Chapter 17
Bayesian Mixture Models with Weight-Dependent Component Priors

Elaheh Oftadeh and Jian Zhang

Abstract In the conventional Bayesian mixture models, independent priors are often assigned to weights and component parameters. This may cause bias in estimation of missing group memberships due to the domination of these priors for some components when there is a big variation across component weights. To tackle this issue, we propose weight-dependent priors for component parameters. To implement the proposal, we develop a simple coordinate-wise updating algorithm for finding empirical Bayesian estimator of allocation or labelling vector of observations. We conduct a simulation study to show that the new method can outperform the existing approaches in terms of adjusted Rand index. The proposed method is further demonstrated by a real data analysis.

17.1 Introduction

Finite mixture models are a popular tool for modelling unobserved heterogeneity in many applications including biology, medicine, economics and engineering among many others (e.g., [3, 4]). Suppose that we sample $\mathbf{y} = (y_1, \cdots, y_N)$ from a population with K groups, described by mixture distribution

$$p(y_i|\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k p(y_i|\boldsymbol{\theta}_k),$$

with unknown component parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ and unknown weights $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)$. Given the dataset $\mathbf{y} = (y_1, \cdots, y_N)$, we want to infer parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$

E. Oftadeh (✉) · J. Zhang
School of Mathematics, Statistics and Actuarial Science, University of Kent,
Canterbury CT2 7FS, UK
e-mail: eo217@kentforlife.net

J. Zhang
e-mail: jz79@kent.ac.uk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_17

as well as unobserved component origins of these observations, labelled by allocation (or labelling) vector $\boldsymbol{S} = (s_1, ..., s_N)$. In Bayesian inference, we often adopt the following hierarchical setting:

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{S}) = \prod_{k=1}^{K} \prod_{S_i=k} p(y_i|\boldsymbol{\theta}_k),\ p(\boldsymbol{S}|\boldsymbol{\eta}) = \prod_{k=1}^{K} \eta_k^{\sum_{i=1}^{N} I(S_i=k)},$$

$$p(\boldsymbol{\eta}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \prod_{k=1}^{K} \eta_k^{e_0-1}, e_0 > 0, \quad (\boldsymbol{\theta}, \boldsymbol{\eta})) \sim p(\boldsymbol{\theta})p(\boldsymbol{\eta}),$$

where $I(\cdot)$ is an indicator function, $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{S})p(\boldsymbol{S}|\boldsymbol{\eta})$ is the complete likelihood and $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are of independent priors. The above setting is useful for fitting finite mixture models to data, because they enable the uncertainty in the model parameters to be directly quantified by the posterior distribution. However, it is difficult to make an objective prior setting for the component parameters (such as the component means and variances, in univariate Gaussian mixtures), when there is no subjective information available on which a prior could be based. For example, when some component weights are small, only a small proportion of observations are expected to obtain from these components. In such a situation, the priors can easily dominate the data for these components. Such a prior domination in the inference can cause a bias. To reduce the bias, we need to set these priors compatible to the available information from the data. Ideally, the priors are set to be close to non-informative. On the other hand, standard non-informative priors such as the Jeffreys prior generally cannot be used here, because placing independent improper priors on the component parameters will cause the posterior to be improper as well [9]. This motivates us to explore the advantage of the weight-dependent component priors. In this paper, we propose a new weight-dependent prior specification for finite mixture models in the form $(\boldsymbol{\theta}, \boldsymbol{\eta}) \sim p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})$. We develop a coordinate-wise updating algorithm for conducting Bayesian inference: First, given the data, derive a marginal posterior distribution for allocation vector $\boldsymbol{S}$ and optimize it over the labelling space to obtain an optimal allocation estimate $\widehat{\boldsymbol{S}}$. Then, conditional on $\widehat{\boldsymbol{S}}$, calculate the posterior distribution of parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$. We conduct a simulation study to show that the new approach can outperform the existing methods in terms of adjusted Rand index. The proposed method is further demonstrated by a real data analysis.

The remaining of the paper is organized as follows. The details of the proposed methodology and algorithm are provided in Sect. 17.2. A comparison to the existing methods are made through a simulation study in Sect. 17.3. A real data application is presented in Sect. 17.4. The conclusion is made in Sect. 17.5.

17.2 Methodology

In Bayesian inference, the main task is to calculate the posterior distribution of unknown parameters by combining the prior information about the parameters of interest with the data. Let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\eta})$. By augmenting the missing allocation vector $\boldsymbol{S}$ into the finite mixture model and letting $p(\boldsymbol{S}|\boldsymbol{\vartheta}) = \prod_{k=1}^{K} \eta_k^{N_k(\boldsymbol{S})}$ with $N_k(\mathbf{S})$ being the size of group k, we can link the incomplete likelihood to the complete likelihood as follows:

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

Denote the complete data by $(\mathbf{y}, \mathbf{S})$ and the complete-data likelihood by

$$p(\mathbf{y}, \boldsymbol{S}|\boldsymbol{\vartheta}) = p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\vartheta}, S_i)p(S_i|\boldsymbol{\vartheta}).$$

Note that $p(y_i|S_i = k, \boldsymbol{\vartheta}) = p(y_i|\boldsymbol{\theta}_k)$ and $P(S_i = k|\boldsymbol{\vartheta}) = \eta_k$. Therefore, the complete-data likelihood function can be rewritten as

$$p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}(p(y_i|\boldsymbol{\theta}_k)\eta_k)^{I(S_i=k)} = \left(\prod_{k=1}^{K} \eta_k^{N_k(\mathbf{S})}\right)\prod_{k=1}^{K}\left(\prod_{i:S_i=k} p(y_i|\boldsymbol{\theta}_k)\right). \tag{17.1}$$

We assign a Dirichlet prior to the weights with the concentration parameter e_0 in the form

$$p(\boldsymbol{\eta}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\prod_{k=1}^{K}\eta_k^{e_0-1}.$$

By integrating out $\boldsymbol{\eta}$ in $p(\boldsymbol{S}|\boldsymbol{\eta})p(\boldsymbol{\eta}|e_0)$, we obtain the marginal prior on $\boldsymbol{S}$ and posterior of $\boldsymbol{\eta}$ as follows

$$p(\mathbf{S}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\int\prod_{k=1}^{K}\eta_k^{N_k(S)+e_0-1}d\eta_k, \quad p(\boldsymbol{\eta}|\mathbf{S}) = \frac{p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\mathbf{S})}.$$

Once we have an estimate of $\boldsymbol{S}$, using the above formulas we are able to calculate the posterior of $\boldsymbol{\eta}$. So, in the following, we focus on Bayesian clustering, i.e., Bayesian estimation of allocation vector $\boldsymbol{S}$.

One of the pioneering works in Bayesian clustering was done by [1], where the problem was formulated in a Bayesian decision theoretic framework with a loss function $R(\mathbf{S}, \widehat{\mathbf{S}})$. This loss function measures the difference between the estimate $\widehat{\mathbf{S}}$ and the true grouping $\boldsymbol{S}$. Here, we take an empirical Bayesian method by optimizing the marginal posterior of allocation vector of $\boldsymbol{S}$, $p(\boldsymbol{S}|\mathbf{y})$. In the simulation study, we evaluate the accuracy of the clustering by calculating the similarity between the estimated and the true labelling by the so-called adjusted Rand index [5, 7]. We consider two different sets of hierarchical priors and derive the corresponding posteriors. In the Bayesian inference for Gaussian mixtures, it is common to choose the component parameter priors to be independent of weights. We derive the posteriors for Bayesian mixture models with independent priors in Sect. 17.2.1.1 and for the models with dependent priors in Sect. 17.2.1.2 below. Although from now on we focus on univariate normal mixtures, the method can be extended to other mixtures such as multivariate normal or non-normal mixtures. For simplicity, we assume that K is known. Otherwise, we can take a Poisson distribution as a prior for K.

17.2.1 Mixture of Univariate Normals

Suppose that $y_i \sim N(\mu_k, \sigma_k^2), i = 1, \cdots, N,$ with $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2), k = 1, 2, ..., K$. For the univariate normal mixtures, we first derive the posterior distribution for mean μ_k and variance $\sigma_k^2, k = 1, ..., K$, given the complete data ($\mathbf{y}$, $\boldsymbol{S}$). We then work out the formulas for calculating and optimizing $p(\boldsymbol{S}|\mathbf{y})$.

17.2.1.1 Weight-Independent Component Priors

We start with a review of the conventional hierarchical model with weight-independent priors on (μ_k, σ_k^2) in [2, 3]:

$$y_i \sim N(\mu_k, \sigma_k^2), \quad \mu_k \sim N(\mu_{k0}, \sigma_{k0}^2), \quad \sigma_k^2 \sim IG(a_0, b_0),$$

where $IG(a_0, b_0)$ is an inverse Gamma density with hyperparameters (a_0, b_0).

The posterior probability of μ_k given the complete data $(\mathbf{S}, \mathbf{y})$ and σ_k^2 can be written as

$$\begin{aligned} p(\mu_k|\mathbf{y}, \mathbf{S}, \sigma_k^2) &\propto p(\mathbf{y}|\mu_k, \sigma_k^2, \mathbf{S})p(\mu_k) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{N_k(\mathbf{S})}{\sigma_k^2} + \frac{1}{\sigma_{k0}^2}\right)\left(\mu_k - \frac{\sum y_i}{\sigma_k^2} + \frac{\mu_{k0}}{\sigma_{k0}^2}\right)^2\right\}. \end{aligned}$$

Thus the posterior distribution of μ_k is the following normal distribution

$$\begin{aligned} p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2) &\sim \mathcal{N}(b_k(\boldsymbol{S}), B_k(\boldsymbol{S})), \quad B_k(\boldsymbol{S})^{-1} = \sigma_{k0}^{-2} + \sigma_k^{-2} N_k(\boldsymbol{S}) \\ b_k(\boldsymbol{S}) &= B_k(\boldsymbol{S})\left(\sigma_k^{-2} N_k(\boldsymbol{S})\bar{y}_k(\boldsymbol{S}) + \sigma_{k0}^{-2}\mu_{k0}\right), \end{aligned}$$

where the sample mean and variance in the kth group are denoted by

$$\bar{y}_k(\boldsymbol{S}) = \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} y_i, \quad S^2_{y,k} = \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} (y_i - \bar{y}_k(\boldsymbol{S}))^2.$$

Similarly, for σ_k^2 we have

$$\sigma_k^2 | \mathbf{y}, \boldsymbol{S}, \mu_k \sim \mathscr{G}^{-1}(c_k(\boldsymbol{S}), C_k(\boldsymbol{S})), \quad c_k(\boldsymbol{S}) = a_0 + \frac{1}{2} N_k(\boldsymbol{S}),$$
$$C_k(\boldsymbol{S}) = b_0 + \frac{1}{2} \sum_{i:S_i=k} (y_i - \mu_k)^2.$$

If we are able to calculate the maximum marginal posterior estimator of the allocation vector, $\widehat{\boldsymbol{S}} = \text{argmax}_{\boldsymbol{S}} p(\boldsymbol{S}|\mathbf{y})$, then we can directly calculate posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\eta})$. To derive the marginal posterior distribution of allocations, we integrate out $(\boldsymbol{\theta}, \boldsymbol{\eta})$ from the model, i.e., consider the following integration

$$\begin{aligned} p(\mathbf{S}|\mathbf{y}) &= \iint p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{S}, \boldsymbol{\theta}) p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) p(\boldsymbol{\theta}) d\theta d\boldsymbol{\eta} \\ &= 2^{Ka_0} N_0^{K/2} \left(\frac{1}{\pi}\right)^{N/2} \left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^K \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \prod_{k=1}^K \sqrt{\frac{1}{N_k(\mathbf{S}) + N_0}} \\ &\times \frac{\prod_{k=1}^K \Gamma(N_k(S) + e_0)}{\Gamma(Ke_0 + N)} \prod_{k=1}^K \mathscr{B}^{-(a_0 + \frac{N_k(\mathbf{S})}{2})} \prod_{k=1}^K \Gamma(a_0 + \frac{N_k(\mathbf{S})}{2}), \end{aligned}$$

where

$$S^2_{y,k} = \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} (y_i - \bar{y}_k(\boldsymbol{S}))^2$$
$$\mathscr{B} = N_k(\mathbf{S}) S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S}) N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}(\mathbf{S}) - \mu_{k0})^2.$$

Taking logarithm, we have

$$\begin{aligned} \log(p(\mathbf{S}|\mathbf{y})) \propto &\sum_{k=1}^K \log \Gamma(N_k(\mathbf{S}) + e_0) \sum_{k=1}^K \log \Gamma(a_0 + \frac{N_k(\mathbf{S})}{2}) \\ &- \sum_{k=1}^K \frac{1}{2} \log(N_k(\mathbf{S}) + N_0) - \sum_{k=1}^K (a_0 + N_k(\mathbf{S})/2) \log \mathscr{B}. \end{aligned} \tag{17.2}$$

17.2.1.2 Weight-Dependent Component Priors

Although we consider the same normal mixture model as in the previous section, we allow certain dependency of the hierarchical priors on component weights as follows:

$$\mu_k|\sigma_k^2, \eta_k \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0\eta_k}), \quad \sigma_k^2 \sim IG(a_0, b_0), k = 1, ..., K. \quad \boldsymbol{\eta} \sim D(e_0, \cdots, e_0),$$

where $D(e_0, \cdots, e_0)$ is a Dirichlet density with concentration parameter e_0. Since N_0 is the total number of prior units we assign to the model, $N_0\eta_k$ is the number of prior units we assign to μ_k. Unlike the weight-independent priors, the prior of μ_k is adaptive to η_k in the sense that the amount of priors will be varying in η_k, in particular, it will be nearly non-informative when η_k tends to zero. The posterior of μ_k given $(\boldsymbol{S}, \mathbf{y})$, σ_k^2 and η_k can then be written as

$$\begin{aligned}
p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2, \eta_k) &\propto p(\mathbf{y}|\mu_k, \sigma_k^2, \mathbf{S})p(\mu_k|\eta_k) \\
&\propto \prod_{k=1}^{K}(\frac{1}{\sigma_k^2})^{-N_k(\mathbf{S})/2} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i - \mu_k)^2\right\} \\
&\times (\frac{1}{\sigma_{k0}^2}\eta_k)^{1/2} \exp\left\{-\frac{1}{2\sigma_{k0}^2\eta_k}(\mu_k - \mu_{k0})^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{N_k(\mathbf{S})}{\sigma_k^2} + \frac{1}{\sigma_{k0}^2\eta_k}\right)\left(\mu_k - \frac{\sum y_i}{\sigma_k^2} + \frac{\mu_{k0}}{\sigma_{k0}^2\eta_k}\right)^2\right\}.
\end{aligned}$$

Thus the posterior distribution of μ_k is the following normal distribution

$$\begin{aligned}
p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2, \eta_k) &\sim \mathcal{N}(b_k(\boldsymbol{S}), B_k(\boldsymbol{S})), \\
B_k(\boldsymbol{S})^{-1} &= \sigma_{k0}^{-2}\eta_k^{-1} + \sigma_k^{-2}N_k(\boldsymbol{S}) \\
b_k(\boldsymbol{S}) &= B_k(\boldsymbol{S})\left(\sigma_k^{-2}N_k(\boldsymbol{S})\bar{y}_k(\boldsymbol{S}) + \eta_k^{-1}\sigma_{k0}^{-2}\mu_{k0}\right),
\end{aligned}$$

where the sample mean and variance in the kth group are denoted by

$$\bar{y}_k(\boldsymbol{S}) = \frac{1}{N_k(\boldsymbol{S})}\sum_{i:S_i=k} y_i, \quad s_{y,k}^2 = \frac{1}{N_k(\boldsymbol{S})}\sum_{i:S_i=k}(y_i - \bar{y}_k(\boldsymbol{S}))^2.$$

Similarly, for σ_k^2 we have

$$\begin{aligned}\sigma_k^2|\mathbf{y}, \boldsymbol{S}, \mu_k &\sim \mathscr{G}^{-1}(c_k(\boldsymbol{S}), C_k(\boldsymbol{S})),\\ c_k(\boldsymbol{S}) &= a_0 + \frac{1}{2}N_k(\boldsymbol{S}),\\ C_k(\boldsymbol{S}) &= b_0 + \frac{1}{2}\sum_{i:S_i=k}(y_i-\mu_k)^2.\end{aligned}$$

According to the above hierarchical prior setting, the joint distribution of the data and the model parameters can be expressed as

$$\begin{aligned}&p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})\\ &= \prod_{i=1}^{N}\prod_{k=1}^{K}\left(p(\mathbf{y}_i|\mu_k, \sigma_k^2)\eta_k\right)^{I_{S_i=k}}\prod_{k=1}^{K}p(\mu_k|\sigma_k^2, \eta_k)p(\sigma_k^2)p(\eta_k)\\ &= \prod_{k=1}^{K}\left(\prod_{i:S_i=k}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\sum\limits_{i:S_i=k}(y_i-\mu_k)^2}{2\sigma_k^2}\right\}\right)\left(\prod_{k=1}^{K}\eta_k^{\sum\limits_{i=1}^{N}I_{S_i=k}}\right)\end{aligned}$$

$$\begin{aligned}&\times\prod_{k=1}^{K}\left(\frac{N_0\eta_k}{2\pi\sigma_k^2}\right)^{1/2}\exp\left\{-\frac{N_0\eta_k}{2\sigma_k^2}(\mu_k-\mu_{k0})^2\right\}\\ &\times\prod_{k=1}^{K}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma_k^2)^{-a_0-1}\exp\left\{-b_0/\sigma_k^2\right\}\times\frac{\Gamma(\sum_{k=1}^{K}e_0)}{\prod_{k=1}^{K}\Gamma(e_0)}\prod_{k=1}^{K}\eta_k^{e_0-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}&p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})\\ &= \left(\frac{1}{2\pi}\right)^{\frac{\sum\limits_{k=1}^{K}N_k(S)}{2}}\left(\frac{N_0}{2\pi}\right)^{K/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\\ &\times\prod_{k=1}^{K}\exp\left\{-\frac{\sum\limits_{i:S_i=k}(y_i-\mu_k)^2+N_0\eta_k(\mu_k-\mu_0)^2+2b_0}{2\sigma_k^2}\right\}\\ &\times\prod_{k=1}^{K}\eta_k^{e_0+N_k(S)-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(S)+1}}\end{aligned}$$

After doing some simple algebra we get

$$
\begin{aligned}
&p(\mathbf{y}|\mathbf{S},\boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})\\
&=\left(\frac{1}{2\pi}\right)^{N/2}\left(\frac{N_0}{2\pi}\right)^{K/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{(N_k(\mathbf{S})+N_0\eta_k)\left[\mu_k-\frac{N_k(\mathbf{S})\bar{y}_k(\mathbf{S})+N_0\eta_k\mu_{k0}}{N_k(\mathbf{S})+N_0\eta_k}\right]^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S})+2b_0+\frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\eta_k^{e_0+N_k(S)-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(S)+1}}
\end{aligned}
$$

Now we are going to find the marginal posterior distribution of the allocation vector $p(\mathbf{S}|\mathbf{y})$ by integrating out all parameters. We first integrate out μ_k from the above expression and we get

$$
\begin{aligned}
&\prod_{k=1}^{K}\int p(\mathbf{y},\mathbf{S}|\boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta})p(\boldsymbol{\eta})d\mu_k\\
&=N_0^{K/2}\left(\frac{1}{2\pi}\right)^{N/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\prod_{k=1}^{K}\sqrt{\frac{1}{N_k(\mathbf{S})+N_0\eta_k}}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S})+2b_0+\frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\eta_k^{e_0+N_k(S)-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(S)+2}}
\end{aligned}
$$

Finally integrating out σ_k and η_k, the posterior $p(\mathbf{S}|\mathbf{y})$ is obtained as

$$
\begin{aligned}
p(\mathbf{S}|\mathbf{y})&=\int_0^1\iint p(\mathbf{y}|\boldsymbol{\eta},\mathbf{S},\boldsymbol{\theta})p(\mathbf{S}|\eta)\mathbf{p}(\boldsymbol{\eta})\mathbf{p}(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}\mathbf{d}\boldsymbol{\eta}\\
&=N_0^{K/2}\left(\frac{1}{\pi}\right)^{N/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}2^{Ka_0}\\
&\times\frac{\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})+e_0+1/2)}{\Gamma(N+Ke_0+K/2)}\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})/2+a_0)\\
&\times\prod_{k=1}^{K}\int_0^1\frac{\mathscr{B}(\eta_k)^{-a_0-\frac{N_k(\mathbf{S})}{2}}}{(N_k(\mathbf{S})+N_0\eta_k)^{1/2}}d\eta_k,
\end{aligned}
\tag{17.3}
$$

where

$$\mathscr{B}(\eta_k) = N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S}) + N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2.$$

As we can see for the case with dependent hierarchical priors there is no explicit form for the posterior $p(\boldsymbol{S}|\mathbf{y})$. Due to this formulation, we faced some challenges in calculating the integration in the expression (17.3). Calculating this integration is not always possible in a usual way as a result of overflow or underflow, depending on simulation settings. To address this issue we calculate this definite integral by calculating Riemann sums over a partition of [0, 1].

Note that

$$\int_0^1 f(\eta_k)d\eta_k = \int_0^1 \frac{(N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2)^{-a_0-\frac{N_k(\mathbf{S})}{2}}}{(N_k(\mathbf{S}) + N_0\eta_k)^{1/2}}.$$

We rearrange the above integrand as follows:

$$f(\eta_k) = \frac{(N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}))^{-a_0-N_k(\mathbf{S})/2}D(\eta_k)^{-a_0-N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}\left(1 + \frac{N_0\eta_k}{N_k(\mathbf{S})}\right)^{1/2}}, \tag{17.4}$$

where

$$D(\eta_k) = 1 + \frac{1}{N_k(\mathbf{S})S^2_{y,k}(\mathbf{S})}\left(2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S}) + N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2\right) \tag{17.5}$$

We partition [0, 1] into subintervals $[x_0, x_1], [x_1, x_2], \cdots, [x_{n-1}, x_n]$ with $\varDelta x_i = x_i - x_{i-1} = 1/n$ and $x_i^* = i\varDelta x_i$. This leads to

$$\int_0^1 f(\eta_k)d\eta_k \approx \frac{1}{n}\sum_{i=1}^n f_k(x_i^*).$$

Even using the above approximation did not completely solve the problem of overflow and underflow and we still got some infinity values in numerical calculations. To tackle this problem we divide all summands by the largest element which is $f_k(x_n^*) = f_k(1)$. Therefore we calculate

$$\int_0^1 f(\eta_k)d\eta_k \approx \frac{f_k(x_n^*)}{n}\sum_{i=1}^n \frac{f_k(x_i^*)}{f_k(x_n^*)}, \tag{17.6}$$

where according to the equation (17.4) we have

$$\frac{f_k(x_i^*)}{f_k(x_n^*)} = \frac{(1 + \frac{N_0}{N_k(\mathbf{S})})^{1/2}}{(1 + \frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2} (\frac{D_k(x_i^*)}{D_k(1)})^{a_0 + N_k(\mathbf{S})/2}}.$$

Now according to the expression (17.5) we have

$$\frac{D_k(x_i^*)}{D_k(1)} \approx \frac{N_k(\mathbf{S}) S_{y,k}^2(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S}) N_0 x_i^*}{N_k(\mathbf{S}) + N_0 x_i^*} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}{N_k(\mathbf{S}) S_{y,k}^2(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S}) N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}$$

$$= \frac{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0 x_i^*}{N_k(\mathbf{S}) + N_0 x_i^*} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}.$$

In order to use the latter expression in computational programming and avoid any possible underflow issue, we further rearrange the latter expression to get

$$\frac{D_k(x_i^*)}{D_k(1)} = 1 - \frac{(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2 \left(\frac{N_0}{N_k(\mathbf{S}) + N_0} - \frac{N_0 x_i^*}{N_k(\mathbf{S}) + N_0 x_i^*} \right)}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}$$

$$= 1 - \frac{\frac{N_0}{N_k(\mathbf{S}) + N_0} \left(1 - \frac{(N_k(\mathbf{S}) + N_0) x_i^*}{N_k(\mathbf{S}) + N_0 x_i^*} \right) (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}_k(\mathbf{S}) - \mu_{k0})^2}.$$

Now the integration in (17.6) can be approximated by the following summation

$$\frac{f_k(x_n^*)}{n} \sum_{i=1}^{n} \frac{f_k(x_i^*)}{f_k(x_n^*)} = \frac{1}{n} \frac{(N_k(\mathbf{S}) S_{y,k}^2(\mathbf{S}))^{-a_0 - \frac{N_k(\mathbf{S})}{2}} D(1)^{-a_0 - N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}}$$

$$\times \sum_{i=1}^{n} \frac{1}{(1 + \frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2} (\frac{D_k(x_i^*)}{D_k(1)})^{a_0 + N_k(\mathbf{S})/2}}$$

Substituting the above expression in the allocation posterior results in

$$p(\mathbf{S}|\mathbf{y}) \approx N_0^{K/2} \left(\frac{1}{\pi} \right)^{N/2} \left(\frac{b_0^{a_0}}{\Gamma(a_0)} \right)^K \frac{\Gamma(K e_0)}{\Gamma(e_0)^K} 2^{K a_0}$$

$$\times \frac{\prod_{k=1}^{K} \Gamma(N_k(\mathbf{S}) + e_0 + 1/2)}{\Gamma(N + K e_0 + K/2)} \prod_{k=1}^{K} \Gamma(N_k(\mathbf{S})/2 + a_0)$$

$$\times \frac{1}{n} \frac{(N_k(\mathbf{S}) S_{y,k}^2(\mathbf{S}))^{-a_0 - N_k(\mathbf{S})/2} D(1)^{-a_0 - N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}}$$

$$\times \sum_{i=1}^{n} \frac{1}{(1 + \frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2} (\frac{D_k(x_i^*)}{D_k(1)})^{a_0 + N_k(\mathbf{S})/2}}$$

Taking the logarithm, we have

$$\begin{aligned}\log(p(\mathbf{S}|\mathbf{y})) \approx\ & K/2\log(N_0) - N/2\log(\pi) + Ka_0\log(b_0) - K\log\Gamma(a_0)\\ & + \log\Gamma(Ke_0) - K\log\Gamma(e_0) + Ka_0\log(2) - \Gamma(N + Ke_0 + K/2)\\ & + \sum_{k=1}^{K}\log\Gamma(N_k(\mathbf{S}) + e_0 + 1/2) + \sum_{k=1}^{K}\log\Gamma(a_0 + \frac{N_k(\mathbf{S})}{2})\\ & - \sum_{k=1}^{K}(a_0 + N_k(\mathbf{S})/2)\left[\log(N_k(\mathbf{S})) + \log(S^2_{y,k}) + \log(D(1))\right]\end{aligned}$$

$$- \sum_{k=1}^{K}\log(n) + \sum_{k=1}^{K}\log\sum_{i=1}^{n}\frac{1}{(1 + \frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2}(\frac{D_k(x_i^*)}{D_k(1)})^{a_0+N_k(\mathbf{S})/2}} \quad (17.7)$$

17.3 A Simulation Study

In this section, we conduct simulations to compare the classification accuracy of Bayesian normal mixture model with that of normal mixture models. We implement the Bayesian normal mixture model based on both independent priors and dependent priors. In order to compare the performance of the Bayesian mixture model to the frequentist model, we use the Mclust software, where the optimal allocation estimate is obtained by using the Expectation-Maximization algorithm.

17.3.1 Adjusted Rand Index

We review one of the widely used methods called adjusted Rand index for quantifying the degree of the agreement between partitions derived from different methods. Suppose we have n objects to classify and $P_1 = \{C_1, \cdots, C_r\}$ is a partition that assigns these objects into r classes and $P_2 = \{C_1, \cdots, C_s\}$ assigns them into s classes. Each pair of objects, either have the same class label or a different one. Since the number of classified objects is n, we have the total number of $n(n-1)/2$ pairs to compare. Let a be the number of pairs that the two partitions agree by assigning the elements to the same classes and b be the number of pairs that partitions agree by assigning them to different classes. Considering all pairs, the proportion of agreement between P_1 and P_2 is evaluated by the following Rand index (RI)

$$\mathrm{RI}(P_1, P_2) = \frac{a+b}{n(n-1)/2}$$

Since the expectation of Rand index for two random partitions is not a constant, [5] proposed a normalized Rand index which is defined by

$$\text{ARI} = \frac{\text{Rand index} - \text{Expected value of Rand index}}{\text{Maximum value of Rand index} - \text{Expected value of Rand index}}.$$

When two partitions completely agree, the adjusted Rand index reaches the maximum value 1. The higher ARI value, the greater degree of agreement between two partitions is.

17.3.2 Simulated Data

We generated data from a normal mixture model with three components. We used the same setting as used in one of the examples in [3] to generate the data. The underlying weights (0.3, 0.2, 0.5). The underlying component means and variances are $(-3, 0, 2)$ and $(1, 0.5, 0.8)$ respectively.

17.3.3 Results

We utilized the Bayesian mixture model under the following hierarchical priors where the component mean depends on the weight corresponding to that component

$$\mu_k|\sigma_k^2, \eta_k \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0\eta_k}), \quad \sigma_k^2 \sim IG(a_0, b_0), \quad \eta_k \sim D(e_0, \cdots, e_0),$$

which results in the log-allocation posterior in equation (17.7). We also implemented the Bayesian mixture model with hierarchical priors where the mean of each component was independent of the weight as following

$$\mu_k|\sigma_k^2 \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0}), \quad \sigma_k^2 \sim IG(a_0, b_0), \quad \eta_k \sim D(e_0, \cdots, e_0).$$

The allocation posterior can be regarded as a function of hyperparameters $N_0, a_0, b_0, e_0, \mu_{k0}$. Following [8], we set μ_{k0} to the median of the data. The hyperparameters are chosen as $a_0 = 2$ and $e_0 = 1$ and for the parameter b_0 they consider the prior $b_0 \sim G(0.2, 10/R^2)$ where R^2 is the length of the interval of the variation of the data. In order to choose N_0, following [6], we set $N_0 = 2.6/(y_{\max} - y_{\min})$.

We found the optimal classification by maximizing the logarithm of the allocation posterior. The optimization was carried out by the following iterative algorithm: We updated the coordinates of the allocation vector in one-by-one and calculated the corresponding posterior. The algorithm started with an initial allocation vector $\boldsymbol{S}^{(0)} = \boldsymbol{S}_{\text{current}}$ by the result derived from MClust. For example, to update the coordinate S_1 corresponding to y_1 while other coordinates were fixed, we generated a random number U from the uniform distribution $\mathscr{U}[0, 1]$. If $U < \eta_1$, assign the observation y_1 to the first component. If $\eta_1 \leq U < \eta_1 + \eta_2$, assign the observation to the second component. Otherwise, assign y_1 to the third component. This resulted in an updated allocation vector $\boldsymbol{S}^{(1)} = \boldsymbol{S}_{\text{updated}}$. The number of elements in each component changed. If $S_1^{\text{new}} = S_1 = k$, then no moving occurred whereas, if the observation moved to another component, say l, then the number of observations in each component was updated as

$$N_k(S_1^{\text{new}}, \boldsymbol{S}_{-1}) = N_k(\boldsymbol{S}) - 1, \quad N_l(S_1^{\text{new}}, \boldsymbol{S}_{-1}) = N_l(\boldsymbol{S}) + 1,$$

where $\boldsymbol{S}_{-1} = (S_2, ..., S_N)$. Correspondingly, the mean $\bar{y}_k(\boldsymbol{S})$ and the variance $S_{y,k}(\boldsymbol{S})$ of each component were updated. Then the log-posterior $p((S_1^{\text{new}}, \boldsymbol{S}_{-1})|\mathbf{y})$ of the updated allocation vector was calculated using the expression (17.7). The updated allocation for the first observation was accepted if the updated posterior was greater than the current posterior, i.e. $p((S_1^{\text{new}}, \boldsymbol{S}_{-1})|\mathbf{y}) > p(\boldsymbol{S}^{(0)}|\mathbf{y})$. If the new allocation was accepted, then this updated allocation was used as the current allocation in the next iteration $\boldsymbol{S}_{\text{current}} = \boldsymbol{S}_{\text{updated}}$ and the observation was moved to the component l. Otherwise, the observation was kept in the current component k and the algorithm moved to the next observation y_2. These steps were repeated until all observations $i = 1 \cdots , N$ were updated and until the posterior reaches a local maximum. Then this optimal allocation vector was recorded and compared with Mclust by computing their adjusted Rand index.

We simulated 300 datasets from a three component mixture of normals. We applied the above algorithm to find the optimal grouping for each of these data. We applied both the weight-dependent (17.7) and weight-independent (17.2) prior approaches.

Results displayed in Fig. 17.1 show that the Bayesian clustering outperformed the Mclust particularly when the component priors were weight-dependent. The results illustrated that imposing dependency of component priors on weights can reduce the bias of clustering due to the effect of weight heterogeneity. Note that if we used a more refined optimization algorithm such as evolutionary Markov chain Monte Carlo algorithms rather than a simple coordinate-wise updating optimization, then the result would be further improved. See [10, 11].

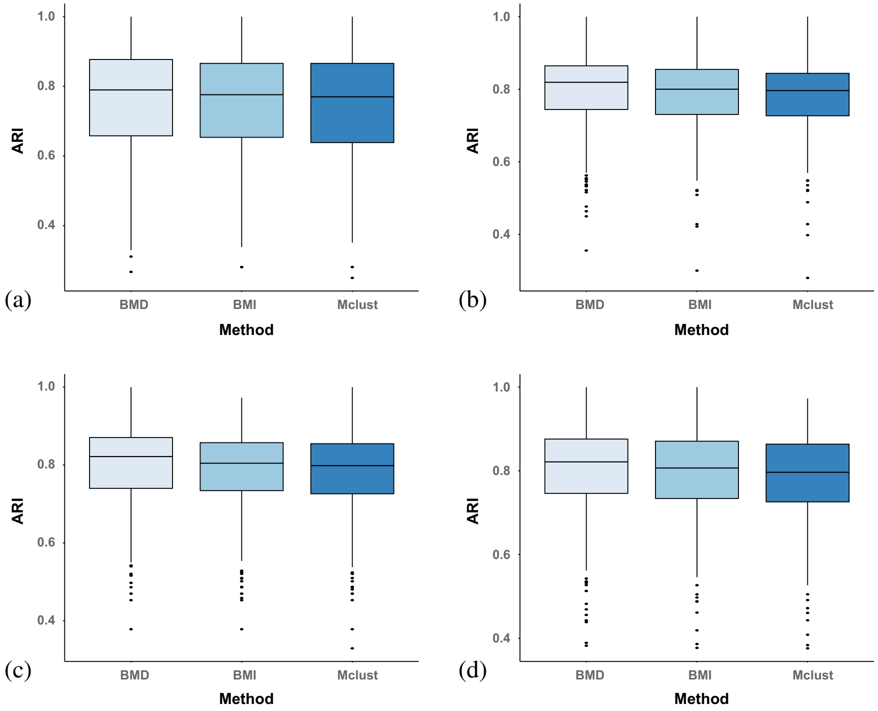

Fig. 17.1 Box plots of adjusted Rand index values corresponding to the classifications performed by applying Bayesian mixture model with weight-dependent priors (BMD), Bayesian mixture model with weight-independent priors (BMI) and the non-Bayesian mixture of normals (Mclust), where $N_0 = 2.6/(y_{\max} - y_{\min})$ and $b_0 \sim G(0.2, 10/R^2)$ where R^2 is the length of the interval of the variation of the data. Other hyperparameters and sample size are chosen as follows: (a) $N = 50$, $a_0 = 2$, $e_0 = 1$, (b) $N = 100$, $a_0 = 2$, $e_0 = 1$, (c) $N = 100$, $a_0 = 5$, $e_0 = 1$, (d) $N = 100$, $a_0 = 5$, $e_0 = 2$

17.4 Application to a Real Dataset

We applied to the so-called ’acidity data’, which concerns an acidity index measured in a sample of 155 lakes in north-central Wisconsin and was previously analysed using a Bayesian mixture of Gaussian distributions on the log-scale by [8]. These authors calculated the posterior for K (the number of components) favours $3 \sim 5$ components. Here, letting $K = 3$, we applied the BMD, BMI and Mclust to the dataset respectively. The three clustering results presented in Fig. 17.2 reveal that BMD performed better in dealing with outliers in the dataset: Unlike BMD, both Cluster 2 derived from BMI or Mclust contained 3 outliers which should belong to Cluster 1.

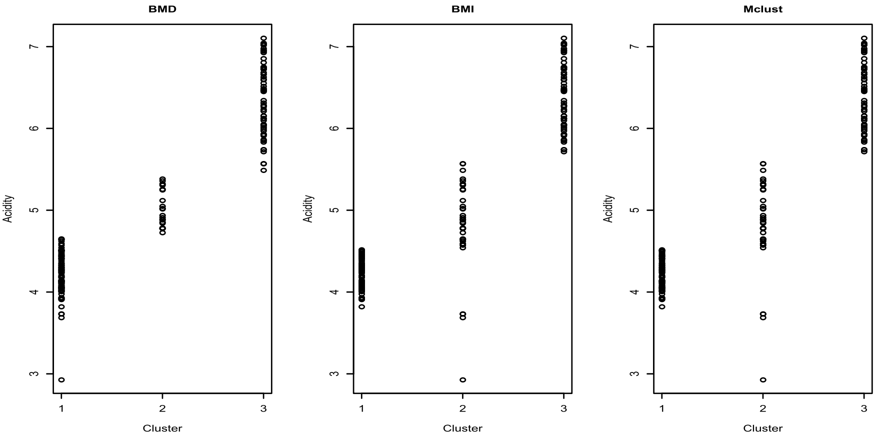

Fig. 17.2 From left to right, the panel presented the clusters derived from BMD, BMI and Mclust. We used the same approach to set hyperparameters as in our simulation study

17.5 Conclusion

In this paper, we have developed a novel prior scheme for Bayesian mixture models. Unlike the classical prior specification, we allow the component priors to depend on their weights (i.e., mixing proportions). This help us tackle the effect of varying weights on estimation of hidden group memberships of the observations. We have conducted a simulation study to compare the proposed method to the existing approaches. The simulation results have shown that the new method can performed better than its competitors in terms of adjusted Rand index. A real data application has suggested that the proposal method is more robust to outliers than the existing methods.

Acknowledgements The research of Elaheh Oftadeh is supported by the 50th anniversary PhD scholarship from the University of Kent.

References

1. Binder, D.A.: Bayesian cluster analysis. Biometrika **65**, 31–38 (1978)
2. Diebolt, J., Robert, C.P.: Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. Series B **56**, 363–375 (1994)
3. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes. Springer, New York (2006)
4. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **971**, 611–631 (2002)
5. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

6. Raftery, A.E.: Hypothesis testing and model selection. In: Gilks, W.R., Spiegelhalter, D.J., Richardson, S. (eds.) Markov Chain Monte Carlo in Practice, pp. 163–188. Chapman and Hall, London (1996)
7. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
8. Richardson, S., Green, P. J.: On Bayesian analysis of mixtures with an unknown number of components (with discussions), J. R. Stat. Soc. Series B (statistical methodology) **5**9, 731–792 (1997)
9. Roeder, K., Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**, 894–902 (1997)
10. Zhang, J.: A Bayesian model for biclustering with applications. JRSSC (Applied Statistics) **59**, 635–656 (2010)
11. Zhang, J.: Genralized plaid models. Neurocomputing **79**, 95–104 (2012)