ_B, the more the clusters are separated, as illustrated in Figure 1 for a three-component standard Gaussian mixture with varying values of ϕ_B.

For a mixture of mixtures model, the heterogeneity $(1 - \phi_B)Cov(\mathbf{Y})$ explained within a cluster can be split further into two sources of variability, namely the proportion ϕ_W explained by the variability of the subcomponent means $\boldsymbol{\mu}_{kl}$ around the cluster center $\boldsymbol{\mu}_k$, and the proportion $(1 - \phi_W)$ explained by the average variability within the subcomponents:

$$\begin{aligned} Cov(\mathbf{Y}) &= \sum_{k=1}^{K} \eta_k \boldsymbol{\Sigma}_k + \sum_{k=1}^{K} \eta_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k' - \boldsymbol{\mu}\boldsymbol{\mu}' \\ &= \sum_{k=1}^{K} \eta_k \sum_{l=1}^{L} w_{kl} \boldsymbol{\Sigma}_{kl} + \sum_{k=1}^{K} \eta_k \left(\sum_{l=1}^{L} w_{kl} \boldsymbol{\mu}_{kl} \boldsymbol{\mu}_{kl}' - \boldsymbol{\mu}_k \boldsymbol{\mu}_k' \right) + \sum_{k=1}^{K} \eta_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k' - \boldsymbol{\mu}\boldsymbol{\mu}' \\ &= (1 - \phi_W)(1 - \phi_B) Cov(\mathbf{Y}) + \phi_W (1 - \phi_B) Cov(\mathbf{Y}) + \phi_B Cov(\mathbf{Y}). \end{aligned} \tag{10}$$

Based on this variance decomposition we select the proportions ϕ_B and ϕ_W and incorporate them into the specification of the hyperparameters of our hierarchical prior.

ϕ_B defines the proportion of variability explained by the different cluster means. We suggest to specify ϕ_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 ϕ_B large. However, if ϕ_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 ϕ_B, estimation of

tight subcomponent densities would result, undermining our modeling aims.

ϕ_W defines the proportion of within-cluster variability explained by the subcomponent means. ϕ_W also controls how strongly the subcomponent means are pulled together and influences the overlap of the subcomponent densities. To achieve strong shrinkage of the subcomponent means toward the cluster center, we select small values of ϕ_W, e.g. $\phi_W = 0.1$. Larger values of ϕ_W may introduce gaps within a cluster, which we want to avoid.

Given ϕ_B and ϕ_W, we specify the scale matrix $\mathbf{G}_0$ of the prior on $\mathbf{C}_{0k}$ such that the a priori expectation of the first term in the variance decomposition (10), given by

$$E\left(\sum_{k=1}^{K}\eta_k\sum_{l=1}^{L}w_{kl}\mathbf{\Sigma}_{kl}\right)=\sum_{k=1}^{K}E(\eta_k)\sum_{l=1}^{L}E(w_{kl})E(E(\mathbf{\Sigma}_{kl}|\mathbf{C}_{0k}))=g_0/(c_0-(r+1)/2)\mathbf{G}_0^{-1},$$

matches the desired amount of heterogeneity explained by a subcomponent:

$$g_0/(c_0-(r+1)/2)\mathbf{G}_0^{-1}=(1-\phi_W)(1-\phi_B)Cov(\mathbf{Y}). \tag{11}$$

We replace $Cov(\mathbf{Y})$ in (11) with the main diagonal of the sample covariance $\mathbf{S}_y$ to take only the scaling of the data into account (see e.g. Frühwirth-Schnatter, 2006). This gives the following specification for $\mathbf{G}_0$:

$$\mathbf{G}_0^{-1}=(1-\phi_W)(1-\phi_B)(c_0-(r+1)/2)/g_0\cdot\mathrm{diag}(\mathbf{S}_y). \tag{12}$$

Specification of the prior of the subcomponent covariance matrices $\mathbf{\Sigma}_{k1},\ldots,\mathbf{\Sigma}_{kL}$ is completed by defining the scalar prior hyperparameters c_0 and g_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 $\mathbf{\Sigma}_{kl}\mathbf{\Sigma}_{km}^{-1}$ are bounded away from 0 avoiding singular matrices. We set $c_0 = 2.5+(r-1)/2$ to allow for a large variability of $\mathbf{\Sigma}_{kl}$. The Wishart density is regular if $g_0 > (r-1)/2$ and in the following we set $g_0 = 0.5+(r-1)/2$.

Regarding the prior specification of the subcomponent means $\boldsymbol{\mu}_{k1},\ldots,\boldsymbol{\mu}_{kL}$, we select the scale matrix $\mathbf{B}_0$ in order to concentrate a lot of mass near the cluster center $\mathbf{b}_{0k}$, pulling $\boldsymbol{\mu}_{kl}$ toward $\mathbf{b}_{0k}$. Matching the a priori expectation of the second term in the variance decomposition (10), given by

$$E\left(\sum_{k=1}^{K}\eta_k\left(\sum_{l=1}^{L}w_{kl}\boldsymbol{\mu}_{kl}\boldsymbol{\mu}_{kl}'-\boldsymbol{\mu}_k\boldsymbol{\mu}_k'\right)\right)=\sum_{k=1}^{K}E(\eta_k)\sum_{l=1}^{L}E(w_{kl})E(\boldsymbol{\mu}_{kl}\boldsymbol{\mu}_{kl}'-\boldsymbol{\mu}_k\boldsymbol{\mu}_k')=\mathbf{B}_0,$$

to the desired proportion of explained heterogeneity and, using once more only the main diagonal

of $\mathbf{S}_y$ we obtain $\mathbf{B}_0 = \phi_W(1-\phi_B)\text{diag}(\mathbf{S}_y)$, which incorporates our idea that only a small proportion ϕ_W of the within-cluster variability should be explained by the variability of the subcomponent means.

After having chosen ϕ_B and ϕ_W, basically the cluster structure and shape is a priori determined. However, in order to allow for more flexibility in capturing the unknown cluster shapes in the sense that within each cluster the amount of shrinkage of the subcomponent means $\boldsymbol{\mu}_{kl}$ toward the cluster center $\mathbf{b}_{0k}$ need not to be the same for all dimensions, for each cluster k and each dimension j additionally a random adaptation factor λ_{kj} is introduced in (8) which adjusts $\mathbf{B}_0$. The gamma prior for λ_{kj} in (7) implies that the prior expectation of the covariance matrix of $\boldsymbol{\mu}_{kl}$ equals $\mathbf{B}_0$. However, λ_{kj} acts as a local adjustment factor for cluster k which allows to shrink (or inflate) the variance of subcomponent means μ_{klj} in dimension j in order to adapt to a more (or less) dense cluster distribution as specified by $\mathbf{B}_0$. In order to allow only for small adjustments of the specified $\mathbf{B}_0$, we choose $\nu = 10$, in this way almost 90% of the a priori values of λ_{kj} are between 0.5 and 1.5. This hierarchical prior specification for $\boldsymbol{\mu}_{kl}$ corresponds to the normal gamma prior (Griffin and Brown, 2010) which has been applied by Frühwirth-Schnatter (2011) and Malsiner-Walli et al. (2016) in the context of finite mixture models for variable selection.

2.4 Relation to BNP mixtures

Our approach bears resemblance to various approaches in BNP modeling. First of all, the concept of sparse finite mixtures as used in Malsiner-Walli et al. (2016) is related to Dirichlet process (DP) mixtures (Müller and Mitra, 2013) where the discrete mixing distribution in the finite mixture (1) is substituted by a random distribution $G \sim DP(\alpha, H)$, drawn from a DP prior with precision parameter α and base measure H. As a draw G from a DP is almost surely discrete, the corresponding model has a representation as an infinite mixture:

$$p(\mathbf{y}) = \sum_{k=1}^{\infty} \eta_k p_k(\mathbf{y}|\boldsymbol{\theta}_k), \tag{13}$$

with i.i.d. atoms $\boldsymbol{\theta}_k \overset{iid}{\sim} H$ drawn from the base measure H and weights $\eta_k = v_k \prod_{j=1}^{k-1}(1-v_j)$ obeying the stick breaking representation with $v_k \overset{iid}{\sim} \mathcal{B}(1, \alpha)$ (Sethuraman, 1994).

If the hyperparameter in the weight distribution $\boldsymbol{\eta}$ of a sparse finite mixture is chosen as $e_0 = \alpha/K$, i.e. $\boldsymbol{\eta} \sim Dir_K(\alpha/K)$, and the component parameters $\boldsymbol{\theta}_k \overset{iid}{\sim} H$ are i.i.d. draws from H, then as K increases, the sparse finite mixture in Equation (1) converges to a DP mixture with mixing distribution $G \sim DP(\alpha, H)$, see Green and Richardson (2001). For example, the sparse finite Gaussian mixture introduced in Malsiner-Walli et al. (2016) converges to a Dirichlet process Gaussian mixture as K increases, with $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ being i.i.d. draws from the appropriate base measure H.

The more general sparse finite mixture of mixtures model introduced in this paper also converges to a Dirichlet process mixture where the atoms are finite mixtures indexed by the parameter $\boldsymbol{\theta}_k$ defined in (3). The parameters $\boldsymbol{\theta}_k$ are i.i.d. draws from the base measure (6), with strong dependence among the means $\boldsymbol{\mu}_{k1}, \ldots, \boldsymbol{\mu}_{kL}$ and covariances $\boldsymbol{\Sigma}_{k1}, \ldots, \boldsymbol{\Sigma}_{kL}$ within each cluster k. This dependence is achieved through the two-layer hierarchical prior described in (7) and (8) and is essential to create well-connected clusters from the subcomponents, as outlined in Section 2.3.

Also in the BNP framework models have been introduced that create dependence, either in the atoms and/or in the weights attached to the atoms. For instance, the nested DP process of Rodriguez et al. (2008) allows to cluster distributions across N units. Within each unit i, $i = 1, \ldots, N$, repeated (univariate) measurements $y_{it}, t = 1, \ldots, N_i$ arise as independent realizations of a DP Gaussian mixture with random mixing distribution G_i. The G_is are i.i.d. draws from a DP, in which the base measure is itself a Dirichlet process $DP(\beta, H)$, i.e. $G_i \overset{iid}{\sim} DP(\alpha, DP(\beta, H))$. Hence, two distributions G_i and G_j either share the same weights and atoms sampled from H, or the weights and atoms are entirely different. If only a single observation y_i is available in each unit, i.e. $N_i = 1$, then the nested DP is related to our model. In particular, it has a two-layer representation as in (1) and (2), however with both K and L being infinite. The nested DP can, in principal, be extended to multivariate observations $\mathbf{y}_i$. In this case, $p(\mathbf{y}_i)$ takes the same form as in (13), with the same stick breaking representation for the cluster weights $\eta_1, \eta_2, \ldots$. On the lower level, each cluster distribution $p_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is a DP Gaussian mixture:

$$p_k(\mathbf{y}_i|\boldsymbol{\theta}_k) = \sum_{l=1}^{\infty} w_{kl} f_{\mathcal{N}}(\mathbf{y}_i|\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}), \tag{14}$$

where the component weights w_{kl} are derived from the stick breaking representation $w_{kl} = u_{kl} \prod_{j=1}^{l-1}(1-$

$u_{kj}), l = 1, 2, \ldots$ where $u_{kl} \overset{iid}{\sim} \mathcal{B}(1, \beta)$. For the nested DP, dependence is introduced only on the level of the weights and sticks, as the component parameters $\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl} \overset{iid}{\sim} H$ are i.i.d. draws from the base measure H. This lack of prior dependence among the atoms $(\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl})$ is likely to be an obstacle in a clustering context.

The BNP approach most closely related to our model is the infinite mixture of infinite Gaussian mixtures (I^{2}GMM) model of Yerebakan et al. (2014) which also deals with clustering multivariate observations from non-Gaussian component densities.[1] The I^{2}GMM model has a two-layer hierarchical representation like the nested DP. On the top level, i.i.d. cluster specific locations $\mathbf{b}_{0k}$ and covariances $\boldsymbol{\Sigma}_k$ are drawn from a random distribution $G \sim DP(\alpha, H)$ arising from a DP prior with base measure H being equal to the conjugate normal-inverse-Wishart distribution. A cluster specific DP is introduced on the lower level as for the nested DP; however, the I^{2}GMM model is more flexible, as prior dependence is also introduced among the atoms belonging to the same cluster. More precisely, $\mathbf{y}_i \sim \mathcal{N}_r(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_k)$, with $\boldsymbol{\mu}_i \overset{iid}{\sim} G_k$, where $G_k \sim DP(\beta, H_k)$ is a draw from a DP with cluster specific base measure $H_k = \mathcal{N}_r(\mathbf{b}_{0k}, \boldsymbol{\Sigma}_k/\kappa_1)$.

It is easy to show that the I^{2}GMM model has an infinite two-layer representation as in (13) and (14), with exactly the same stick breaking representation.[2] However, the I^{2}GMM model has a constrained form on the lower level, with homoscedastic covariances $\boldsymbol{\Sigma}_{kl} \equiv \boldsymbol{\Sigma}_k$, whereas the locations $\boldsymbol{\mu}_{kl}$ scatter around the cluster centers $\mathbf{b}_{0k}$ as in our model:

$$(\mathbf{b}_{0k}, \boldsymbol{\Sigma}_k) \overset{iid}{\sim} H, \qquad \boldsymbol{\mu}_{kl} | \mathbf{b}_{0k}, \boldsymbol{\Sigma}_k \overset{iid}{\sim} H_k. \tag{15}$$

In our sparse mixture of mixtures model, we found it useful to base the density estimator on heteroscedastic covariances $\boldsymbol{\Sigma}_{kl}$, to better accommodate the non-Gaussianity of the cluster densities with a fairly small number L of subcomponents. It should be noted that our semi-parametric density estimator is allowed to display non-convex shapes, as illustrated in Figure C.2 in the Appendix. Nevertheless, we could have considered a mixture in (2) where $\boldsymbol{\Sigma}_{kl} \equiv \boldsymbol{\Sigma}_k$, with the same base measure for the atoms $(\boldsymbol{\mu}_{k1}, \ldots, \boldsymbol{\mu}_{kL}, \boldsymbol{\Sigma}_k)$ as in (15). In this case, the relationship between our sparse finite mixture and the I^{2}GMM model would become even more apparent: by choosing $e_0 = \alpha/K$

[1] We would like to thank a reviewer for pointing us to this paper.

[2] Note that the notation in Yerebakan et al. (2014) is slightly different, with γ and α corresponding to α and β introduced above.

and $d_0 = \beta/L$ and letting K and L go to infinity, our model would converge to the I^{2}GMM model.

3 Clustering and posterior inference

3.1 Clustering and selecting the number of clusters

For posterior inference, two sequences of allocation variables are introduced, namely the cluster assignment indicators $\mathbf{S} = (S_1, \ldots, S_N)$ and the within-cluster allocation variables $\mathbf{I} = (I_1, \ldots, I_N)$. More specifically, $S_i \in \{1, \ldots, K\}$ assigns each observation $\mathbf{y}_i$ to cluster S_i on the upper level of the mixture of mixtures model. On the lower level, $I_i \in \{1, \ldots, L\}$ assigns observation $\mathbf{y}_i$ to subcomponent I_i. Hence, the pair (S_i, I_i) carries all the information needed to assign each observation to a unique component in the expanded mixture (4).

Note that for all observations $\mathbf{y}_i$ and $\mathbf{y}_j$ belonging to the same cluster, the upper level indicators $S_i = S_j$ will be the same, while the lower level indicators $I_i \neq I_j$ might be different, meaning that they belong to different subcomponents within the same cluster. It should be noted that the Dirichlet prior $\mathbf{w}_k \sim Dir_L(d_0)$, with $d_0 > d/2$, on the weight distribution ensures overlapping densities within each cluster, in particular if L is overfitting. Hence the indicators I_i will typically cover all possible values $\{1, \ldots, L\}$ within each cluster.

For clustering, only the upper level indicators $\mathbf{S}$ are explored, integrating implicitly over the uncertainty of assignment to the subcomponents on the lower level. A cluster $C_k = \{i | S_i = k\}$ is thus a subset of the data indices $\{1, \ldots, N\}$, containing all observations with identical upper level indicators. Hence, the indicators $\mathbf{S}$ define a random partition $\mathcal{P} = \{C_1, \ldots, C_{K_0}\}$ of the N data points in the sense of Lau and Green (2007), as $\mathbf{y}_i$ and $\mathbf{y}_j$ belong to the same cluster, if and only if $S_i = S_j$. The partition $\mathcal{P}$ contains $K_0 = |\mathcal{P}|$ clusters, where $|\mathcal{P}|$ is the cardinality of $\mathcal{P}$. Due to the Dirichlet prior $\boldsymbol{\eta} \sim Dir_K(e_0)$, with e_0 close to 0 to obtain a sparse finite mixture, K_0 is a random number being a priori much smaller than K.

For a sparse finite mixture model with K clusters, the prior distribution over all random partitions $\mathcal{P}$ of N observations is derived from the joint (marginal) prior $p(\mathbf{S}) = \int \prod_{i=1}^{N} p(S_i|\boldsymbol{\eta}) d\boldsymbol{\eta}$

which is given, e.g., in Frühwirth-Schnatter (2006, p. 66):

$$p(\mathbf{S}) = \frac{\Gamma(Ke_0)}{\Gamma(N+Ke_0)\Gamma(e_0)^{K_0}} \prod_{k:N_k>0} \Gamma(N_k + e_0), \tag{16}$$

where $N_k = \#\{S_i = k\}$. For a given partition $\mathcal{P}$ with K_0 data clusters, there are $K!/(K-K_0)!$ assignment vectors $\mathbf{S}$ that belong to the equivalence class defined by $\mathcal{P}$. The prior distribution over all random partitions $\mathcal{P}$ is then obtained by summing over all assignment vectors $\mathbf{S}$ that belong to the equivalence class defined by $\mathcal{P}$:

$$p(\mathcal{P}|K_0) = \frac{K!}{(K-K_0)!} \frac{\Gamma(Ke_0)}{\Gamma(N+Ke_0)\Gamma(e_0)^{K_0}} \prod_{k:N_k>0} \Gamma(N_k + e_0), \tag{17}$$

which takes the form of a product partition model and therefore is invariant to permuting the cluster labels. Hence, it is possible to derive the prior predictive distribution $p(S_i|\mathbf{S}_{-i})$, where $\mathbf{S}_{-i}$ denote all indicators, excluding S_i. Let K_0^{-i} be the number of non-empty clusters implied by $\mathbf{S}_{-i}$ and let N_k^{-i} be the corresponding cluster sizes. From (16), we obtain the following probability that S_i is assigned to an existing cluster k:

$$\Pr\{S_i = k|\mathbf{S}_{-i}, N_k^{-i} > 0\} = \frac{N_k^{-i} + e_0}{N - 1 + e_0 K}. \tag{18}$$

The prior probability that S_i creates a new cluster with $S_i \in I = \{k|N_k^{-i} = 0\}$ is equal to

$$\Pr\{S_i \in I|\mathbf{S}_{-i}\} = (K - K_0^{-i})\Pr\{S_i = k^*|\mathbf{S}_{-i}, k^* \in I\} = \frac{e_0(K - K_0^{-i})}{N - 1 + e_0 K}. \tag{19}$$

It is illuminating to investigate the prior probability to create new clusters in detail. First of all, for e_0 independent of K, this probability not only depends on e_0, but also increases with K. Hence a sparse finite mixture model based on the prior $\boldsymbol{\eta} \sim \mathcal{D}_K(e_0)$ can be regarded as a two-parameter model, where both e_0 and K influence the a priori expected number of data clusters K_0 which is determined for a DP mixture solely by α. A BNP two-parameter mixture is obtained from the Pitman-Yor process (PYP) prior $PY(\beta, \alpha)$ with $\beta \in [0, 1), \alpha > -\beta$ (Pitman and Yor, 1997), with stickbreaking representation $v_k \overset{iid}{\sim} \mathcal{B}(1-\beta, \alpha + k\beta)$. The DP prior results as that special case where $\beta = 0$.

Second, the prior probability (19) to create new clusters in a sparse finite mixture model decreases, as the number K_0^{-i} of non-empty clusters increases. This is in sharp contrast to DP mixtures

where this probability is constant and PYP mixtures where this probability increases, see e.g., Fall and Barat (2014).

Finally, what distinguishes a sparse finite mixture model, both from a DP as well as a PYP mixture, is the a priori expected number of data clusters K_0, as the number N of observations increases. For K and e_0 independent of N, the probability to create new clusters decreases, as N increases, and converges to 0, as N goes to infinity. Therefore, K_0 is asymptotically independent of N for sparse finite mixtures, whereas for the DP process $K_0 \sim \alpha \log(N)$ (Korwar and Hollander, 1973) and $K_0 \sim N^{\beta}$ obeys a power law for PYP mixtures (Fall and Barat, 2014). This leads to quite different clustering behavior for these three types of mixtures.

A well-known limitation of DP priors is that a priori the cluster sizes are expected to be geometrically ordered, with one big cluster, geometrically smaller clusters, and many singleton clusters (Müller and Mitra, 2013). PYP mixtures are known to be more useful than the DP mixture for data with many significant, but small clusters. A common criticism concerning finite mixtures is that the number of clusters needs to be known a priori. Since this is not the case for sparse finite mixtures, they are useful in the context of clustering, in particular in cases where the data arise from a moderate number of clusters, that does not increase as the number of data points N increases.

3.2 MCMC estimation and posterior inference

Bayesian estimation of the sparse hierarchical mixture of mixtures model is performed using MCMC methods based on data augmentation and Gibbs sampling. We only need standard Gibbs sampling steps, see the detailed MCMC sampling scheme in Appendix A.

In order to perform inference based on the MCMC draws, i.e. to cluster the data, to estimate the number of clusters, to solve the label switching problem on the higher level and to estimate cluster-specific parameters, several existing procedures can be easily adapted and applied to post-process the posterior draws of a mixture of mixtures model, e.g., those which are, for instance, implemented in the R packages **PReMiuM** (Liverani et al., 2015) and **label.switching** (Papastamoulis, 2015).

For instance, the approach in **PReMiuM** is based on the posterior probabilities of co-clustering, expressed through the similarity matrix $\Pr\{S_i = S_j|\mathbf{y}\}$ which can be estimated from the M poste-

rior draws $\mathbf{S}^{(m)}, m = 1, \ldots, M$, see Appendix B for details. The methods implemented in **label.switching** aim at resolving the label switching problem when fitting a finite mixture model using Bayesian estimation. Note that in the case of the mixture of mixtures model label switching occurs on two levels. 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. On the subcomponent level, label switching happens due to the invariance of Equation (2) with respect to reordering of the subcomponents. As we are only interested in estimating the entire cluster distributions, it is not necessary to identify the single subcomponents. Therefore, the label switching problem can be ignored on this level.

In this paper, the post-processing approach employed first performs a model selection step. The posterior draws of the indicators $\mathbf{S}^{(m)}, m = 1, \ldots, M$ are used to infer the number of non-empty clusters $K_0^{(m)}$ on the upper level of the mixture of mixtures model and the number of data clusters is then estimated as the mode. Conditional on the selected model, an identified model is obtained based on the point process representation of the estimated mixture. This method was introduced in Frühwirth-Schnatter (2006, p. 96) and successfully applied to model-based clustering in various applied research, see e.g. Frühwirth-Schnatter (2011) for some review. This procedure has been adapted to sparse finite mixtures in Frühwirth-Schnatter (2011) and Malsiner-Walli et al. (2016) and is easily extended to deal with sparse mixture of mixtures models, see Appendix B for more details. We will use this post-processing approach in our simulation studies and the applications in Section 4 and Appendices C, D and F to determine a partition of the data based on the maximum a posteriori (MAP) estimates of the relabeled cluster assignments.

4 Simulation studies and applications

The performance of the proposed strategy for selecting the unknown number of clusters and identifying the cluster distributions is illustrated in two simulation studies. 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 maximum number of clus-

ters K and subcomponents L on the clustering results is studied. In the second simulation study the sensitivity of the a priori defined proportions ϕ_B and ϕ_W on the clustering result is investigated. For a detailed description of the simulation design and results see Appendix C. Overall, the results indicated that our approach performed well and yielded promising results.

To further evaluate our approach, we fit the sparse hierarchical mixture of mixtures model on benchmark data sets and real data. First, we consider five data sets which were previously used to benchmark algorithms in cluster analysis. For these data sets we additionally apply the "merging strategy" proposed by Baudry et al. (2010) in order to compare the results to those of our approach. For these benchmark data sets class labels are available and we assess the performance by comparing how well our approach is able to predict the class labels using the cluster assignments, measured by the misclassification rate as well as the adjusted Rand index.

To assess how the algorithm scales to larger data sets we investigate the application to two flow cytometry data sets. The three-dimensional DLBCL data set (Lee and McLachlan, 2013) consists of around 8000 observations and comes with manual class labels which can be used as benchmark. The GvHD data set (Brinkman et al., 2007) consists of 12441 observations, but no class labels are available. We compare the clusters detected for this data set qualitatively to solutions previously reported in the literature.

The detailed description of all investigated data sets as well as of the derivation of the performance measures are given in Appendix D. For the benchmark data sets, the number of estimated clusters $\hat{K}_0$, the adjusted Rand index (*adj*), and misclassification rate (*er*) are reported in Table 1 for all estimated models. In the first columns of Table 1, the name of the data set, the number of observations N, the number of variables r and the number of true 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 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 based on the BIC is reported in the column `Mclust`. Then the selected components are combined hierarchically to clusters by calling function `clustCombi` from the same package (column `clustCombi`). 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 with $K = 10$ (Malsiner-Walli et al., 2016) is fitted to the data sets (column *SparseMix*). The results of fitting a sparse hierarchical mixture of mixtures model with $K = 10$ are given in column *SparseMixMix*, where $L = 5$ is compared to our default choice of $L = 4$ to investigate robustness with respect to the choice of L. For each estimation, MCMC sampling is run for 4000 iterations after a burn-in of 4000 iterations.

As can be seen in Table 1, for all data sets the sparse hierarchical mixture of mixtures model is able to capture the data clusters quite well both in terms of the estimated number of clusters and the clustering quality measured by the misclassification rate as well as the adjusted Rand index. In general, our approach is not only outperforming the standard model-based clustering model using mixtures of Gaussians regarding both measures, but also the approach proposed by Baudry et al. (2010). In addition, it can be noted that for all data sets the estimation results remain quite stable, if the number of subcomponents L is increased to 5, see the last column in Table 1. The results for the Yeast data set are of particular interest as they indicate that `clustCombi` completely fails. Although the misclassification rate of 25% implies that only a quarter of the observations is assigned to "wrong" clusters, inspection of the clustering obtained reveals that almost all observations are lumped together in a single, very large cluster, whereas the few remaining observations are split into five very small clusters. This bad clustering quality is better reflected by the adjusted Rand index which takes a negative value ($adj = -0.02$), i.e. is "worse than would be expected by guessing" (Franczak et al., 2012). For the flower data set, more results are given in Appendix D where the obtained clustering and cluster distributions are illustrated.

In order to investigate the performance of our approach on larger data sets with a slightly different cluster structure, we fit the sparse hierarchical mixture of mixtures model to two flow cytometry data sets. These applications also allow us to indicate how the prior settings need to be adapted if a different cluster structure is assumed to be present in the data. As generally known, flow cytometry data exhibit non-Gaussian characteristics such as skewness, multimodality and a large number of outliers, as can be seen in the scatter plot of two variables of the GvHD data set in Figure 3. Thus, we specified a sparse hierarchical mixture of mixtures model with $K = 30$

clusters and increased the number of subcomponents forming a cluster to $L = 15$ in order to handle more complex shapes of the cluster distributions given the large amount of data. Since the flow cytometry data clusters have a lot of outliers similar to the clusters generated by shifted asymmetric Laplace (*SAL*) distributions (see Appendix F), we substitute the hyperprior $\mathbf{C}_{0k} \sim \mathcal{W}_r(g_0, \mathbf{G}_0)$ by the fixed value $\mathbf{C}_{0k} = g_0\mathbf{G}_0^{-1}$ and set $\lambda_{kj} \equiv 1$, $j = 1, \ldots, r$ to prevent that within a cluster the subcomponent covariance matrices are overly shrunken and become too similar. In this way, subcomponent covariance matrices are allowed to vary considerably within a cluster and capture both a dense cluster region around the cluster center and scattered regions at the boundary of the cluster.

We fit this sparse hierarchical mixture of mixtures model to the DLBCL data after removing 251 dead cells. For most MCMC runs after a few hundred iterations all but four clusters become empty during MCMC sampling. The estimated four cluster solution coincides almost exactly with the cluster solution obtained with manual gating; the adjusted Rand index is 0.95 and the error rate equals 0.03. This error rate outperforms the error rate of 0.056 reported by Lee and McLachlan (2013). In Figure 2 the estimated four cluster solution is visualized.

When fitting a sparse hierarchical mixture of mixtures model to the GvHD data, the classifications resulting from different runs of the MCMC algorithm seemed to be rather stable. The obtained solutions differ mainly in the size of the two large clusters with low expressions. These, however, are supposed to not contain any information regarding the development of the disease. On the right hand side of Figure 3, the results of one specific run are shown in a heatmap. In this run, we found eight clusters which are similar to those reported by Frühwirth-Schnatter and Pyne (2010) when fitting a skew-*t* mixture model to these data. In the heatmap each row represents the location of a six-dimensional cluster, and each column represents a particular marker (variable). The red, white and blue colors denote high, medium and low expressions.

As in Frühwirth-Schnatter and Pyne (2010), we identified two larger clusters (43% and 20.4%, first two rows in the heatmap) with rather low expressions in the last four variables. We also identified a smaller cluster (3.8%, forth row from the bottom) representing live cells (high values in the first two variables) with a unique signature in the other four variables (high values in all four variables). Also two other small clusters can be identified (second and third row from the bottom)

which have a signature very similar to the clusters found by Frühwirth-Schnatter and Pyne (2010), and thus our results confirm their findings.

5 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 structured informative priors on all model parameters. This imposes a rigid hierarchical structure on the normal subcomponents and allows for simultaneous estimation of the number 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 taking the data clusters into account which are 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 by the subcomponent densities is crucial for the clustering result. Enforcing overlapping subcomponent densities is essential in order to avoid that a single subcomponent becomes too narrow thus leading to a small a posteriori cluster probability for observations from this subcomponent. Also, enforcing that observations are assigned to *all* subcomponents during MCMC sampling is important as the estimation of empty subcomponents would bias the resulting cluster 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 inverse Wishart prior for the subcomponent covariance matrices. We select both crucial hyperparameters based on 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, 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 Section 2 are not able to capture the clusters and need to be tuned to capture also this special kind of data clusters, see the example given in Appendix F.

Although our approach to estimate the number of clusters worked well for many data sets, we encountered mixing problems with the blocked conditional Gibbs sampler outlined in Appendix A, in particular in high dimensional spaces with large data sets. To alleviate this problem, a collapsed sampler similar to Fall and Barat (2014) could be derived for finite mixtures. However, we leave this for future research.

SUPPLEMENTARY MATERIAL

Appendix containing (A) the MCMC scheme to estimate a mixture of mixtures model, (B) a detailed description of the post-processing strategy based on the point process representation, (C) the simulation studies described in Section 4, (D) a description of the data sets studied in Section 4, (E) issues with the merging approach, and (F) estimation of data clusters generated by a *SAL*-distribution (Franczak et al., 2012). (Appendix.pdf)

R code implementing the sparse hierarchical mixture of mixtures model (Code.zip).

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.

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). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics 2*(19), 332–353.

Brinkman, R. R., M. Gasparetto, S.-J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation 13*(6), 691–700.

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.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics*, pp. 201–218. Cambridge University Press.

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.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Fall, M. D. and E. Barat (2014). Gibbs sampling methods for Pitman-Yor mixture models. <hal-00740770v2>.

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. (2011). Panel data analysis – a survey on model-based clustering of time series. *Advances in Data Analysis and Classification 5*251–280.

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.

Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics 28*, 355–375.

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

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

Korwar, R. M. and M. Hollander (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability 1*(4), 705–711.

Lau, J. W. and P. Green (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics 16*, 526–558.

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.

Lee, S. X. and G. J. McLachlan (2013). EMMIX-uskew: an R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. *Journal of Statistical Software 55*(12).

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

Liverani, S., D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software 64*(7), 1–30.

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

Medvedovic, M., K. Y. Yeung, and R. E. Bumgarner (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics 20*(8), 1222–1232.

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

Müller, P. and R. Mitra (2013). Bayesian nonparametric inference - Why and how. *Bayesian Analysis 8*(2), 269–302.

Papastamoulis, P. (2015). *label.switching: Relabelling MCMC Outputs of Mixture Models*. R package version 1.4.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability 25*, 855–900.

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

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

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.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

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

Yerebakan, H. Z., B. Rajwa, and M. Dundar (2014). The infinite mixture of infinite Gaussian mixtures. In *Advances in Neural Information Processing Systems*, pp. 28–36.

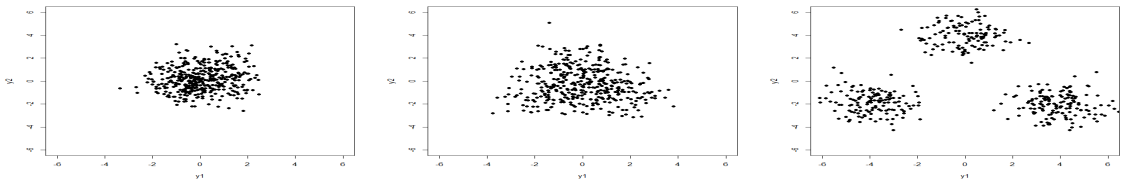

Figure 1: Variance decomposition of a mixture distribution. Scatter plots of samples from a standard normal mixture distribution with three components and equal weights, with a varying amount of heterogeneity ϕ_B explained by the variation of the component means, $\phi_B = 0.1$, $\phi_B = 0.5$ and $\phi_B = 0.9$ (from left to right).

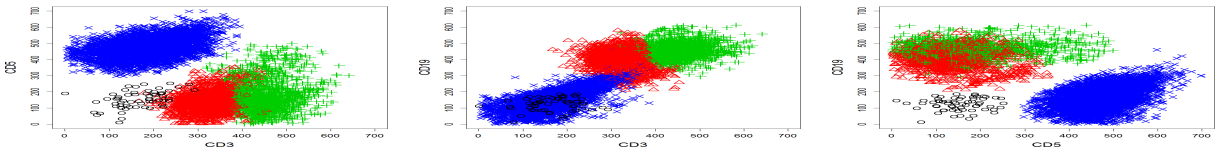

Figure 2: Flow cytometry data set DLBCL. Scatterplot of the clustering results.

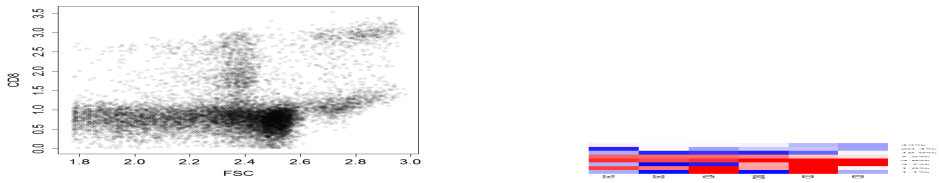

Figure 3: Flow cytometry data set GvHD. Scatter plot of two variables ("FSC", "CD8") (left-hand side), and heatmap of the clustering results by fitting a sparse hierarchical mixture of mixtures model (right-hand side). In the heatmap, each row represents the location of a six-dimensional cluster, and each column represents a particular marker. The red, white and blue colors denote high, medium and low expression, respectively.

Table 1: Results for the estimated number of data clusters $\hat{K}_0$ for various benchmark data sets, using the functions `Mclust` to fit a standard mixture model with $K = 10$ and `clustCombi` to estimate a mixture with combined components (column *Mclust*), using a sparse finite mixture model with $K = 10$ (column *SparseMix*), and estimating a sparse hierarchical mixture of mixtures model with $K = 10$, $\phi_B = 0.5$ and $\phi_W = 0.1$, and $L = 4, 5$ (column *SparseMixMix*). Priors and hyperparameter specifications are selected as described in Section 2. In parentheses, the adjusted Rand index ("1" corresponds to perfect classification) and the proportion of misclassified observations ("0" corresponds to perfect classification) are reported.

				Mclust $K = 10$		*SparseMix* $K = 10$	*SparseMixMix* $K = 10$	
Data set	N	r	K^{true}	`Mclust`	`clustCombi`	$L = 1$	$L = 4$	$L = 5$
Yeast	626	3	2	8 *(.50, .20)*	6 *(-.02, 0.25)*	6 *(.48, .23)*	**2** *(.68, .08)*	2 *(.71, .07)*
Flea beetles	74	6	3	5 *(.77, .18)*	4 *(.97, .03)*	3 *(1.00, .00)*	**3** *(1.00, .00)*	3 *(1, .00)*
AIS	202	3	2	3 *(.73, .13)*	2 *(.66, .09)*	3 *(.76, .11)*	**2** *(.81, .05)*	2 *(.76, .06)*
Wisconsin	569	3	2	4 *(.55, .30)*	4 *(.55, .30)*	4 *(.62, .21)*	**2** *(.82, .05)*	2 *(.82, .05)*
Flower	400	2	4	6 *(.52, .35)*	4 *(.99, .01)*	5 *(.67, .20)*	**4** *(.97, .01)*	4 *(.97, .02)*