3

Practical Bayesian Inference for a Finite Mixture Model with Known Number of Components

3.1 Introduction

Assume as in Chapter 2 that N observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, drawn randomly from a finite mixture of $\mathcal{T}(\boldsymbol{\theta})$ distributions with density $p(\mathbf{y}|\boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta} \in \Theta$, are available, which should be used to make inferences about the underlying mixture structure. In this chapter we outline in detail Bayesian inference for the standard finite mixture model,

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k p(\mathbf{y}_i|\boldsymbol{\theta}_k), \tag{3.1}$$

when the number of components is known.

If $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\eta})$ are unknown parameters that need to be estimated from the data then, as noted earlier, from a Bayesian perspective all information contained in the data $\mathbf{y}$ about $\boldsymbol{\vartheta}$ is summarized in terms of the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, which is derived using Bayes' theorem:

$$p(\boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}). \tag{3.2}$$

By Bayes' theorem, the data-dependent mixture likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$, defined earlier, is combined with a prior density $p(\boldsymbol{\vartheta})$ in order to obtain the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$. For Bayesian estimation, we have to assume that such a prior distribution $p(\boldsymbol{\vartheta})$ is available. For finite mixture models it is not possible to choose an improper prior such as $p(\boldsymbol{\vartheta}) \propto$ constant, because this leads to an improper mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$. This problem and choosing proper priors are discussed in Section 3.2.

Within a Bayesian analysis of a finite mixture model we are interested in the entire mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, which to a large extent is dominated by the mixture likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$, and, as discussed in Section 3.3, inherits all of its properties, in particular the invariance to relabeling the mixture components.

In Section 3.4 we discuss inference for the group indicators $\mathbf{S}$ without parameter estimation which is interesting in its own right and provides an opportunity to introduce recent concepts of computational Bayesian statistics such as Gibbs sampling and the Metropolis–Hastings algorithm. Gibbs sampling, together with data augmentation, is also useful for drawing Bayesian inference about the parameters of a mixture model (reviewed in detail in Section 3.5) and is the most commonly used approach for obtaining draws from the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$; other sampling-based approaches such as the Metropolis–Hastings algorithm are briefly discussed in Section 3.6. Finally, it is discussed in Section 3.7 how draws from the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ could be used within a Bayesian approach to obtain inference on quantities of interest such as the unknown component parameters.

3.2 Choosing the Prior for the Parameters of a Mixture Model

3.2.1 Objective and Subjective Priors

For Bayesian estimation of a finite mixture model a prior $p(\boldsymbol{\vartheta})$ has to be selected for the unknown parameters $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\eta})$. As in Press (2003, Chapter 5), one may distinguish between objective and subjective priors.

Objective priors should reflect the notion of having no prior information, however, there exists no general agreement about how knowing little about a parameter $\boldsymbol{\vartheta}$ should be expressed in terms of a probability distribution $p(\boldsymbol{\vartheta})$. Very often improper priors, which are not integrable over the parameter space, are used to express complete ignorance, in the hope that the data are informative enough to turn the improper prior $p(\boldsymbol{\vartheta})$ into a proper posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$. The choice of objective priors is particularly difficult for finite mixture models, as common improper priors will lead to improper posteriors; see Subsection 3.2.2.

Subjective priors bring prior knowledge into the analysis, and offer the advantage of being proper. For finite mixture models, such priors are usually obtained by choosing priors that are conjugate for the complete-data likelihood function; see Subsection 3.2.3. It is common to assume that the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are independent of the weight distribution $\boldsymbol{\eta}$:

$$p(\boldsymbol{\vartheta}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)p(\boldsymbol{\eta}). \tag{3.3}$$

For finite mixture models, the standard prior for the weight distribution $\boldsymbol{\eta}$ is the $\mathcal{D}(e_0, \ldots, e_0)$-distribution, which arises from the same prior distribution family as for complete-data Bayesian inference considered in Subsection 2.3.4, however, the hyperparameters of the prior are assumed to be the same, in order to obtain an invariant prior. The precise prior on the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ depends on the distribution family underlying the mixture distribution.

It is not always easy to assess the parameters of a subjective prior, also called hyperparameters. Results from a Bayesian analysis of finite mixture models using subjective prior information is often highly dependent on particular choices of hyperparameters. To reduce this sensitivity, it is common practice in the context of finite mixture modeling to use hierarchical priors where the hyperparameter is equipped with a prior of its own; see Subsections 3.2.4.

In any case, for a Bayesian analysis of finite mixture models the prior distribution has to be selected with some care.

3.2.2 Improper Priors May Cause Improper Mixture Posteriors

Assume that in (3.3), complete ignorance about $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ is expressed in terms of the product of independent improper priors:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \propto \prod_{k=1}^{K} p^\star(\boldsymbol{\theta}_k), \tag{3.4}$$

with $\int p^\star(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k = \infty$. Roeder and Wasserman (1997b) show that the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$ is improper under prior (3.4), by rewriting the mixture likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ as a sum over complete-data likelihoods:

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{\mathbf{S} \in \mathcal{S}_K} p(\mathbf{y}|\mathbf{S}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) p(\mathbf{S}|\boldsymbol{\eta}), \tag{3.5}$$

where summation runs over all K^N possible classifications $\mathbf{S}$. Under prior (3.4), the mixture posterior is proportional to

$$p(\boldsymbol{\vartheta}|\mathbf{y}) \propto \sum_{\mathbf{S} \in \mathcal{S}_K} p(\mathbf{y}|\mathbf{S}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \prod_{k=1}^{K} p^\star(\boldsymbol{\theta}_k) p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}), \tag{3.6}$$

and is proper, if the integral over the right-hand side is finite. The normalizing constant turns out to be

$$\begin{aligned} &\sum_{\mathbf{S} \in \mathcal{S}_K} c_1(\mathbf{S}) c_2(\mathbf{S}), \qquad (3.7) \\ &c_1(\mathbf{S}) = \prod_{k=1}^{K} \int \left(\prod_{i: S_i = k} p(\mathbf{y}_i|\boldsymbol{\theta}_k) \right) p^\star(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k, \\ &c_2(\mathbf{S}) = \int p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta}. \end{aligned}$$

To obtain a proper posterior distribution, $c_1(\mathbf{S})$ and $c_2(\mathbf{S})$ have to be finite for *all* classifications $\mathbf{S}$. Note that the hidden multinomial prior on $\mathbf{S}$ assigns positive probability to partitions $\mathbf{S}$, where one component, say j, is empty.

In this case, the complete-data likelihood does not contain any information about $\boldsymbol{\theta}_j$ and $c_1(\mathbf{S})$ is not finite under the improper prior (3.4) because

$$\int \left(\prod_{i:S_i=j} p(\mathbf{y}_i|\boldsymbol{\theta}_j) \right) p^\star(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j = \int p^\star(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j = \infty.$$

To obtain proper posterior distributions under the prior (3.4), Wasserman (2000) modifies the prior distribution $p(\mathbf{S})$ of the allocations $\mathbf{S}$, by restricting $\mathcal{S}_K$ to allocations with nonempty components.

3.2.3 Conditionally Conjugate Priors

Whereas it is not possible to choose simple conjugate priors for the mixture likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$, a conjugate analysis is possible for the complete-data likelihood $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta})$, if the component densities in the mixture come from the exponential family as in (1.11),

$$p(\mathbf{y}_i|\boldsymbol{\theta}_k) = \exp\Big\{ \phi(\boldsymbol{\theta}_k)' u(\mathbf{y}_i) - g(\boldsymbol{\theta}_k) + c(\mathbf{y}_i) \Big\};$$

see also Subsection 2.3.3. To formulate a joint prior for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, the component parameters are assumed to be independent a priori, given a hyperparameter $\boldsymbol{\delta}$:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\boldsymbol{\delta}) = \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\boldsymbol{\delta}). \tag{3.8}$$

If for each component the prior $p(\boldsymbol{\theta}_k|\boldsymbol{\delta})$ takes the form

$$p(\boldsymbol{\theta}_k|\boldsymbol{\delta}) \propto \exp\Big\{ \phi(\boldsymbol{\theta}_k)' a_0 - g(\boldsymbol{\theta}_k) b_0 \Big\}, \tag{3.9}$$

with hyperparameter $\boldsymbol{\delta} = (a_0, b_0)$, then the conditional posterior $p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y})$ is given by:

$$p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y}) \propto \exp\Big\{ \phi(\boldsymbol{\theta}_k)' a_k - g(\boldsymbol{\theta}_k) b_k \Big\}, \tag{3.10}$$

which is again a density from the chosen exponential family with

$$a_k = a_0 + \sum_{i:S_i=k} u(\mathbf{y}_i), \qquad b_k = b_0 + N_k(\mathbf{S}),$$

where $N_k(\mathbf{S}) = \#\{S_i = k\}$. For mixtures of Poisson distributions, for instance, Bayesian inference for the complete data problem, already studied in Subsection 2.3.3, leads to the conditionally conjugate prior $\mu_k \sim \mathcal{G}(a_0, b_0)$, where a_0 as well as b_0 have to be positive to obtain a proper posterior distribution.

3.2.4 Hierarchical Priors and Partially Proper Priors

For practical Bayesian inference, prior (3.8) is assessed by choosing the hyperparameter $\boldsymbol{\delta}$. Prior (3.8) acts as a kind of shrinkage prior, pulling all component parameters $\boldsymbol{\theta}_k$ toward a common center defined by $\mathrm{E}(\boldsymbol{\theta}_k|\boldsymbol{\delta})$, where both the center of the prior as well as the amount of shrinkage may crucially depend on $\boldsymbol{\delta}$. For illustration, consider a mixture of Poisson distributions, and rewrite the conditionally conjugate $\mathcal{G}(a_0, b_0)$-prior introduced in Subsection 3.2.3 as $\mu_k \sim (a_0/b_0)W_k$, $W_k \sim \mathcal{G}(a_0, a_0)$. Evidently, this prior induces shrinkage of μ_k toward the prior mean $\mathrm{E}(\mu_k) = a_0/b_0$,, with shrinkage being more pronounced the larger a_0.

In particular for mixtures with small components, the posterior distribution may be sensitive to specific choices of $\boldsymbol{\delta}$. To reduce sensitivity to specific choices of $\boldsymbol{\delta}$, it is common practice to use hierarchical priors, which treat $\boldsymbol{\delta}$ as an unknown quantity with a prior $p(\boldsymbol{\delta})$:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\delta}) = p(\boldsymbol{\delta}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\boldsymbol{\delta}). \tag{3.11}$$

As a result, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are dependent a priori:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = \int p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K|\boldsymbol{\delta})p(\boldsymbol{\delta})d\boldsymbol{\delta} \neq \prod_{k=1}^{K} p(\boldsymbol{\theta}_k).$$

Such priors have been applied to finite mixtures of normal distributions in Mengersen and Robert (1996), Richardson and Green (1997), and Roeder and Wasserman (1997b).

Partially proper priors (Roeder and Wasserman, 1997b) are hierarchical priors where the prior $p(\boldsymbol{\delta})$ of the hyperparameter $\boldsymbol{\delta}$ is improper. Although, marginally, the prior $p(\boldsymbol{\theta}_k)$ is improper, the posterior distribution is proper.

A Hierarchical Prior for Poisson Mixtures

For a mixture of Poisson distributions, a hierarchical prior is obtained by assuming that b_0 is a random parameter with a prior of its own:

$$\mu_k|b_0 \sim \mathcal{G}(a_0, b_0), \qquad b_0 \sim \mathcal{G}(g_0, G_0). \tag{3.12}$$

Then the component means $\mu_1, \ldots, \mu_K$ are dependent a priori, and the joint prior $p(\mu_1, \ldots, \mu_K)$, where b_0 is integrated out, is available in closed form, if the $\mathcal{G}(g_0, G_0)$-prior is proper:

$$p(\mu_1,\ldots,\mu_K) = \int p(\mu_1,\ldots,\mu_K|b_0)p(b_0)db_0 \tag{3.13}$$
$$= \frac{G_0^{g_0}\Gamma(g_0+Ka_0)\left(\prod_{k=1}^{K}\mu_k\right)^{a_0-1}}{\Gamma(a_0)^K\Gamma(g_0)\left(G_0+\sum_{k=1}^{K}\mu_k\right)^{g_0+Ka_0}}.$$

A partially proper prior results if the $\mathcal{G}(g_0, G_0)$-prior is improper, for example, if $g_0 = 0.5$ and $G_0 = 0$.

3.2.5 Other Priors

Reference priors were suggested by Bernardo (1979) as prior distributions having a minimal effect on the final inference, relative to the data. The derivation of such a reference prior, however, is less than obvious for mixture models. Reference priors depend on the asymptotic behavior of the relevant posterior distributions. Although several papers have established the limiting properties of maximum likelihood estimators in finite mixture models (see Subsection 2.4.4), the derivation of reference priors for general finite mixture models still seems infeasible.

Some investigations appear in Bernardo and Girón (1988) for a mixture model where only the weight distribution $\boldsymbol{\eta}$ is unknown. For a mixture of two known densities, the reference prior for η_1 is virtually Jeffrey's $\mathcal{B}\left(\frac{1}{2}, \frac{1}{2}\right)$-prior, when the two densities are well separated, whereas the uniform $\mathcal{B}(1,1)$ would approximate the reference prior when the two densities are very close. For a mixture of more than two known densities Bernardo and Girón (1988) suggest that a Dirichlet distribution with parameters ranging in the interval $[\frac{1}{2}, 1]$ is a reasonable approximation to the reference prior.

3.2.6 Invariant Prior Distributions

Because the components in a mixture density may be arbitrarily arranged, it is usual to choose priors that reflect this information, by being invariant to relabeling the components. Consider all $s = 1,\ldots,K!$ different permutations ρ_s:$\{1,\ldots,K\} \to \{1,\ldots,K\}$, where the value $\rho_s(k)$ is assigned to each value $k \in \{1,\ldots,K\}$. Let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K,\eta_1,\ldots,\eta_K)$ be an arbitrary parameter in $\Theta_K = \Theta^K \times \mathcal{E}_K$, and define for each permutation ρ_s the parameter $\tilde{\boldsymbol{\vartheta}}_s$ by

$$\tilde{\boldsymbol{\vartheta}}_s = (\boldsymbol{\theta}_{\rho_s(1)},\ldots,\boldsymbol{\theta}_{\rho_s(K)},\eta_{\rho_s(1)},\ldots,\eta_{\rho_s(K)}). \tag{3.14}$$

A prior density $p(\boldsymbol{\vartheta})$ is invariant to relabeling the components of the mixture model, if the following identity holds for all $\boldsymbol{\vartheta} \in \Theta_K$, for any of the $K!$ permutations $\rho_s(\cdot)$;

$$p(\tilde{\boldsymbol{\vartheta}}_s) = p(\boldsymbol{\vartheta}). \tag{3.15}$$

Any of the prior distributions discussed so far in this section is invariant by construction.

Nonsymmetric priors have been applied in the hope that this eliminates all modes of the mixture likelihood function but one and Bayesian inference leads to a unimodal posterior distribution. Because this is not necessarily the case (see, for instance, Chib, 1995), the recommendation is to use an invariant prior unless there is a structural asymmetry in the mixture distribution. One example is Bayesian outlier modeling based on finite mixture, where it is sensible to choose priors that are not invariant, because the outlier component is much smaller than the other components by definition; see Section 7.2 for more detail.

3.3 Some Properties of the Mixture Posterior Density

3.3.1 Invariance of the Posterior Distribution

The mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ defined in (3.2) is to a large extent dominated by the mixture likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$, which is invariant to relabeling the components of the mixture distribution. Under an invariant prior, the mixture posterior distribution inherits the invariance of the mixture likelihood to relabeling the components of the mixture, and the following identity holds for all $\boldsymbol{\vartheta} \in \Theta_K$, for any of the $K!$ permutations $\rho_s(\cdot)$;

$$p(\tilde{\boldsymbol{\vartheta}}_s|\mathbf{y}) = p(\boldsymbol{\vartheta}|\mathbf{y}). \tag{3.16}$$

It is quite illuminating to study the behavior of the posterior density as N increases. The following considerations are purely heuristic, without providing a formal proof.

Let $\boldsymbol{\vartheta}^{\text{true}} = (\boldsymbol{\theta}_1^{\text{true}}, \ldots, \boldsymbol{\theta}_K^{\text{true}}, \eta_1^{\text{true}}, \ldots, \eta_K^{\text{true}})$ denote the true value of $\boldsymbol{\vartheta}$. Assume that $\boldsymbol{\vartheta}^{\text{true}}$ fulfills the formal identifiability constraints of Subsection 1.3.3, with $\eta_k^{\text{true}} > 0$, and $\boldsymbol{\theta}_k^{\text{true}} \neq \boldsymbol{\theta}_l^{\text{true}}$, for all $k \neq l$, where in a multiparameter setting not all components of all parameters need to be different. Let $\mathcal{U}^P(\boldsymbol{\vartheta}^{\text{true}})$ be the set defined in (1.27). Due to the formal identifiability constraints the mixture model is not overfitting and the set $\mathcal{U}^P(\boldsymbol{\vartheta}^{\text{true}})$ contains $K!$ distinct points, obtained from relabeling all components of $\boldsymbol{\vartheta}^{\text{true}}$ through all possible permutations of $\{1, \ldots, K\}$.

Then with increasing number of observations, the posterior density has $K!$ equivalent modes and becomes proportional to an invariant mixture of asymptotic normal distributions, with the modes lying in the set $\mathcal{U}^P(\boldsymbol{\vartheta}^{\text{true}})$:

$$p(\boldsymbol{\vartheta}|\mathbf{y}) \approx \frac{1}{K!} \sum_{s=1}^{K!} f_N(\tilde{\boldsymbol{\vartheta}}_s; \boldsymbol{\vartheta}^{\text{true}}, \mathcal{I}(\boldsymbol{\vartheta}^{\text{true}})).$$

3.3.2 Invariance of Seemingly Component-Specific Functionals

The invariance property of the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, discussed in the previous subsection, causes state independence of many functionals derived from the posterior distribution, which are seemingly component specific, like the posterior mean $\mathrm{E}(\boldsymbol{\theta}_k|\mathbf{y})$.

Marginal Distributions of Component-Specific Parameters

Consider, as an example the marginal distribution of the component parameter $\boldsymbol{\theta}_k$, which is defined in the usual way as

$$p(\boldsymbol{\theta}_k|\mathbf{y}) = \int_{\Theta^{K-1}\times\mathcal{E}_K} p(\boldsymbol{\vartheta}|\mathbf{y}) d(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k-1},\boldsymbol{\theta}_{k+1},\ldots,\boldsymbol{\theta}_K,\eta_1,\ldots,\eta_K).$$

Consider an arbitrary permutation $\rho_s(1),\ldots,\rho_s(K)$ of $\{1,\ldots,K\}$, which is different from the identity, to transform the parameter in this integration. The Jacobian of the transformation being 1, the area of integration being unchanged, one obtains:

$$p(\boldsymbol{\theta}_k|\mathbf{y}) = \int_{\Theta^{K-1}\times\mathcal{E}_K} p(\tilde{\boldsymbol{\vartheta}}_s|\mathbf{y}) \\ d(\boldsymbol{\theta}_{\rho_s(1)},\ldots,\boldsymbol{\theta}_{\rho_s(k-1)},\boldsymbol{\theta}_{\rho_s(k+1)},\ldots,\boldsymbol{\theta}_{\rho_s(K)},\eta_{\rho_s(1)},\ldots,\eta_{\rho_s(K)}).$$

By the invariance property (3.16) this is equal to:

$$p(\boldsymbol{\theta}_k|\mathbf{y}) = \int_{\Theta^{K-1}\times\mathcal{E}_K} p(\boldsymbol{\vartheta}|\mathbf{y}) \\ d(\boldsymbol{\theta}_{\rho_s(1)},\ldots,\boldsymbol{\theta}_{\rho_s(k-1)},\boldsymbol{\theta}_{\rho_s(k+1)},\ldots,\boldsymbol{\theta}_{\rho_s(K)},\eta_{\rho_s(1)},\ldots,\eta_{\rho_s(K)}).$$

Marginalization is with respect to all unknown parameters except $\boldsymbol{\theta}_{\rho_s(k)}$, therefore $p(\boldsymbol{\theta}_k|\mathbf{y}) = p(\boldsymbol{\theta}_{\rho_s(k)}|\mathbf{y})$. Because this holds all permutations $s = 1,\ldots,K!$, the seemingly component-specific marginal posterior densities $p(\boldsymbol{\theta}_k|\mathbf{y})$ are actually state-independent and the same for all $k \neq k'$:

$$p(\boldsymbol{\theta}_k|\mathbf{y}) = p(\boldsymbol{\theta}_{k'}|\mathbf{y}). \tag{3.17}$$

It could be proven in a similar way that the marginal posterior density of the component weight η_k is state-independent:

$$p(\eta_k|\mathbf{y}) = p(\eta_{k'}|\mathbf{y}), \tag{3.18}$$

for all $k \neq k'$. State independence holds for other marginal densities, such as the marginal distribution of any two parameters from different components where $k \neq k'$ and ρ_s arbitrary:

$$p(\boldsymbol{\theta}_k,\boldsymbol{\theta}_{k'}|\mathbf{y}) = p(\boldsymbol{\theta}_{\rho_s(k)},\boldsymbol{\theta}_{\rho_s(k')}|\mathbf{y}). \tag{3.19}$$

As this relation holds in particular for $\rho_s(k) = k'$ and $\rho_s(k') = k$, the posterior in (3.19) is symmetric:

$$p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k'}|\mathbf{y}) = p(\boldsymbol{\theta}_{k'}, \boldsymbol{\theta}_k|\mathbf{y}). \tag{3.20}$$

For a mixture of univariate normal distributions, for instance, we obtain $\forall k, k' = 1, \ldots, K, k \neq k'$:

$$\begin{aligned}
&p(\mu_k|\mathbf{y}) = p(\mu_{k'}|\mathbf{y}), \qquad p(\sigma_k^2|\mathbf{y}) = p(\sigma_{k'}^2|\mathbf{y}),\\
&p(\mu_k, \sigma_k^2|\mathbf{y}) = p(\mu_{k'}, \sigma_{k'}^2|\mathbf{y}),\\
&p(\mu_k, \mu_{k'}|\mathbf{y}) = p(\mu_1, \mu_2|\mathbf{y}) = p(\mu_2, \mu_1|\mathbf{y}),\\
&p(\sigma_k^2, \sigma_{k'}^2|\mathbf{y}) = p(\sigma_1^2, \sigma_2^2|\mathbf{y}) = p(\sigma_2^2, \sigma_1^2|\mathbf{y}).
\end{aligned}$$

The Posterior Mean

The posterior mean is a commonly used point estimator, which is optimal with respect to a quadratic loss function; see, for instance, Zellner (1971) and Berger (1985). From the mixture posterior distribution, the following result may be derived,

$$\mathrm{E}(\tilde{\boldsymbol{\vartheta}}_s|\mathbf{y}) = \mathrm{E}(\boldsymbol{\vartheta}|\mathbf{y}), \tag{3.21}$$

where the parameter $\tilde{\boldsymbol{\vartheta}}_s$ has been defined for each permutation ρ_s in (3.14). Identity (3.21) follows from the invariance property (3.16).

As (3.21) holds for all permutations, it follows that the seemingly component-specific posterior mean of $\boldsymbol{\theta}_k$ and η_k is actually state-independent:

$$\mathrm{E}(\boldsymbol{\theta}_k|\mathbf{y}) = \mathrm{E}(\boldsymbol{\theta}_{k'}|\mathbf{y}), \qquad \mathrm{E}(\eta_k|\mathbf{y}) = \mathrm{E}(\eta_{k'}|\mathbf{y}),$$

for any $k \neq k'$. Consequently, the mean $\mathrm{E}(\boldsymbol{\vartheta}|\mathbf{y})$ of the mixture posterior is not a sensible point estimator for the component parameters and the weight distribution. More sensible point estimators are discussed in Subsection 3.7.6.

3.3.3 The Marginal Posterior Distribution of the Allocations

We now turn to the posterior density $p(\mathbf{S}|\mathbf{y})$ of the allocations $\mathbf{S}$, which is of importance when dealing with Bayesian clustering in Section 7.1. $p(\mathbf{S}|\mathbf{y})$ is a discrete distribution over the lattice

$$\mathcal{S}_K = \{(S_1, \ldots, S_N) : S_i \in \{1, \ldots, K\}, i = 1, \ldots, N\}. \tag{3.22}$$

As noted by Chen and Liu (1996) and Casella et al. (2000), for many mixture models it is possible to derive an explicit form for the marginal posterior $p(\mathbf{S}|\mathbf{y})$ of the indicators $\mathbf{S}$, where dependence on the parameter $\boldsymbol{\vartheta}$ is integrated out. By Bayes' theorem, the marginal posterior $p(\mathbf{S}|\mathbf{y})$ is given by

$$p(\mathbf{S}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{S})p(\mathbf{S}), \tag{3.23}$$

where the integrated likelihood $p(\mathbf{y}|\mathbf{S})$ and the integrated prior $p(\mathbf{S})$ are equal to

$$p(\mathbf{y}|\mathbf{S}) = \int p(\mathbf{y}|\mathbf{S}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) d(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K),$$
$$p(\mathbf{S}) = \int p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta}.$$

Assume that the prior $p(\boldsymbol{\vartheta})$ takes exactly the same form as (3.3) and (3.8). Then:

$$p(\mathbf{y}|\mathbf{S}) = \prod_{k=1}^{K} \int \prod_{i:S_i=k} p(\mathbf{y}_i|\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k,$$
$$p(\mathbf{S}) = \int p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta}.$$

Under the conditionally conjugate prior $\boldsymbol{\eta} \sim \mathcal{D}(e_0, \ldots, e_0)$ we obtain:

$$p(\mathbf{S}) = \frac{\Gamma(Ke_0) \prod_{k=1}^{K} \Gamma(N_k(\mathbf{S}) + e_0)}{\Gamma(N + Ke_0)\Gamma(e_0)^K}, \tag{3.24}$$

where $N_k(\mathbf{S}) = \#\{S_i = k\}$. If the component densities in the mixture come from the exponential family as in (1.11), then under a conditionally conjugate prior $p(\boldsymbol{\theta}_k)$, the integrated likelihood $p(\mathbf{y}|\mathbf{S})$ is the product of the normalizing constants of each nonnormalized complete-data posterior, which are easily derived from (2.19):

$$p(\mathbf{y}|\mathbf{S}) = \prod_{k=1}^{K} \left(\frac{p(\boldsymbol{\theta}_k)}{p(\boldsymbol{\theta}_k|\mathbf{y}, \mathbf{S})} \prod_{i:S_i=k} p(\mathbf{y}_i|\boldsymbol{\theta}_k) \right). \tag{3.25}$$

For a mixture of Poisson distributions, for instance, this yields:

$$p(\mathbf{y}|\mathbf{S}) = \prod_{i=1}^{N} \frac{1}{\Gamma(y_i + 1)} \frac{b_0^{Ka_0}}{\Gamma(a_{0,k})^K} \prod_{k=1}^{K} \frac{\Gamma(a_k(\mathbf{S}))}{b_k(\mathbf{S})^{a_k(\mathbf{S})}},$$

where $a_k(\mathbf{S})$ and $b_k(\mathbf{S})$ are the posterior moments of the complete-data posterior densities given in (2.18):

$$a_k(\mathbf{S}) = a_0 + N_k(\mathbf{S})\overline{y}_k(\mathbf{S}),$$
$$b_k(\mathbf{S}) = b_0 + N_k(\mathbf{S}).$$

3.3.4 Invariance of the Posterior Distribution of the Allocations

State invariance occurs also for the seemingly component dependent allocations $\mathbf{S}$. $p(\mathbf{S}|\mathbf{y})$ is a marginal density obtained from integrating the joint posterior $p(\mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y})$ with respect to $\boldsymbol{\vartheta}$:

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

Because this holds for any $\mathbf{S}$, it also holds for $\tilde{\mathbf{S}}_s = (\rho_s(S_1), \ldots, \rho_s(S_N))$ for an arbitrary permutation. When using the same permutation for transforming the parameter $\boldsymbol{\vartheta}$ in this integration, we obtain:

$$p(\tilde{\mathbf{S}}_s|\mathbf{y}) = \int_{\Theta_K} p(\tilde{\mathbf{S}}_s, \tilde{\boldsymbol{\vartheta}}_s|\mathbf{y}) d\tilde{\boldsymbol{\vartheta}}_s = \int_{\Theta_K} p(\mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y}) d\tilde{\boldsymbol{\vartheta}}_s = p(\mathbf{S}|\mathbf{y}),$$

because the joint posterior is invariant to relabeling, and the order of integration may be rearranged arbitrarily. Therefore, for an arbitrary permutation $\rho_s(\cdot)$ of $\{1, \ldots, K\}$, the posterior density $p(\mathbf{S}|\mathbf{y})$ is invariant to relabeling:

$$p(S_1, \ldots, S_N|\mathbf{y}) = p(\rho_s(S_1), \ldots, \rho_s(S_N)|\mathbf{y}). \tag{3.26}$$

It follows that any two sequences $\mathbf{S}$ and $\mathbf{S}'$ that imply the same partition of the data obtain the same posterior probability. Consider, as a simple example, $N = 3$ and $K = 2$; then there are only four different partitions, each of which has the same posterior probability:

$$\begin{aligned} p(1,1,1|\mathbf{y}) &= p(2,2,2|\mathbf{y}), \qquad & p(2,1,1|\mathbf{y}) &= p(1,2,2|\mathbf{y}), \\ p(1,2,1|\mathbf{y}) &= p(2,1,2|\mathbf{y}), \qquad & p(1,1,2|\mathbf{y}) &= p(2,2,1|\mathbf{y}). \end{aligned}$$

The Marginal Posterior of a Single Allocation

When a finite mixture model is fitted to data with the aim of performing posterior clustering, one would hope to infer how likely the event $\{S_i = k\}$ is in light of the data. A natural candidate appears to be the posterior probability $\Pr(S_i = k|\mathbf{y})$. Somewhat surprisingly, it turns out that this marginal posterior probability is state-independent and equal to $1/K$, regardless of the data:

$$\Pr(S_i = k|\mathbf{y}) = \frac{1}{K}. \tag{3.27}$$

This follows from (3.26), by integrating both sides with respect to the indicators $(S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_N)$, which yields that the seemingly component-specific posterior probability $\Pr(S_i = k|\mathbf{y})$ is actually state invariant:

$$\Pr(S_i = k|\mathbf{y}) = \Pr(\rho_s(S_i) = k|\mathbf{y}) = \Pr(S_i = \rho_s^{-1}(k)|\mathbf{y}).$$

As this holds for all permutations, (3.27) follows immediately.

3.4 Classification Without Parameter Estimation

One of the most the challenging inference problems in finite mixture modeling, commonly known as the clustering problem, is classifying observations from a mixture distribution into K groups without knowing the component parameters. This interesting issue is studied in detail in Section 7.1; some aspects, however, are addressed at this point because they provide a good opportunity to introduce two important MCMC technique, namely Gibbs sampling and the Metropolis–Hastings algorithm, which are of relevance not only for classification, but also for Bayesian parameter estimation.

Bayesian clustering without parameter estimation is based on the marginal posterior distribution $p(\mathbf{S}|\mathbf{y})$ of the hidden allocation vector $\mathbf{S}$, where the mixture parameter $\boldsymbol{\vartheta}$ is integrated out, which is known up to a normalizing constant explicitly for mixtures from the exponential family; see again Subsection 3.3.3. $p(\mathbf{S}|\mathbf{y})$ is a discrete distribution over the lattice $\mathcal{S}_K$, defined in (3.22), which increases rapidly with the number of observations and the number of components. For $N = 10$ and $K = 3$, for instance, there are 59,049 different allocations $\mathbf{S}$, whereas for $N = 100$ and $K = 3$ the number of different allocations is of the order $5 \cdot 10^{47}$. For a very small data set from a mixture with very few components it would be possible to determine $p(\mathbf{S}|\mathbf{y})$ for all K^N possible allocations, and to find the allocation with the highest posterior probability $p(\mathbf{S}|\mathbf{y})$. With increasing sample size and increasing number of components, however, this is infeasible, and some search strategy has to be implemented to find an optimal allocation. Exploring the space $\mathcal{S}_K$, however, is in general quite a challenge.

Common search strategies that are applied in a Bayesian context are based on sampling allocations $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(M)}$ from the marginal posterior distribution $p(\mathbf{S}|\mathbf{y})$, which are then used for further inference, as explained in Subsection 7.1.7. Direct sampling of $\mathbf{S}$ from $p(\mathbf{S}|\mathbf{y})$ is not simple, as unconditionally the allocations $S_1, \ldots, S_N$ are correlated. Chen and Liu (1996) showed how sampling of the allocation through Markov chain Monte Carlo methods is feasible. An MCMC sampler starts from some preliminary classification $\mathbf{S}^{(0)}$. During sweep m, $m \geq 1$, of the MCMC sampler, the allocation S_i of each observation $\mathbf{y}_i$ is resampled in an appropriate manner, and the updated allocations are then stored as $\mathbf{S}^{(m)}$. Two common methods to implement an MCMC sampler are single-move Gibbs sampling and the Metropolis–Hastings algorithm. Both methods are described in Subsection 3.4.1 and Subsection 3.4.2, respectively.

For a detailed account we refer to the relevant literature on Markov chain Monte Carlo methods, in particular Gamerman (1997), Liu (2001), and Robert and Casella (1999).

3.4.1 Single-Move Gibbs Sampling

In this subsection we briefly introduce Gibbs sampling in the context of classification without parameter estimation, following Chen and Liu (1996) who used single-move Gibbs sampling to sample allocations $\mathbf{S}$ from the posterior distribution $p(\mathbf{S}|\mathbf{y})$ given in Subsection 3.3.3.

The single-move Gibbs sampler starts from some preliminary classification $\mathbf{S}^{(0)}$. Within each sweep $m, m \geq 1$, of the Gibbs sampler, the old allocations $\mathbf{S} = \mathbf{S}^{(m-1)}$ are updated for each observation $\mathbf{y}_i$, for $i = 1, \ldots, N$. Starting with $i = 1$, a new classification S_i^{new} is sampled, while holding the classifications $\mathbf{S}_{-i} = (S_1^{new}, \ldots, S_{i-1}^{new}, S_{i+1}, \ldots, S_N)$ of all other observations fixed. As not only $\mathbf{y}$, but also $\mathbf{S}_{-i}$ are assumed to be known, the appropriate posterior distribution for sampling S_i^{new} is the conditional posterior distribution $p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})$. Well-known properties of conditional distributions yield:

$$\begin{aligned} p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y}) &= \frac{p(S_i^{new}, \mathbf{S}_{-i}|\mathbf{y})}{p(\mathbf{S}_{-i}|\mathbf{y})} \propto p(\mathbf{y}|S_i^{new}, \mathbf{S}_{-i}) p(S_i^{new}, \mathbf{S}_{-i}) \\ &\propto p(\mathbf{y}_i|S_i^{new}, \mathbf{S}_{-i}) p(S_i^{new}|\mathbf{S}_{-i}), \end{aligned}$$

where constants independent of S_i^{new} were dropped. This is a univariate discrete density with K categories, which is easily sampled. Once S_i^{new} has been simulated, the Gibbs sampler proceeds with sampling the next indicator S_i^{new} after increasing i by 1, until $i = N$. Then the new allocations are stored as $\mathbf{S}^{(m)} = (S_1^{new}, \ldots, S_N^{new})$, m is increased by 1, and the whole procedure is repeated.

This sampling algorithm generates a sequence $\mathbf{S}^{(m)}, m = 1, 2, \ldots$ of classifications, which are obviously a Markov chain, as the distribution of $\mathbf{S}^{(m)}$ depends on $\mathbf{S}^{(m-1)}$, only:

$$p(\mathbf{S}^{(m)}|\mathbf{S}^{(m-1)}, \mathbf{y}) = \prod_{i=1}^{N} p(S_i^{(m)}|\mathbf{S}_{1:i-1}^{(m)}, \mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y}),$$

where $\mathbf{S}_{i:j}$ denotes the whole sequence $S_i, S_{i+1}, \ldots, S_j$. Well-known results from Markov chain theory guarantee that in the long run, as $m \to \infty$, the distribution of $\mathbf{S}^{(m)}$ converges to a stationary distribution, which could be shown to be equal to the desired marginal posterior $p(\mathbf{S}|\mathbf{y})$. When starting from an arbitrary allocation, the Markov chain will not be in equilibrium at the beginning, but will reach the stationary distribution after a suitable burn-in phase. Thus the first M_0 simulations are discarded before the simulated allocations may be used for posterior inference.

Algorithm 3.1: Single-Move Gibbs Sampling of the Allocations Start with some classification $\mathbf{S}$ and repeat the following steps for $m = 1, \ldots, M_0, \ldots, M + M_0$.

(a) Choose a certain observation $\mathbf{y}_i$, $i \in \{1, \ldots, N\}$, hold the most recent allocation of all observations but $\mathbf{y}_i$ fixed, and let $\mathbf{S}_{-i}$ be the sequence containing these allocations.

(b) Find a new allocation S_i^{new} for the observation $\mathbf{y}_i$ in the following way. Determine the univariate discrete distribution

$$p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y}) \propto p(\mathbf{y}_i|S_i^{new}, \mathbf{S}_{-i})p(S_i^{new}|\mathbf{S}_{-i}), \tag{3.28}$$

for all possible values $S_i^{new} = 1, \ldots, K$. Sample S_i^{new} from this distribution, and substitute the old allocation S_i by the new allocation S_i^{new}.

Repeat these steps until the allocations of all observations are updated. Store the actual values of all allocations as $\mathbf{S}^{(m)}$, increase m by one, and return to step (a).

Assume the current allocation of $\mathbf{y}_i$ is equal to $k : S_i = k$. Before sampling S_i^{new} from the posterior given in (3.28), the likelihood $p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})$, given by (3.25), and the prior $p(S_i^{new}|\mathbf{S}_{-i})$, given by (3.24), have to be evaluated for all values $S_i^{new} = l$, $l = 1, \ldots, K$. This is straightforward for $S_i^{new} = S_i = k$. Whenever the allocation changes (i.e., $S_i^{new} = l$ with $l \neq k$), the number of observations attached to component k and l need to be updated before applying (3.24):

$$N_k(S_i^{new}, \mathbf{S}_{-i}) = N_k(\mathbf{S}) - 1, \qquad N_l(S_i^{new}, \mathbf{S}_{-i}) = N_l(\mathbf{S}) + 1.$$

In a similar way, the statistics of the complete-data likelihood have to be updated before evaluating the likelihood $p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})$ from (3.25) for $S_i^{new} = l$, where $l \neq k$. For mixtures of Poisson distributions, for instance, this reads:

$$\begin{aligned} b_k(S_i^{new}, \mathbf{S}_{-i}) &= b_k(\mathbf{S}) - 1, \qquad & b_l(S_i^{new}, \mathbf{S}_{-i}) &= b_l(\mathbf{S}) + 1, \\ a_k(S_i^{new}, \mathbf{S}_{-i}) &= a_k(\mathbf{S}) - y_i, \qquad & a_l(S_i^{new}, \mathbf{S}_{-i}) &= a_l(\mathbf{S}) + y_i. \end{aligned}$$

Similar simple updates are available for many other standard finite mixture models. For various other more complex mixture models, such as mixtures of regression models, Chen and Liu (1996) developed an efficient algorithm to compute the likelihood $p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})$ recursively from $p(\mathbf{y}|\mathbf{S}_{-i}, S_i)$.

Why Single-Move Gibbs Sampling Works

It is instructive to verify that single-move Gibbs sampling works, by showing that sampling $\mathbf{S}^{(m)}$ from $p(\mathbf{S}^{(m)}|\mathbf{S}^{(m-1)})$ yields a sample from $p(\mathbf{S}|\mathbf{y})$, once the chain reaches equilibrium, and $\mathbf{S}^{(m-1)}$ is drawn from $p(\mathbf{S}|\mathbf{y})$. Let $f(\mathbf{S}^{(m)})$ denote the density of the distribution of $\mathbf{S}^{(m)}$, which is given by

$$f(\mathbf{S}^{(m)}) = \sum_{\mathbf{S}^{(m-1)} \in \mathcal{S}_K} p(\mathbf{S}^{(m)}|\mathbf{S}^{(m-1)}, \mathbf{y})p(\mathbf{S}^{(m-1)}|\mathbf{y}) =$$

$$= \sum_{S_N^{(m-1)}=1}^{K} \cdots \sum_{S_2^{(m-1)}=1}^{K} \prod_{i=1}^{N} p(S_i^{(m)}|\mathbf{S}_{1:i-1}^{(m)}, \mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y}) \prod_{i=2}^{N} p(S_i^{(m-1)}|\mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y})$$

$$\cdot \left(\sum_{S_1^{(m-1)}=1}^{K} p(S_1^{(m-1)}|\mathbf{S}_{2:N}^{(m-1)}, \mathbf{y}) \right),$$

where the innermost term is obviously equal to 1. Therefore

$$f(\mathbf{S}^{(m)}) = \sum_{S_N^{(m-1)}=1}^{K} \cdots \sum_{S_3^{(m-1)}=1}^{K} \prod_{i=2}^{N} p(S_i^{(m)}|\mathbf{S}_{1:i-1}^{(m)}, \mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y})$$
$$\cdot \prod_{i=3}^{N} p(S_i^{(m-1)}|\mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y}) \left(\sum_{S_2^{(m-1)}=1}^{K} p(S_2^{(m-1)}|\mathbf{S}_{3:N}^{(m-1)}, \mathbf{y}) p(S_1^{(m)}|\mathbf{S}_{2:N}^{(m-1)}, \mathbf{y}) \right).$$

The innermost term is equal to $p(S_1^{(m)}|\mathbf{S}_{3:N}^{(m-1)}, \mathbf{y})$, therefore

$$f(\mathbf{S}^{(m)}) = \sum_{S_N^{(m-1)}=1}^{K} \cdots \sum_{S_4^{(m-1)}=1}^{K} \prod_{i=3}^{N} p(S_i^{(m)}|\mathbf{S}_{1:i-1}^{(m)}, \mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y})$$
$$\cdot \prod_{i=4}^{N} p(S_i^{(m-1)}|\mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y}) \left(\sum_{S_3^{(m-1)}=1}^{K} p(S_3^{(m-1)}|\mathbf{S}_{4:N}^{(m-1)}, \mathbf{y}) p(\mathbf{S}_{1:2}^{(m)}|\mathbf{S}_{3:N}^{(m-1)}, \mathbf{y}) \right).$$

The innermost term is equal to $p(\mathbf{S}_{1:2}^{(m)}|\mathbf{S}_{4:N}^{(m-1)}, \mathbf{y})$, therefore:

$$f(\mathbf{S}^{(m)}) = \sum_{S_N^{(m-1)}=1}^{K} \cdots \sum_{S_5^{(m-1)}=1}^{K} \prod_{i=4}^{N} p(S_i^{(m)}|\mathbf{S}_{1:i-1}^{(m)}, \mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y})$$
$$\cdot \prod_{i=5}^{N} p(S_i^{(m-1)}|\mathbf{S}_{i+1:N}^{(m-1)}, \mathbf{y}) \cdot \left(\sum_{S_4^{(m-1)}=1}^{K} p(S_4^{(m-1)}|\mathbf{S}_{5:N}^{(m-1)}, \mathbf{y}) p(\mathbf{S}_{1:3}^{(m)}|\mathbf{S}_{4:N}^{(m-1)}, \mathbf{y}) \right),$$

where the innermost term is equal to $p(\mathbf{S}_{1:3}^{(m)}|\mathbf{S}_{5:N}^{(m-1)}, \mathbf{y})$. This is repeated until we obtain:

$$f(\mathbf{S}^{(m)}) = \sum_{S_N^{(m-1)}=1}^{K} p(S_{N-1}^{(m)}|\mathbf{S}_{1:N-2}^{(m)}, S_N^{(m-1)}, \mathbf{y}) p(S_N^{(m)}|\mathbf{S}_{1:N-1}^{(m)}, \mathbf{y}) p(S_N^{(m-1)}|\mathbf{y})$$
$$\cdot \left(\sum_{S_{N-1}^{(m-1)}=1}^{K} p(S_{N-1}^{(m-1)}|S_N^{(m-1)}, \mathbf{y}) p(\mathbf{S}_{1:N-2}^{(m)}|\mathbf{S}_{N-1:N}^{(m-1)}, \mathbf{y}) \right),$$

which yields the desired result:

$$f(\mathbf{S}^{(m)}) = p(S_N^{(m)}|\mathbf{S}_{1:N-1}^{(m)}, \mathbf{y}) \sum_{S_N^{(m-1)}=1}^{K} p(\mathbf{S}_{1:N-1}^{(m)}|S_N^{(m-1)}, \mathbf{y}) p(S_N^{(m-1)}|\mathbf{y})$$
$$= p(\mathbf{S}^{(m)}|\mathbf{y}).$$

3.4.2 The Metropolis–Hastings Algorithm

Alternatively to the Gibbs sampler described in *Algorithm 3.1*, the Metropolis–Hastings algorithm may be applied to draw from the density $p(\mathbf{S}|\mathbf{y})$. Running a Gibbs sampler may be impractical if K is large, as in (3.28) the probability $p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})$ needs to be evaluated for all $S_i^{new} = 1, \ldots, K$.

Whereas the Gibbs sampler used the density $p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})$ for proposing S_i^{new}, the Metropolis–Hastings algorithm uses an arbitrary discrete density $q(S_i^{new}|S_i)$, where S_i is the current allocation, to propose S_i^{new}. Without modifications, the resulting Markov chain $\mathbf{S}^{(m)}$ would not draw from the desired posterior distribution $p(\mathbf{S}|\mathbf{y})$. To obtain draws from the desired distribution, the proposed allocation S_i^{new} is not accepted in any case, but only with a certain probability $\alpha(S_i^{new}|S_i)$. If the new value is accepted, then $S_i^{(m)} = S_i^{new}$, otherwise S_i^{new} is rejected and the chain does not move: $S_i^{(m)} = S_i$.

As pointed out by Chib and Greenberg (1995), the accept–reject step is necessary as $q(S_i^{new}|S_i)$ is not likely to fulfill the detailed balance condition. For instance, it may happen that

$$p(S_i|\mathbf{S}_{-i}, \mathbf{y})q(S_i^{new}|S_i) > p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})q(S_i|S_i^{new}), \tag{3.29}$$

meaning that too many moves from S_i to S_i^{new}, and too few moves from S_i^{new} to S_i are made. The probability $\alpha(S_i^{new}|S_i)$ of accepting a move from S_i to S_i^{new} is introduced, in order to ensure detailed balance. The acceptance probability $\alpha(S_i^{new}|S_i)$ is chosen precisely to ensure that the Markov chain $S_i^{(m)}$ is reversible with respect to $p(S_i|\mathbf{S}_{-i}, \mathbf{y})$. Following Chib and Greenberg (1995), $\alpha(S_i|S_i^{new})$ should be set to 1, if (3.29) holds, as moves from S_i^{new} to S_i are too rare. The reverse probability $\alpha(S_i^{new}|S_i)$ is then determined by forcing a detailed balance in (3.29),

$$p(S_i|\mathbf{S}_{-i}, \mathbf{y})q(S_i^{new}|S_i)\alpha(S_i^{new}|S_i) = p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})q(S_i|S_i^{new}). \tag{3.30}$$

Thus $\alpha(S_i^{new}|S_i)$ which could not be larger than 1, is given by

$$\alpha(S_i^{new}|S_i) = \min\left(1, \frac{p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})q(S_i|S_i^{new})}{p(S_i|\mathbf{S}_{-i}, \mathbf{y})q(S_i^{new}|S_i)}\right), \tag{3.31}$$

if $p(S_i|\mathbf{S}_{-i}, \mathbf{y})q(S_i^{new}|S_i) > 0$. Interestingly, other acceptance rules are possible (see Liu, 2001, Section 5), however, Peskun (1973) proves superiority of (3.31) in terms of statistical efficiency.

Algorithm 3.2: Sampling the Allocations Through a Metropolis–Hastings Algorithm Start with some classification $\mathbf{S}$ and repeat the following steps for $m = 1, \ldots, M_0, \ldots, M + M_0$.

(a) Choose a certain observation $\mathbf{y}_i$, $i \in \{1, \ldots, N\}$, hold the most recent allocations of all observations but $\mathbf{y}_i$ fixed, and let $\mathbf{S}_{-i}$ be the sequence containing these allocations.

(b) Find a new allocation S_i^{new} for the observation $\mathbf{y}_i$ in the following way. Sample S_i^{new} from a proposal density $q(S_i^{new}|S_i)$ and substitute the old allocation S_i by the new allocation S_i^{new} with probability $\min(1, r_i)$, where

$$r_i = \frac{p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})p(S_i^{new}|\mathbf{S}_{-i})q(S_i|S_i^{new})}{p(\mathbf{y}|\mathbf{S}_{-i}, S_i)p(S_i|\mathbf{S}_{-i})q(S_i^{new}|S_i)}. \tag{3.32}$$

If $U_i < \min(1, r_i)$, where U_i is random number from the $\mathcal{U}[0, 1]$-distribution, then S_i is substituted by S_i^{new}, otherwise leave S_i unchanged.

Repeat these steps until the allocations of all observations are updated. Store the actual values of all allocations as $\mathbf{S}^{(m)}$, increase m by one, and return to step (a).

If $q(S_i^{new}|S_i) = p(S_i^{new}|\mathbf{S}_{-i}, \mathbf{y})$, then $r_i = 1$, and the Metropolis–Hastings algorithm reduces to the Gibbs sampler described in *Algorithm 3.1.* To avoid the functional evaluations that are necessary to sample from this specific proposal density, much simpler proposal densities are used for the Metropolis–Hastings algorithm.

Some simplifications are possible when evaluating r_i. If $S_i^{new} = S_i$, the likelihood and the prior cancel, and r_i is equal to the proposal ratio. If $S_i^{new} = l$ while $S_i = k$ with $k \neq l$, the acceptance ratio r_i simplifies to

$$r_i = \frac{p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})(N_l(\mathbf{S}) + 1 + e_{0,l})q(S_i|S_i^{new})}{p(\mathbf{y}|\mathbf{S}_{-i}, S_i)(N_k(\mathbf{S}) + e_{0,k})q(S_i^{new}|S_i)},$$

where $N_k(\mathbf{S})$ and $N_l(\mathbf{S})$ are the current numbers of allocations. For mixtures of Poisson distributions the likelihood ratio reduces to:

$$\frac{p(\mathbf{y}|\mathbf{S}_{-i}, S_i^{new})}{p(\mathbf{y}|\mathbf{S}_{-i}, S_i)} = \frac{\Gamma(a_k(\mathbf{S}) - y_i)\Gamma(a_l(\mathbf{S}) + y_i)b_k(\mathbf{S})^{a_k(\mathbf{S})}b_l(\mathbf{S})^{a_l(\mathbf{S})}}{\Gamma(a_k(\mathbf{S}))\Gamma(a_l(\mathbf{S}))(b_k(\mathbf{S}) - 1)^{a_k(\mathbf{S})-y_i}(b_l(\mathbf{S}) + 1)^{a_l(\mathbf{S})+y_i}}.$$

3.5 Parameter Estimation Through Data Augmentation and MCMC

Markov chain Monte Carlo sampling is not only useful for the purpose of sampling allocations, but also for parameter estimation.

3.5.1 Treating Mixture Models as a Missing Data Problem

As already discussed in Subsection 2.3.3, for mixture models from exponential families such as mixtures of Poisson distributions or mixtures of normal distributions, a conjugate analysis is feasible for the complete-data likelihood function (2.8) when the allocations $\mathbf{S} = (S_1, \ldots, S_N)$ are observed. For unknown allocations, however, this is not the case.

Following the seminal paper by Dempster et al. (1977), a mixture model may be seen as an incomplete data problem by introducing the allocations $\mathbf{S}$ as missing data. The benefit of this data augmentation (Tanner and Wong, 1987) is that conditional on $\mathbf{S}$ we are back in the conjugate setting of complete-data Bayesian estimation considered in Subsection 2.3.3. On the other hand, conditional on knowing the parameter $\boldsymbol{\vartheta}$, we are back to the classification problem studied in Section 2.2, where the posterior distribution of the allocations takes a very simple form. It is then rather straightforward to sample from the posterior (3.2) using Markov chain Monte Carlo methods, in particular Gibbs sampling. Early papers realizing the importance of Gibbs sampling for Bayesian estimation of mixture models are Evans et al. (1992), West (1992), Smith and Roberts (1993), Diebolt and Robert (1994), Escobar and West (1995), Mengersen and Robert (1996), and Raftery (1996b). We first give specific results for a mixture of Poisson distributions in Subsection 3.5.2 and then proceed with a discussion for more general finite mixture models in Subsection 3.5.3.

3.5.2 Data Augmentation and MCMC for a Mixture of Poisson Distributions

For N observations $\mathbf{y} = (y_1, \ldots, y_N)$, assumed to arise from a finite mixture of K Poisson distributions, the mixture likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$ is given by

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \eta_k f_P(y_i; \mu_k) \right), \tag{3.33}$$

where $f_P(y_i; \mu_k)$ is the density of a Poisson distribution with mean μ_k. Although direct sampling from (3.33) is not easy, a straightforward method of sampling from (3.33) based on data augmentation is possible.

For each observation y_i, $i = 1, \ldots, N$, the group indicator S_i taking a value in $\{1, \ldots, K\}$ is introduced as a missing observation. Conditional on knowing the group indicator S_i, the observation model for observation y_i is a Poisson distribution with mean μ_{S_i}:

$$y_i|\mu_1, \ldots, \mu_K, S_i \sim \mathcal{P}(\mu_{S_i}). \tag{3.34}$$

All observations with the same group indicator S_i equal to k, say, arise from the same $\mathcal{P}(\mu_k)$-distribution. Therefore the complete-data likelihood $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta})$, which has been defined in (2.8), reads:

$$p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}) = \prod_{k=1}^{K} \left(\prod_{i:S_i=k} f_P(y_i; \mu_k) \right) \left(\prod_{k=1}^{K} \eta_k^{N_k(\mathbf{S})} \right),$$

where $N_k(\mathbf{S}) = \#\{S_i = k\}$. $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta})$, considered as a function of $\boldsymbol{\vartheta}$, is the product of $K + 1$ independent factors. Each of the first K factors depends

only on μ_k, whereas the last factor depends on $\boldsymbol{\eta}$. Assuming independence a priori, the parameters $\mu_1, \ldots, \mu_K, \boldsymbol{\eta}$ are independent a posteriori given the complete data $(\mathbf{y}, \mathbf{S})$:

$$p(\mu_1, \ldots, \mu_K, \boldsymbol{\eta}|\mathbf{S}, \mathbf{y}) = \prod_{k=1}^{K} p(\mu_k|\mathbf{S}, \mathbf{y}) p(\boldsymbol{\eta}|\mathbf{S}).$$

Each of the conditional posteriors can be handled within the conjugate setting discussed in Subsection 2.3.3. We express prior knowledge about μ_k as a $\mathcal{G}(a_0, b_0)$-distribution. Then from Bayes' theorem:

$$p(\mu_k|\mathbf{S}, \mathbf{y}) \propto \left(\prod_{i:S_i=k} f_P(y_i; \mu_k) \right) p(\mu_k). \tag{3.35}$$

The posterior distribution $p(\mu_k|\mathbf{S}, \mathbf{y})$ is a $\mathcal{G}(a_k(\mathbf{S}), b_k(\mathbf{S}))$-distribution, where

$$a_k(\mathbf{S}) = a_0 + N_k(\mathbf{S})\overline{y}_k(\mathbf{S}), \qquad b_k(\mathbf{S}) = b_0 + N_k(\mathbf{S}), \tag{3.36}$$

and $N_k(\mathbf{S}) = \#\{S_i = k\}$ and $\overline{y}_k(\mathbf{S})$ are the number of observations and the mean in group k.

Based on assuming a Dirichlet $\mathcal{D}(e_0, \ldots, e_0)$-distribution for $\boldsymbol{\eta}$, the posterior distribution of the weight distribution $\boldsymbol{\eta}$ given $\mathbf{S}$ is a $\mathcal{D}(e_1(\mathbf{S}), \ldots, e_K(\mathbf{S}))$-distribution, where

$$e_k(\mathbf{S}) = e_0 + N_k(\mathbf{S}), \qquad k = 1, \ldots, K. \tag{3.37}$$

MCMC Estimation Using Gibbs Sampling

MCMC estimation of a mixture of Poisson distributions under fixed hyperparameters a_0 and b_0 consists of the following steps.

Algorithm 3.3: Gibbs Sampling for a Poisson Mixture Start with some classification $\mathbf{S}^{(0)}$ and repeat the following steps for $m = 1, \ldots, M_0, \ldots, M + M_0$.

(a) Parameter simulation conditional on the classification $\mathbf{S}^{(m-1)}$:
   (a1) Sample $\eta_1, \ldots, \eta_K$ from a $\mathcal{D}\left(e_1(\mathbf{S}^{(m-1)}), \ldots, e_K(\mathbf{S}^{(m-1)})\right)$-distribution, where $e_k(\mathbf{S}^{(m-1)})$ is given by (3.37).
   (a2) For each $k = 1, \ldots, K$, sample μ_k from a $\mathcal{G}\left(a_k(\mathbf{S}^{(m-1)}), b_k(\mathbf{S}^{(m-1)})\right)$-distribution, where $a_k(\mathbf{S}^{(m-1)})$ and $b_k(\mathbf{S}^{(m-1)})$ are given by (3.36).

   Store the actual values of all parameters as $\boldsymbol{\vartheta}^{(m)} = (\mu_1^{(m)}, \ldots, \mu_K^{(m)}, \boldsymbol{\eta}^{(m)})$.

(b) Classification of each observation y_i conditional on knowing $\boldsymbol{\vartheta}^{(m)}$: sample S_i independently for each $i = 1, \ldots, N$ from the conditional posterior distribution $p(S_i|\boldsymbol{\vartheta}^{(m)}, y_i)$, which by the results of Subsection 2.2.1 is given by

$$p(S_i = k|\boldsymbol{\vartheta}^{(m)}, y_i) \propto (\mu_k^{(m)})^{y_i} \mathrm{e}^{-\mu_k^{(m)}} \eta_k^{(m)}.$$

Store the actual values of all allocations as $\mathbf{S}^{(m)}$, increase m by one, and return to step (a). Finally, the first M_0 draws are discarded.

Hierarchical Priors

Under the hierarchical prior (3.12), an additional block has to be added in *Algorithm 3.3*, where b_0 is sampled from the conditional posterior distribution $p(b_0|\mu_1, \ldots, \mu_K, \mathbf{S}, \mathbf{y})$, given by Bayes' theorem:

$$p(b_0|\mu_1, \ldots, \mu_K, \mathbf{S}, \mathbf{y}) \propto \prod_{k=1}^{K} p(\mu_k|b_0)p(b_0) \tag{3.38}$$

$$\propto \prod_{k=1}^{K} b_0^{a_0} \exp(-\mu_k b_0)\, p(b_0) \propto b_0^{g_0+Ka_0-1} \exp\left(-(G_0 + \sum_{k=1}^{K} \mu_k)b_0\right).$$

Under a conjugate $\mathcal{G}(g_0, G_0)$-prior for b_0, this posterior is a Gamma distribution, depending on the data only indirectly through the component means.

Gibbs sampling requires the following modification of *Algorithm 3.3*. Select a starting value ${b_0}^{(0)}$ and run step (a2) conditional on ${b_0}^{(m-1)}$. A third step is added to sample the hyperparameter ${b_0}^{(m)}$:

(c) Sample $b_0^{(m)}$ from $p(b_0|\mu_1^{(m)}, \ldots, \mu_K^{(m)})$ given by (3.38):

$$b_0|\mu_1^{(m)}, \ldots, \mu_K^{(m)} \sim \mathcal{G}\left(g_0 + Ka_0, G_0 + \sum_{k=1}^{K} \mu_k^{(m)}\right). \tag{3.39}$$

3.5.3 Data Augmentation and MCMC for General Mixtures

As for the Poisson mixture, Bayesian estimation of a general mixture model through data augmentation estimates the augmented parameter $(\mathbf{S}, \boldsymbol{\vartheta})$ by sampling from the complete-data posterior distribution $p(\mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y})$. This posterior is given by Bayes' theorem,

$$p(\mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}), \tag{3.40}$$

thus the complete-data posterior is proportional to the complete-data likelihood likelihood $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta})$ defined in (2.8) times the prior $p(\boldsymbol{\vartheta})$ on $\boldsymbol{\vartheta}$; see again Subsection 2.3.3 for more details. Sampling from the posterior (3.40) is most commonly carried out by the following MCMC sampling scheme, where $\boldsymbol{\vartheta}$ is sampled conditional on knowing $\mathbf{S}$, and $\mathbf{S}$ is sampled conditional on knowing $\boldsymbol{\vartheta}$. This scheme is formulated for the general case, where the observations $\mathbf{y}_i$ may be multivariate.

Algorithm 3.4: Unconstrained MCMC for a Mixture Model Start with some classification $\mathbf{S}^{(0)}$ and repeat the following steps for $m = 1, \ldots, M_0, \ldots, M + M_0$.

(a) Parameter simulation conditional on the classification $\mathbf{S}^{(m-1)}$:

(a1) Sample $\boldsymbol{\eta}$ from the $\mathcal{D}\left(e_1(\mathbf{S}^{(m-1)}), \ldots, e_K(\mathbf{S}^{(m-1)})\right)$-distribution, where $e_k(\mathbf{S}^{(m-1)})$ is given by (3.37).

(a2) Sample the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ from the complete-data posterior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \mathbf{S}^{(m-1)}, \mathbf{y})$.

Store the actual values of all parameters as $\boldsymbol{\vartheta}^{(m)} = (\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}, \boldsymbol{\eta}^{(m)})$.

(b) Classification of each observation $\mathbf{y}_i$ conditional on knowing $\boldsymbol{\vartheta}^{(m)}$: sample S_i independently for each $i = 1, \ldots, N$ from the conditional posterior distribution $p(S_i | \boldsymbol{\vartheta}^{(m)}, \mathbf{y}_i)$, which by the results of Subsection 2.2.1 is given by

$$p(S_i = k | \boldsymbol{\vartheta}^{(m)}, \mathbf{y}_i) \propto p(\mathbf{y}_i | \boldsymbol{\theta}_k^{(m)}) \eta_k^{(m)}. \tag{3.41}$$

Store the actual values of all allocations as $\mathbf{S}^{(m)}$, increase m by one, and return to step (a). Finally, the first M_0 draws are discarded.

The structure of the posterior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \mathbf{S}, \mathbf{y})$ depends on the specific distribution families appearing in the components of the mixture model and on the chosen priors. If the components come from an exponential family, the results of Subsection 3.2.3 will be helpful. Under the conditionally conjugate prior (3.9), the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are independent given $\mathbf{S}$ and may be sampled from the conditional posterior $p(\boldsymbol{\theta}_k | \mathbf{S}, \mathbf{y})$ given by (3.10) for each $k = 1, \ldots, K$.

The MCMC sampler described in *Algorithm 3.4* starts with sampling the parameter $\boldsymbol{\vartheta}$ based on allocations $\mathbf{S}^{(0)}$ defined by the investigator. Theoretically, it does not make any difference if the sampling steps (a) and (b) are interchanged, in which case the algorithm starts with sampling the allocations $\mathbf{S}$ based on a parameter $\boldsymbol{\vartheta}^{(0)}$. Practical MCMC convergence diagnostics for finite mixture models is considered by Robert et al. (1999).

Hierarchical Priors

Under the hierarchical prior discussed in Subsection 3.2.4, an additional block has to be added in *Algorithm 3.4* to sample the hyperparameter $\boldsymbol{\delta}$ conditional on knowing $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ from

$$p(\boldsymbol{\delta} | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \propto \prod_{k=1}^{K} p(\boldsymbol{\theta}_k | \boldsymbol{\delta}) p(\boldsymbol{\delta}). \tag{3.42}$$

In many cases, this density will be of closed form. This leads to the following modification of *Algorithm 3.4*. Select a starting value $\boldsymbol{\delta}^{(0)}$ and run step (a2) conditional on $\boldsymbol{\delta}^{(m-1)}$. A third step is added to sample the hyperparameter $\boldsymbol{\delta}^{(m)}$ from $p(\boldsymbol{\delta} | \boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)})$, given by (3.42).

3.5.4 MCMC Sampling Under Improper Priors

MCMC sampling under improper priors is possible as long as the conditional posterior $p(\boldsymbol{\vartheta}|\mathbf{S},\mathbf{y})$ is proper for all possible allocations. What happens, if unintentionally MCMC sampling is carried out under a prior like the improper product prior (3.4), where the posterior is improper (Natarajan and McCulloch, 1998)?

For a mixture of Poisson distributions, for instance, an improper product prior based on $\mu_k \sim \mathcal{G}(0,0)$ or $\mu_k \sim \mathcal{G}(0.5,0)$ leads to an improper posterior distribution by the results of Subsection 3.2.2. If $N_k(\mathbf{S}^{(m-1)}) = 0$ for a certain draw, then the conditional posterior $p(\mu_k|\mathbf{S}^{(m-1)},\mathbf{y})$ given by (3.36) is equal to the improper prior and the MCMC sampler breaks down when drawing $\mu_k^{(m)}$, warning us that something is not in order.

In other cases, it is possible to obtain sensible looking results when running data augmentation and MCMC under the product prior (3.4). Consider, for instance, a synthetic data set of size $N = 500$, simulated from a mixture of two Poisson distributions, where $\mu_1 = 1$, $\mu_2 = 5$, and $\eta_1 = 0.4$. We estimated $(\mu_1,\mu_2,\eta_1,\eta_2)$ under the uniform $\mathcal{D}(1,1)$ prior on (η_1,η_2), with an improper $\mathcal{G}(0.5,0)$ as well as a proper $\mathcal{G}(0.01,0.01)$ prior on μ_1 and μ_2, running MCMC for 1 million iterations without problems. Furthermore, for both priors the resulting density estimates were indistinguishable. To understand this, consider the following representation of the posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$,

$$p(\boldsymbol{\vartheta}|\mathbf{y}) = \sum_{\mathbf{S}\in\mathcal{S}_K} p(\boldsymbol{\vartheta}|\mathbf{S},\mathbf{y})p(\mathbf{S}|\mathbf{y}),$$

where the complete-data posterior $p(\boldsymbol{\vartheta}|\mathbf{S},\mathbf{y})$ is weighted by the posterior probability of the corresponding partition $\mathbf{S}$. If partitions $\mathbf{S}$, where the corresponding complete-data posterior is improper, have very low posterior probability, then it is very unlikely (though possible) that such a classification is selected during MCMC sampling. Therefore the *estimated* posterior

$$\hat{p}(\boldsymbol{\vartheta}|\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\vartheta}|\mathbf{S}^{(m)},\mathbf{y})$$

will be proper. Nevertheless, is not recommended to sample from improper posterior distributions in this way, as statistical inference drawn from such a posterior distribution lacks any theoretical justification.

3.5.5 Label Switching

The term *label switching* has been introduced into the literature on mixture models by Redner and Walker (1984) to describe the invariance of the mixture likelihood function under relabeling the components of a mixture model described in Subsection 2.4.2. Label switching is of no concern for maximum

likelihood estimation, where the goal is to find one of the equivalent modes of the likelihood function. In the context of Bayesian estimation, however, label switching has to be addressed explicitly because in the course of sampling from the mixture posterior distribution, the labeling of the unobserved categories changes. Interestingly, the label switching problem was totally neglected in the early papers on MCMC estimation of finite mixture models and was addressed only later on by Celeux (1998), Celeux et al. (2000), Stephens (2000a, 2000b), Casella et al. (2000), and Frühwirth-Schnatter (2001b).

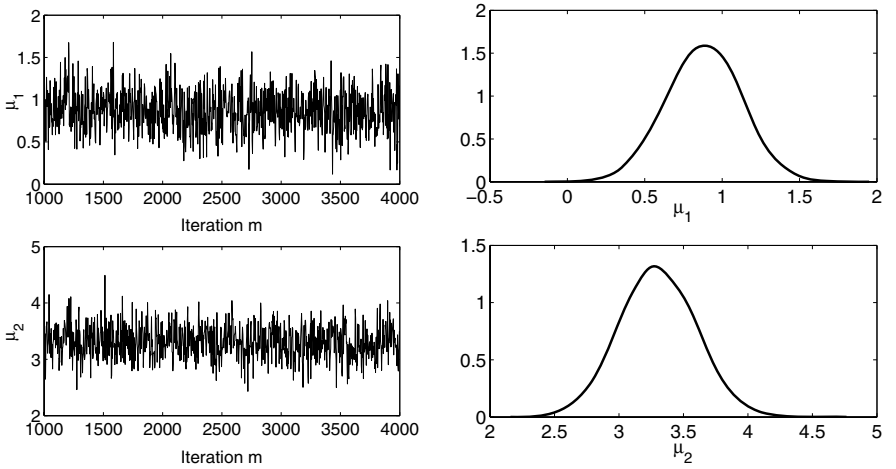

Fig. 3.1. HIDDEN AGE GROUPS — Synthetic Data Set 1; MCMC draws of μ_1 and μ_2 (left-hand side) and estimated marginal posterior densities of μ_1 and μ_2 (right-hand side)

Some Illustration

For illustration, we reconsider the example of Subsection 2.4.3, where we simulated artificial data sets of length $N = 100$ from the following mixture of normals,

$$p(y) = 0.5 f_N(y; \mu_y, 1) + 0.5 f_N(y; \mu_o, 1),$$

where μ_y and μ_o are the mean of a random variable Y in a younger and in an older subgroup in the population. For MCMC estimation of μ_1 and μ_2, we apply data augmentation as in Subsection 3.5.3 under the prior $p(\mu_k) \sim \mathcal{N}(0, 100)$. The details of step (a2) for the specific example of a mixture of normal distributions appear later in Subsection 6.2.4. The MCMC draws of μ_k as well as the estimated marginal densities $p(\mu_k|\mathbf{y})$ are plotted in Figure 3.1

and Figure 3.2 for two artificial data sets, where $\mu_y = 1$ and $\mu_o = 3$ for the first, and $\mu_y = 1$ and $\mu_o = 1.5$ for the second data set.

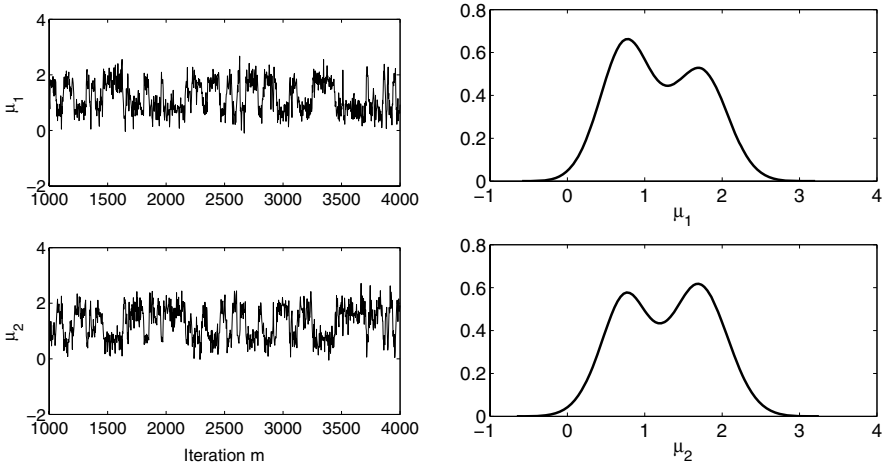

Fig. 3.2. HIDDEN AGE GROUPS — Synthetic Data Set 2; MCMC draws of μ_1 and μ_2 (left-hand side) and estimated marginal posterior densities of μ_1 and μ_2 (right-hand side)

For each data set we started at $\mu_1 = \mu_y$ and $\mu_2 = \mu_o$, which corresponds to labeling 1. For synthetic data set 1 the sampler stays within the modal region corresponding to this labeling, as this region is well separated from the region where the other labeling is valid; see again Figure 2.5. Note that the estimated marginal posterior densities in Figure 3.1 are unimodal and that Gibbs sampling leads implicitly to a unique labeling.

For synthetic data set 2, however, the marginal posterior densities are bimodal and the MCMC draws suffer from label switching. For this data set parameters around the nonidentifiability set $\mathcal{U}^E(\hat{\mu})$, where $\hat{\mu} = \overline{y}$, have considerable likelihood under both labelings; consider again Figure 2.6. Even if we start in the modal region corresponding to labeling 1, where $\mu_1 < \mu_2$, the sampler is likely to move into the area where $\mu_1 > \mu_2$. In this area, however, the parameter (μ_1, μ_2) has higher likelihood if μ_1 is associated with the older subgroup, rather than with the younger one. Therefore, when sampling the group indicators $\mathbf{S}$, there is a certain risk that the labeling changes and now μ_1 is associated with the older subgroup. After such a label switching takes place, the sampler remains in the second modal region for a while until it returns to the area where $\mu_1 < \mu_2$. Then there exists considerable likelihood that the sampler switches back to labeling 1. This occasional change of labeling is obvious from the MCMC draws in Figure 3.2.

3.5.6 Permutation MCMC Sampling

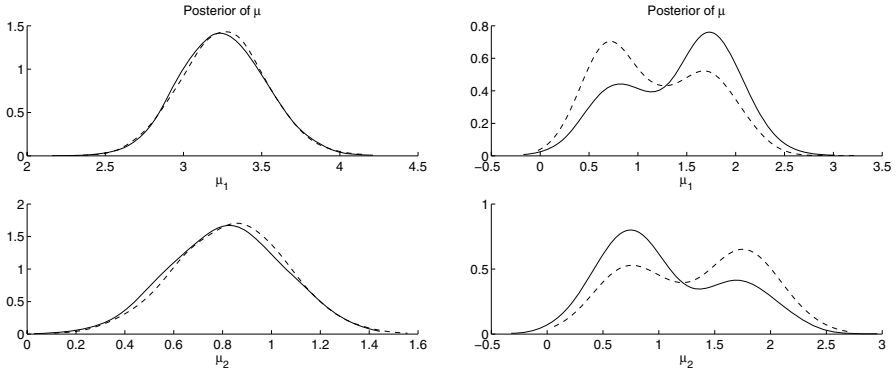

Fig. 3.3. HIDDEN AGE GROUPS — Synthetic Data Set 1 and 2; marginal posterior densities of μ_1 and μ_2 estimated from two different runs of Gibbs sampling under the same prior for Data Set 1 (left-hand side) and Data Set 2 (right-hand side)

The examples of the previous subsection demonstrated that the behavior of the Gibbs sampler described in *Algorithm 3.4* is somewhat unpredictable. For synthetic data set 1 it is trapped at one modal region, whereas it jumps from time to time to the other modal region for data set 2. In both cases the sampler did not explore the full mixture posterior distribution.

This matters especially when estimating marginal densities. Assume that we want to assess the influence of the prior $p(\boldsymbol{\vartheta})$ on the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$. To do so, we usually compare the marginal posterior densities $p(\boldsymbol{\theta}_k|\mathbf{y})$ obtained under different prior distributions $p(\boldsymbol{\theta}_k)$. There, the marginal density is estimated from the MCMC draws by some kernel smoothing method.

For a mixture model it turns out that estimating the marginal density from the MCMC draws may lead to a poor estimate when *unbalanced* label switching takes place. It may even happen that although we assume the *same* prior distribution $p(\boldsymbol{\theta}_k)$, the marginal posterior densities $p(\boldsymbol{\theta}_k|\mathbf{y})$ estimated from different runs of the MCMC sampler, are very different. For illustration, Figure 3.3 compares estimates of the marginal density obtained from two different runs of full conditional Gibbs sampling for $M = 2000$ under the same prior for the two synthetic data sets considered earlier. The estimated marginal densities are nearly identical for data set 1, where no label switching took place. We observe a substantial difference in these densities for data set 2, the reason being that sampler did not explore the whole mixture posterior distribution as label switching took place only from time to time in an unbalanced manner.

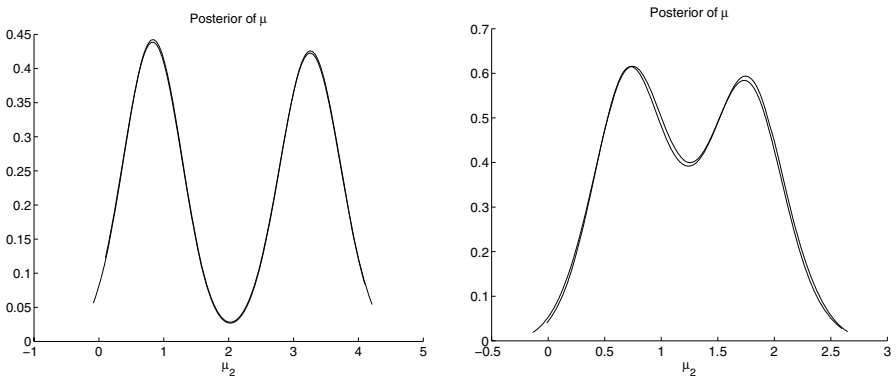

Fig. 3.4. HIDDEN AGE GROUPS — Synthetic Data Set 1 and 2; marginal posterior densities of μ_1 and μ_2 estimated from random permutation Gibbs sampling under the same prior for Data Set 1 (left-hand side) and Data Set 2 (right-hand side)

A simple, but efficient solution to obtain a sampler that explores the full mixture posterior distribution is to force balanced label switching by concluding each MCMC draw by a randomly selected permutation of the labeling. This method is called random permutation MCMC sampling (Frühwirth-Schnatter, 2001b).

Algorithm 3.5: Random Permutation MCMC Sampling for a Finite Mixture Model Start as described in *Algorithm 3.4*.

(a) and (b) are the same steps as in *Algorithm 3.4*.
(c) Conclude each draw by selecting randomly one of the $K!$ possible permutations $\rho_s(1), \ldots, \rho_s(K)$ of the current labeling. This permutation is applied to $\boldsymbol{\eta}^{(m)}$, the component parameters $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$, and the allocations $\mathbf{S}^{(m)}$:
  (c1) The group weights $\eta_1^{(m)}, \ldots, \eta_K^{(m)}$ are substituted by $\eta_{\rho_s(1)}^{(m)}, \ldots, \eta_{\rho_s(K)}^{(m)}$.
  (c2) The component parameters $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ are substituted by $\boldsymbol{\theta}_{\rho_s(1)}^{(m)}$, $\ldots$, $\boldsymbol{\theta}_{\rho_s(K)}^{(m)}$.
  (c3) The allocations $S_i^{(m)}, i = 1, \ldots, N$, are substituted by $\rho_s(S_i^{(m)}), i = 1, \ldots, N$.

For illustration we consider once more the synthetic data sets 1 and 2. Let $(\mu_1^{(m)}, \mu_2^{(m)}, \mathbf{S}^{(m)})$ denote a draw obtained from Gibbs sampling as in *Algorithm 3.4*. To implement the random permutation Gibbs sampler, we perform a random permutation of the labels after each draw. For $K = 2$, there are only two permutations, namely the identity $\rho_1(1) = 1, \rho_1(2) = 2$, and interchanging the labels: $\rho_2(1) = 2, \rho_2(2) = 1$. Thus with a probability of 0.5 the draws remain unchanged, whereas with probability 0.5 the labels are interchanged

by substituting (μ_1, μ_2) by (μ_2, μ_1), and switching the allocations, which take the value 1, if they are 2, and take the value 2, if they are 1. Figure 3.4 shows the marginal posterior densities $p(\mu_k|\mathbf{y})$ estimated from random permutation MCMC sampling for both synthetic data sets. As expected from the theoretical considerations in Subsection 3.3.2, these densities are identical.

3.6 Other Monte Carlo Methods Useful for Mixture Models

In the previous section we focused on data augmentation and MCMC methods, but other Monte Carlo methods have been found to be useful for finite mixture models.

3.6.1 A Metropolis–Hastings Algorithm for the Parameters

Several authors (Celeux et al., 2000; Brooks, 2001; Viallefont et al., 2002) use a Metropolis–Hastings algorithm to generate a sample from the mixture posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$. This is feasible, because the Metropolis–Hastings algorithm requires knowledge of the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ only up to a normalizing constant. The Metropolis–Hastings algorithm, introduced in Subsection 3.4.2 in the context of sampling allocations $\mathbf{S}$ from the posterior $p(\mathbf{S}|\mathbf{y})$, is implemented in the following manner to simulate $\boldsymbol{\vartheta}$ from the mixture posterior $p(\boldsymbol{\vartheta}|\mathbf{y})$.

Algorithm 3.6: Sampling the Parameters of a Finite Mixture Through a Metropolis–Hastings Algorithm Start with some parameter $\boldsymbol{\vartheta}^{(0)}$ and repeat the following steps for $m = 1, \ldots, M_0, \ldots, M + M_0$.

(a) Propose a new parameter $\boldsymbol{\vartheta}^{new}$ by sampling from a proposal density $q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(m-1)})$.
(b) Move the sampler to $\boldsymbol{\vartheta}^{new}$ with probability $\min(1, A)$, where

$$A = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}^{new})p(\boldsymbol{\vartheta}^{new})q(\boldsymbol{\vartheta}^{(m-1)}|\boldsymbol{\vartheta}^{new})}{p(\mathbf{y}|\boldsymbol{\vartheta}^{(m-1)})p(\boldsymbol{\vartheta}^{(m-1)})q(\boldsymbol{\vartheta}^{new}|\boldsymbol{\vartheta}^{(m-1)})}. \tag{3.43}$$

If $U < \min(1, A)$, where U is a random number from the $\mathcal{U}[0, 1]$-distribution, then accept $\boldsymbol{\vartheta}^{new}$ and set $\boldsymbol{\vartheta}^{(m)} = \boldsymbol{\vartheta}^{new}$, otherwise reject $\boldsymbol{\vartheta}^{new}$ and set $\boldsymbol{\vartheta}^{(m)} = \boldsymbol{\vartheta}^{(m-1)}$.

Increase m by one, and return to step (a).

Hurn et al. (2003) use the following multivariate random walk proposal on a suitably transformed parameter $\phi(\boldsymbol{\vartheta})$, which is obtained from a log-transform on variance parameters and a logit transform on the weights,

$$\phi(\boldsymbol{\vartheta}^{new}|\boldsymbol{\vartheta}^{(m-1)}) = \phi(\boldsymbol{\vartheta}^{(m-1)}) + C\epsilon,$$

where ϵ follows a multivariate Cauchy distribution. C is calibrated during a pilot-run to lead to an acceptance rate of about 40%.

An advantage of this method compared to data augmentation and MCMC is that sampling of the indicators is avoided. A disadvantage is that tuning the proposal density may require several pilot runs.

3.6.2 Importance Sampling for the Allocations

An alternative attempt at sampling from $p(\mathbf{S}|\mathbf{y})$ has been investigated in Casella et al. (2000). Rather than drawing from $p(\mathbf{S}|\mathbf{y})$, they draw a sequence $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(L)}$ from an importance density $q(\mathbf{S})$. One way to construct the importance density is to ignore posterior correlation among the indicators, which is actually only introduced through the prior $p(\mathbf{S})$, and to use a density with independent components:

$$q(\mathbf{S}) = \prod_{i=1}^{N} q(S_i|\mathbf{y}_i), \qquad q(S_i|\mathbf{y}_i) \propto p(\mathbf{y}_i|S_i)p(S_i). \tag{3.44}$$

Under the conjugate Dirichlet prior $\boldsymbol{\eta} \sim \mathcal{D}(e_0, \ldots, e_0)$, we obtain the following marginal prior for a single indicator S_i,

$$p(S_i) \propto \Gamma(1 + e_0)\Gamma(e_0)^{K-1}, \tag{3.45}$$

and the marginal likelihood of $\mathbf{y}_i$ given $S_i = k$ results from (3.25),

$$p(\mathbf{y}_i|S_i = k) = \frac{p(\mathbf{y}_i|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k)}{p(\boldsymbol{\theta}_k|\mathbf{y}_i)}, \tag{3.46}$$

where $p(\boldsymbol{\theta}_k|\mathbf{y}_i)$ is the posterior density from the single observation $\mathbf{y}_i$. The right-hand side of (3.46) may be evaluated for arbitrary $\boldsymbol{\theta}_k$, in particular for the posterior mode of $p(\boldsymbol{\theta}_k|\mathbf{y}_i)$. (3.46) is likely to be unstable for high-dimensional parameter $\boldsymbol{\theta}_k$, where the posterior $p(\boldsymbol{\theta}_k|\mathbf{y}_i)$ is not well defined from a single observation.

To improve the efficiency of importance sampling, Casella et al. (2000) use stratified importance sampling by decomposing the space of all possible allocations into all partition sets with identical allocation size $N_k(\mathbf{S})$. Casella et al. (2000) argue that among these partition sets only a few carry most of the weights.

Casella et al. (2000) use draws from the importance density to approximate the posterior expectation of any function $h(\boldsymbol{\vartheta})$ as explained, for instance, in Geweke (1989):

$$\mathrm{E}(h(\boldsymbol{\vartheta})|\mathbf{y}) \approx \frac{1}{L}\sum_{l=1}^{L} \mathrm{E}(h(\boldsymbol{\vartheta})|\mathbf{y}, \mathbf{S}^{(l)})\frac{p(\mathbf{S}^{(l)}|\mathbf{y})}{q(\mathbf{S}^{(l)})}. \tag{3.47}$$

A certain objection to this approach is that the ergodic average (3.47) may be biased due to undetected label switching.

3.6.3 Perfect Sampling

Like MCMC, perfect sampling is based on the idea of constructing a Markov chain where the stationary distribution is equal to an untractable posterior distribution. Whereas MCMC exploits the fact that for an ergodic Markov chain the stationary distribution is also the limiting distribution, perfect sampling is an algorithm for generating independent draws from precisely the exact stationary distribution; see Casella et al. (2001) for an introduction.

The construction of a perfect sampler for mixture models is a delicate issue as the first attempt of Hobert et al. (1999) demonstrates where they applied perfect sampling to two- and three-component mixtures where the component parameters are known. Casella et al. (2002) extend these results to finite mixtures with an arbitrary number of components and unknown component parameters where the marginal posterior $p(\mathbf{S}|\mathbf{y})$ of the allocations is available explicitly up to a constant; see also Subsection 3.3.3.

3.7 Bayesian Inference for Finite Mixture Models Using Posterior Draws

From a Bayesian perspective, the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ contains all information provided by the data, and is the basis for drawing inference on any quantity of interest. If a sampling-based approach as described in Sections 3.5 and 3.6 is pursued for practical estimation, a sequence of draws $\{\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M\}$ from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$ is available, which could be used to approximate all quantities of interest. In what follows, it is assumed that an appropriate amount of initial draws M_0 has been removed, if the draws were produced by an MCMC sampler.

3.7.1 Sampling Representations of the Mixture Posterior Density

It is sometimes helpful to visualize the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, but producing a simple density plot is feasible only for very simple problems, where the unknown parameter $\boldsymbol{\vartheta}$ is at most bivariate. If the dimension of $\boldsymbol{\vartheta}$ exceeds two, other tools have been developed for visualizing the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$. Draws from the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ have been used as a sampling representation of the mixture posterior distribution, which is then visualized in an appropriate manner (Celeux et al., 2000; Frühwirth-Schnatter, 2001b; Hurn et al., 2003).

To illustrate the equivalence of a density plot and the sampling representation, Figure 3.5 compares the contours of the mixture posterior density $p(\mu_1, \mu_2|\mathbf{y})$ with MCMC draws $\mu_1^{(m)}$ and $\mu_2^{(m)}$ from $p(\mu_1, \mu_2|\mathbf{y})$ obtained from random permutation Gibbs sampling using *Algorithm 3.5* for the synthetic

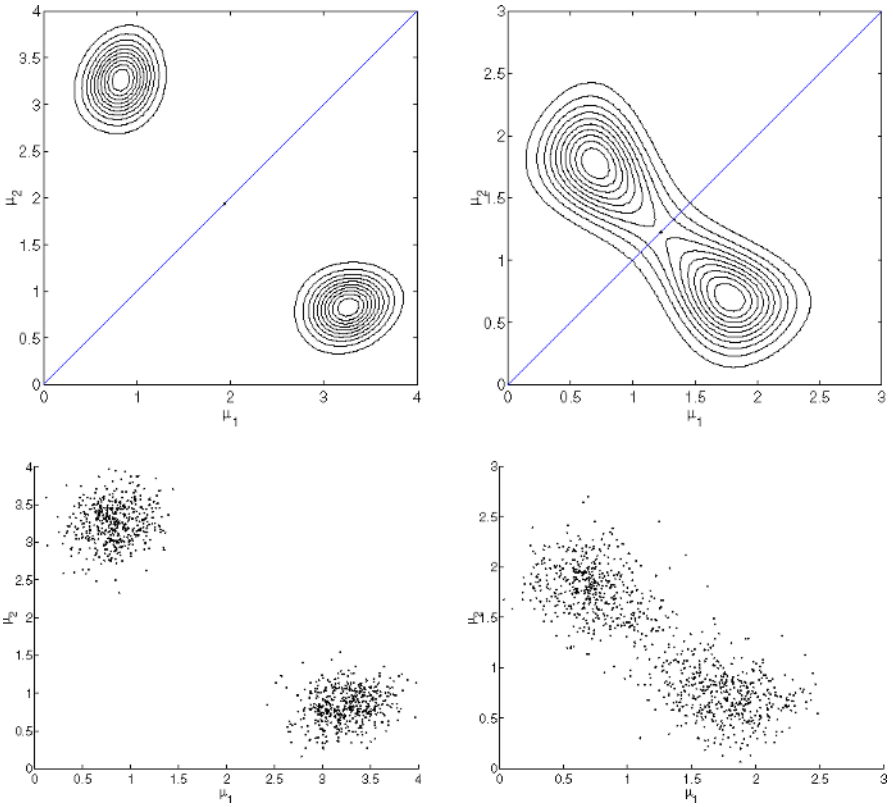

Fig. 3.5. HIDDEN AGE GROUPS — Synthetic Data Set 1 (left-hand side) and Data Set 2 (right-hand side); top: contours of the mixture posterior density $p(\mu_1, \mu_2|\mathbf{y})$, bottom: MCMC draws from the mixture posterior density $p(\mu_1, \mu_2|\mathbf{y})$ obtained from random permutation sampling)

data sets 1 and 2 discussed earlier in Subsection 3.5.6. By using the random permutation Gibbs sampler, rather than standard Gibbs sampling, the exploration of both modes of the posterior distribution is forced.

In particular, for higher-dimensional problems sampling representations are a very useful tool for visualizing the mixture posterior distribution. One interesting view is the bivariate marginal density $p(\theta_{k,j}, \theta_{k',j}|\mathbf{y})$, where $k \neq k'$, visualized for each $j = 1, \ldots, d$, through scatter plots of the MCMC draws $(\theta_{k,j}^{(m)}, \theta_{k',j}^{(m)})$. By the results of Subsection 3.3.2, this density is the same for all pairs of (k, k'), thus $k = 1$ and $k' = 2$, or any other pair, may be selected, provided that the random permutation Gibbs sampler has been used. These figures allow us to study how much the jth element $\theta_{k,j}$ of the component parameter $\boldsymbol{\theta}_k$ differs among the various components. If this element is significantly different among all components, then this plot shows $K^2 - K = K(K-1)$

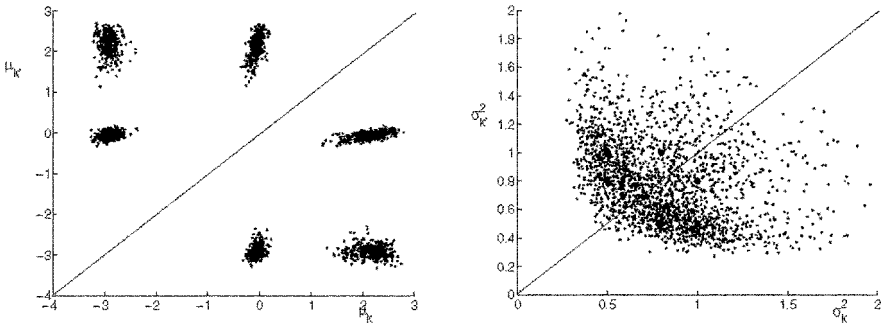

Fig. 3.6. Synthetic data of size $N = 500$ simulated from a mixture of three univariate normal distributions with $\eta_1 = 0.3$, $\eta_2 = 0.5$, $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.8$; sampling representation of $p(\mu_k, \mu_{k'}|\mathbf{y})$ (left-hand side) and $p(\sigma_k^2, \sigma_{k'}^2|\mathbf{y})$ (right-hand side) based on random permutation Gibbs sampling

simulation clusters. If this element is nearly the same in all components, then this plot shows a single simulation cluster; see Figure 3.6 for illustration.

Another useful view is the bivariate marginal density $p(\theta_{k,j}, \theta_{k,j'}|\mathbf{y})$, which is visualized separately for each pair $j, j' = 1, \ldots, d, j \neq j'$ through scatter plots of the MCMC draws $(\theta_{k,j}^{(m)}, \theta_{k,j'}^{(m)})$. By the results of Subsection 3.3.2, this density is the same for all $k = 1, \ldots, K$, thus $k = 1$ or any other value may be selected. If the dimension of $\boldsymbol{\theta}_k$ is equal to two, this scatter plot is closely related to the point process representation of the underlying mixture distribution, discussed in Subsection 1.2.2. The MCMC draws will scatter around the points corresponding to the true point process representation, with the spread of the clouds representing the uncertainty of estimating the points; see Figure 3.7 for illustration. This is also true for multivariate component parameters, where the plots correspond to projections of the point process representation onto bivariate subspaces.

These figures allow us to study the component parameters in relation to each other without having to worry about label switching. In Figure 3.7, for instance, it becomes evident that the components differ mainly in the mean, that two components have nearly the same variance, whereas the third component has a variance which is slightly smaller.

For a mixture with a univariate component parameter θ_k a bivariate plot is not available. In this case $\theta_k^{(m)}$ may be plotted against $\eta_k^{(m)}$ or an auxiliary parameter $\psi^{(m)}$ which is drawn from a standard normal distribution.

3.7.2 Using Posterior Draws for Bayesian Inference

On the basis of the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, inference is drawn on quantities of interest such as the posterior mean $\mathrm{E}(\boldsymbol{\vartheta}|\mathbf{y})$, which commonly is used as a

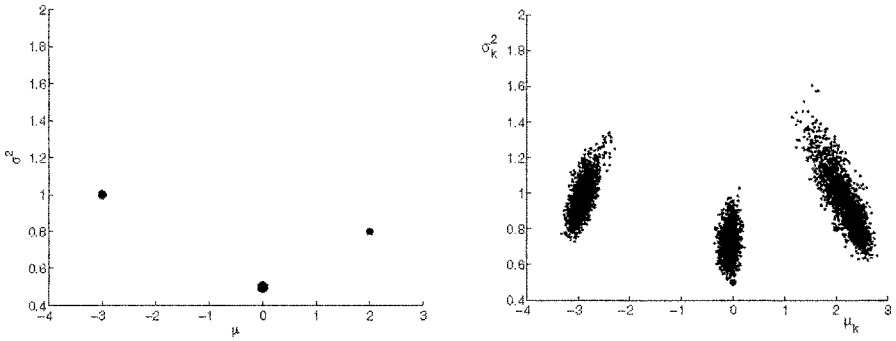

Fig. 3.7. Synthetic data of size $N = 500$ simulated from a mixture of three univariate normal distributions with $\eta_1 = 0.3$, $\eta_2 = 0.5$, $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.8$; point process representation of the finite mixture distribution (left hand side) and point process representation of draws from $p(\mu_k, \sigma_k^2|\mathbf{y})$ based on random permutation Gibbs sampling (right-hand side)

point estimator of $\boldsymbol{\vartheta}$, or the predictive density $p(\mathbf{y}_f|\mathbf{y})$, which is a pointwise estimator of the density of the marginal distribution of the observed random variable $\mathbf{Y}$.

For finite mixture models, as for many other interesting and complex statistical models, no explicit expression is available for most quantities of interest, and draws $\{\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M\}$ from the posterior density are used to approximate all quantities of interest. Consider, as an example, the posterior expectation

$$\mathrm{E}(h(\boldsymbol{\vartheta})|\mathbf{y}) = \int h(\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\mathbf{y}) d\boldsymbol{\vartheta}$$

of a function $h(\boldsymbol{\vartheta})$, which is approximated by averaging over the draws from the posterior distribution in the following way,

$$\overline{h}_M = \frac{1}{M} \sum_{m=1}^{M} h(\boldsymbol{\vartheta}^{(m)}). \tag{3.48}$$

Under mild conditions, $\overline{h}_M$ converges to $\mathrm{E}(h(\boldsymbol{\vartheta})|\mathbf{y})$ by the law of large numbers, even if the draws were generated by a Markov chain Monte Carlo method (Tierney, 1994). There are several questions associated with Bayesian inference based on posterior draws, in particular convergence diagnostics and choosing appropriate values of M, which are beyond the scope of this book, and are addressed, for example, in the excellent books by Robert and Casella (1999) and Liu (2001).

For finite mixture models, a specific issue arises that is related to the invariance of the posterior distribution discussed in Subsection 3.3.2 and the

label switching problem discussed in Subsection 3.5.5. Bayesian inference for finite mixture models using posterior draws may be, but need not, be sensitive to label switching.

Label switching does not matter whenever the function $h(\boldsymbol{\vartheta})$ is invariant to relabeling the components of the mixture:

$$h(\boldsymbol{\vartheta}) = h(\tilde{\boldsymbol{\vartheta}}_s), \tag{3.49}$$

where $\tilde{\boldsymbol{\vartheta}}_s$ is the permuted parameter defined in (3.14). In such a case, averaging over the draws $h(\boldsymbol{\vartheta}^{(m)})$ as in (3.48) is evidently insensitive to label switching, and any of the methods discussed in Section 3.5 such as data augmentation and Gibbs sampling (*Algorithm 3.4*) or data augmentation and random permutation Gibbs sampling (*Algorithm 3.5*) may be used.

It is not always easy to identify functionals that are invariant to relabeling, in particular, if inference concerns the component parameters $(\boldsymbol{\theta}_k, \eta_k)$. Obvious estimators turn out to be sensitive to label switching, in which case it is necessary to identify the model before making an inference, as explained in detail in Subsection 3.7.7. Clustering of a single object $\mathbf{y}_i$, based on the posterior probability distribution $\Pr(S_i = k|\mathbf{y})$, into one of the K hidden groups, is a further example of an inference problem where any kind of label switching matters; see Subsection 7.1.7 for more detail.

3.7.3 Predictive Density Estimation

A quantity that often is of interest when fitting a finite mixture model, is the posterior predictive density $p(\mathbf{y}_f|\mathbf{y})$ of a future realization $\mathbf{y}_f$, given the data $\mathbf{y}$, which is given by

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

This density is the posterior expectation of following function $h(\boldsymbol{\vartheta}) = p(\mathbf{y}_f|\boldsymbol{\vartheta})$,

$$p(\mathbf{y}_f|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k p(\mathbf{y}_f|\boldsymbol{\theta}_k), \tag{3.50}$$

which is invariant to relabeling the components of the mixture. Therefore, the density estimated from the MCMC draws,

$$\hat{p}(\mathbf{y}_f|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{k=1}^{K} \eta_k^{(m)} p(\mathbf{y}_f|\boldsymbol{\theta}_k^{(m)}) \right), \tag{3.51}$$

is robust against label switching. For illustration, consider Figure 3.8 which compares a histogram of the synthetic data sets 1 and 2 discussed earlier in Subsection 3.5.5 with the predictive density estimate $\hat{p}(y_f|\mathbf{y})$.

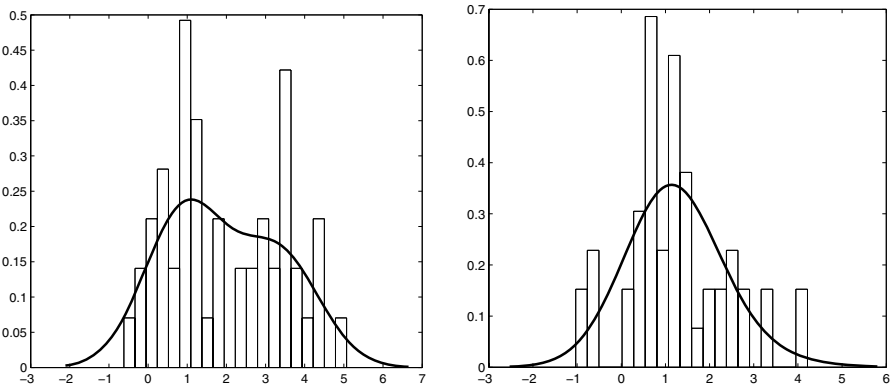

Fig. 3.8. HIDDEN AGE GROUPS — Synthetic Data Set 1 (left-hand side) and Synthetic Data Set 2 (right-hand side); predictive density estimate obtained from fitting a two-component normal mixture with μ_1, μ_2, and η_1 unknown (variance $\sigma_1^2 = \sigma_2^2 = 1$ fixed) in comparison to a histogram of the data

For univariate mixtures of normals, Richardson and Green (1997) studied MCMC estimation under various constraints, and observed that the predictive density estimator $\hat{p}(\mathbf{y}_f|\mathbf{y})$ differed significantly across the constraints, which is not surprising as a poor constraint introduces a bias; see again the discussion in Subsection 3.5.6. For this reason it is recommended to use draws from the unconstrained posterior when the mixture model is used for practical density estimation or as a smoothing device. Due to the invariance to relabeling, the estimator $\hat{p}(\mathbf{y}_f|\mathbf{y})$ could be based on Gibbs sampling (*Algorithm 3.4*) or random permutation Gibbs sampling (*Algorithm 3.5*).

The Posterior Predictive Distribution of a Sequence

It is possible to predict a whole sequence $\mathbf{y}_f = (\mathbf{y}_{f,1}, \ldots, \mathbf{y}_{f,H})$ of length $H \geq 1$, given the data $\mathbf{y}$. The posterior predictive density $p(\mathbf{y}_f|\mathbf{y})$ of $\mathbf{y}_f$, conditional on the observations $\mathbf{y}$ is given by

$$p(\mathbf{y}_f|\mathbf{y}) = \int \prod_{h=1}^{H} p(\mathbf{y}_{f,h}|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\mathbf{y}) d\boldsymbol{\vartheta}. \tag{3.52}$$

Analytical integration is not possible, but one could easily draw a sample from (3.52) if a sequence of draws from the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ is available, using the following algorithm.

Algorithm 3.7: Sampling from the Posterior Predictive Distribution Assume that a sequence of draws $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(M)}$ from the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ is available. Perform the following two steps for $m = 1, \ldots, M$.

(a) Draw H component indicators $S_1, \ldots, S_H$ independently from the discrete distribution $(\eta_1^{(m)}, \ldots, \eta_K^{(m)})$.

(b) For each $h = 1, \ldots, H$, sample $\mathbf{y}_{f,h}^{(m)}$ from the component density $p(\mathbf{y}|\boldsymbol{\theta}_h)$, where $\boldsymbol{\theta}_h = \boldsymbol{\theta}_{S_h}^{(m)}$. Define $\mathbf{y}_f^{(m)} = (\mathbf{y}_{f,1}^{(m)}, \ldots, \mathbf{y}_{f,H}^{(m)})$.

The sample $\mathbf{y}_f^{(1)}, \ldots, \mathbf{y}_f^{(M)}$ produced by this algorithm is a sample from the posterior predictive distribution $p(\mathbf{y}_f|\mathbf{y})$.

3.7.4 Individual Parameter Inference

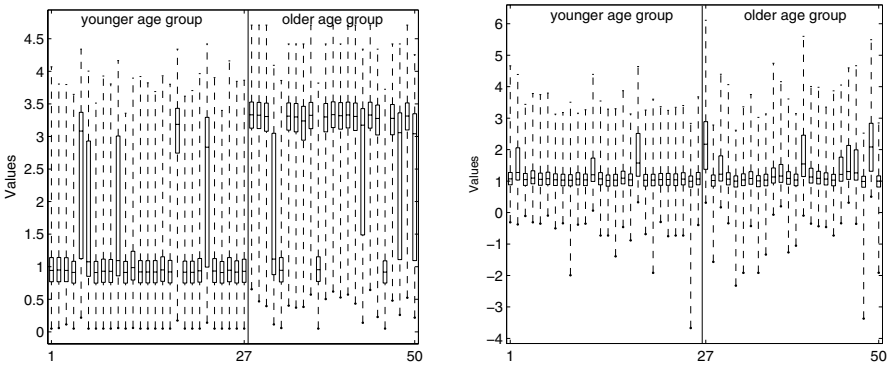

Fig. 3.9. HIDDEN AGE GROUPS — Synthetic Data Set 1 (left-hand side) and Synthetic Data Set 2 (right-hand side); reconstruction of the individual means μ_i^s for the two age groups obtained from fitting a two-component normal mixture with μ_1, μ_2, and η_1 unknown (variance $\sigma_1^2 = \sigma_2^2 = 1$ fixed); box plots of the posterior draws of μ_i^s

Often it is of interest to make an inference about the individual parameters $\boldsymbol{\theta}_i^s$, which are defined for each subject i, $i = 1, \ldots, N$, by

$$\boldsymbol{\theta}_i^s = \sum_{k=1}^{K} \boldsymbol{\theta}_k I_{\{S_i = k\}}. \tag{3.53}$$

Obviously, $\boldsymbol{\theta}_i^s$ is invariant to relabeling the components of the mixture. Consequently, the sequence $\{(\boldsymbol{\theta}_1^{s,(m)}, \ldots, \boldsymbol{\theta}_N^{s,(m)}), m = 1, \ldots, M\}$, which is determined from the posterior draws $\{\boldsymbol{\vartheta}^{(m)}\}, m = 1, \ldots, M$ through the transformation (3.53),

$$\boldsymbol{\theta}_i^{s,(m)} = \boldsymbol{\theta}_{k_m}^{(m)}, \qquad k_m = S_i^{(m)},$$

contains M draws from the joint posterior distribution $p(\boldsymbol{\theta}_1^s, \ldots, \boldsymbol{\theta}_N^s|\mathbf{y})$, which are insensitive to label switching. It is possible to visualize the individual parameters $\boldsymbol{\theta}_i^s$ through box-plots of $\{\boldsymbol{\theta}_i^{s,(m)}, m = 1, \ldots, M\}$ for each $i = 1, \ldots, N$, which estimate the marginal distribution $p(\boldsymbol{\theta}_i^s|\mathbf{y})$. To obtain a point estimator of $\boldsymbol{\theta}_i^s$, the expected value $\mathrm{E}(\boldsymbol{\theta}_i^s|\mathbf{y})$ is estimated from the posterior draws in an obvious way:

$$\hat{\boldsymbol{\theta}}_i^s = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_i^{s,(m)}.$$

An Illustrative Example

For illustration we consider the synthetic data sets 1 and 2, discussed earlier in Subsection 3.5.5. The true value of μ_i^s is equal to μ_y for the younger age group and equal to μ_o for the older age group. In Figure 3.9, box plots of $\mu_i^{s,(m)}$ are shown for both data sets, based on data augmentation and random permutation Gibbs sampling (*Algorithm 3.5*). Reconstruction of μ_i^s is rather precise for data set 1, whereas the lack of separation between the two groups leads to rather imprecise reconstructions for data set 2. This is, however, not due to any deficiencies of the sampling method, but due to a lack of information in the data.

3.7.5 Inference on the Hyperparameter of a Hierarchical Prior

Note that the hyperparameter $\boldsymbol{\delta}$ is invariant by definition, and may be easily estimated from the MCMC output by taking ergodic averages over the posterior draws $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(M)}$.

3.7.6 Inference on Component Parameters

When making an inference about the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, one is actually interested in an inference on the corresponding hidden groups in the population. Only under a unique labeling, does a fixed link exist between a hidden group with group-specific parameter $\boldsymbol{\theta}_G$ and a certain component in the mixture with component parameter $\boldsymbol{\theta}_k$. If this labeling remains the same throughout MCMC sampling, then the draws $\{\boldsymbol{\theta}_k^{(m)}, m = 1, \ldots, M\}$ may be regarded as posterior draws for the parameter $\boldsymbol{\theta}_G$, and it is possible to average over these draws to obtain a point estimator of the group-specific parameter $\boldsymbol{\theta}_G$:

$$\hat{\boldsymbol{\theta}}_G = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_k^{(m)}. \tag{3.54}$$

However, if label switching took place during sampling, then the hidden group parameter $\boldsymbol{\theta}_G$ no longer has to be associated with $\boldsymbol{\theta}_k$, but with another component parameter $\boldsymbol{\theta}_{k'}$. When averaging over the draws of $\boldsymbol{\theta}_k$ as in (3.54), a biased point estimator of the group-specific parameter $\boldsymbol{\theta}_G$ results, which is pulled toward the overall mean $\mathrm{E}(\boldsymbol{\theta}_k|\mathbf{y})$ of the unconstrained posterior.

To draw an inference about hidden groups by averaging over posterior draws, it is essential that these draws arise from a single labeling subspace $\mathcal{L}$. We denote such draws as $\boldsymbol{\vartheta}^{\mathcal{L},(m)} = (\boldsymbol{\theta}_1^{\mathcal{L},(m)}, \ldots, \boldsymbol{\theta}_K^{\mathcal{L},(m)}, \eta_1^{\mathcal{L},(m)}, \ldots, \eta_K^{\mathcal{L},(m)})$. These draws could be used to estimate the parameters in the hidden groups by

$$\hat{\boldsymbol{\theta}}_k = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_k^{\mathcal{L},(m)}, \tag{3.55}$$

as well as the group sizes by

$$\hat{\eta}_k = \frac{1}{M} \sum_{m=1}^{M} \eta_k^{\mathcal{L},(m)}. \tag{3.56}$$

It is discussed in detail in Subsection 3.7.7 how to obtain posterior draws from a unique labeling subspace.

Choosing Invariant Loss Functions

It should be noted that not all point estimators of $\boldsymbol{\vartheta}$ are sensitive to label switching. Whether this is the case depends on the underlying loss function. Within a decision-theoretic framework any point estimator $\boldsymbol{\vartheta}^\star$ is derived as that value which minimizes the expected posterior loss under a certain loss function $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta})$:

$$\boldsymbol{\vartheta}^\star = \arg\min_{\hat{\boldsymbol{\vartheta}}} \mathrm{E}(R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta})|\mathbf{y}) = \int_\Theta R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\mathbf{y}) d\boldsymbol{\vartheta};$$

see Berger (1985) for a full account. If this framework is applied to finite mixture models, sensible estimators are obtained only if the loss function $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta})$, which corresponds to $h(\boldsymbol{\vartheta})$ in (3.48), is invariant to relabeling the components of the mixture.

This leads immediately to problems with the quadratic loss-function $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}) = (\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})'(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$, which yields the posterior mean $\mathrm{E}(\boldsymbol{\vartheta}|\mathbf{y})$ as optimal estimator, and is for many other statistical models the most commonly used loss function. Evidently, the functional value of $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta})$ changes when the components of the mixture are relabeled, leading to an ambiguous definition of the expected risk. Interestingly enough, it was realized earlier that the posterior mean $\mathrm{E}(\boldsymbol{\vartheta}|\mathbf{y})$ is not a sensible point estimator as it does not contain any component-specific information; see again (3.21).

The 0/1 loss function, for which the posterior mode turns out to be the optimal estimator (see, for instance, Zellner, 1971), is easily adapted to finite mixture models by defining that $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}) = 0$ iff $\hat{\boldsymbol{\vartheta}}$ and $\boldsymbol{\vartheta}$ are identical up to permutations, otherwise $R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta})$ is equal to 1. This loss function is invariant to relabeling, and the mode of the mixture posterior may be used for estimation. The posterior mode may be approximated from the posterior draws $\{\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M\}$ through that value which maximizes the nonnormalized mixture posterior density $p^\star(\boldsymbol{\vartheta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$.

Various alternative loss functions have been considered for parameter estimation in mixture models. Celeux et al. (2000) consider loss functions that are based on the predictive density $p(\mathbf{y}_f|\boldsymbol{\vartheta})$ which is invariant to relabeling the components; see again (3.50). Examples include the integrated squared difference

$$R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}) = \int_{\mathcal{Y}} (p(\mathbf{y}_f|\boldsymbol{\vartheta}) - p(\mathbf{y}_f|\hat{\boldsymbol{\vartheta}}))^2 d\mathbf{y}_f,$$

and the symmetrized Kullback–Leibler distance

$$R(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}) = \int_{\mathcal{Y}} \left(p(\mathbf{y}_f|\boldsymbol{\vartheta}) \log \frac{p(\mathbf{y}_f|\boldsymbol{\vartheta})}{p(\mathbf{y}_f|\hat{\boldsymbol{\vartheta}})} + p(\mathbf{y}_f|\hat{\boldsymbol{\vartheta}}) \log \frac{p(\mathbf{y}_f|\hat{\boldsymbol{\vartheta}})}{p(\mathbf{y}_f|\boldsymbol{\vartheta})} \right) d\mathbf{y}_f,$$

where in both cases integration reduces to summation for a discrete sample space $\mathcal{Y}$. In both cases the expected loss is given by an expression that contains expectations of terms such as $p(\mathbf{y}_f|\boldsymbol{\vartheta})$, $p(\mathbf{y}_f|\boldsymbol{\vartheta})^2$, or $\log p(\mathbf{y}_f|\boldsymbol{\vartheta})$, with respect to the posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$. The practical evaluation of these estimators is rather involved and Celeux et al. (2000) follow the two-step procedure of Rue (1995). In a first step, expectations with respect to the posterior density are evaluated using posterior draws and integration with respect to $\mathbf{y}_f$ is carried out using some numerical technique. In a second step, the minimization problem for the estimator $\boldsymbol{\vartheta}^\star$ is solved using simulated annealing. We refer to Celeux et al. (2000) for further computational details.

Dias and Wedel (2004) provide an empirical comparison of EM and MCMC performance, which includes different prior specifications and various procedures to deal with the label switching problem.

3.7.7 Model Identification

The parameter estimation problem discussed in Subsection 3.7.6 illustrated that care must be exercised when using draws from the mixture posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$ to estimate functionals of $\boldsymbol{\vartheta}$, which are not invariant to relabeling the components of the finite mixture. Inference on such functionals is sensible only if the posterior draws come from a unique labeling subspace of the unconstrained parameter space. The discussion of this subsection is devoted to the difficult task of identifying such draws.

Gibbs sampling as described in *Algorithm 3.4* may lead to implicit model identification if the $K!$ modal parts of the mixture posterior density are very well separated, and the sampler is trapped in one of modal regions; see again the discussion in Subsection 3.5.6. In this case the posterior draws obtained by *Algorithm 3.4* may be treated as coming from a unique labeling subspace, $\boldsymbol{\vartheta}^{\mathcal{L},(m)} = \boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M$, as was done for instance in Chib (1996). It is, however, not recommended to rely blindly on this implicit model identification, as the behavior of Gibbs sampling is unpredictable in this respect.

One strategy is to relabel the posterior draws $\{\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M\}$ in such a way that draws $\{\boldsymbol{\vartheta}^{\mathcal{L},(m)}, m = 1, \ldots, M\}$ from a unique labeling subspace result. This may be achieved by isolating a sensible identifiability constraint through exploring the posterior draws (Frühwirth-Schnatter, 2001b) or by unsupervised clustering of the posterior draws (Celeux, 1998).

Model Identification Through Identifiability Constraint

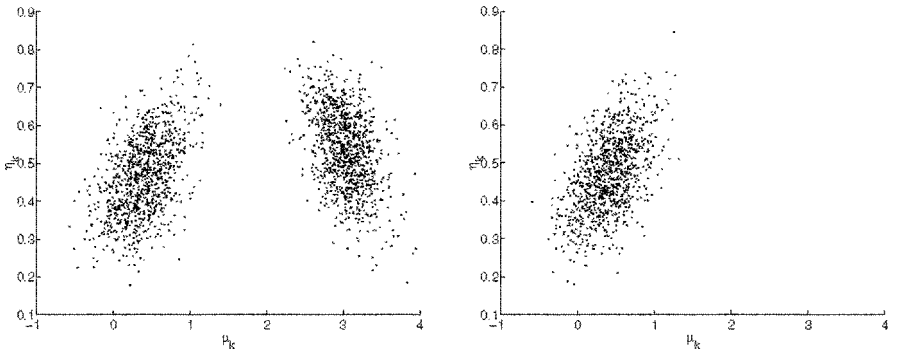

Fig. 3.10. HIDDEN AGE GROUPS — Synthetic Data Set 3; draws from the bivariate marginal distributions $p(\mu_k, \eta_k|\mathbf{y})$ (left-hand side); posterior draws of (μ_1, η_1) under the constraint $\mu_1 < \mu_2$ (right-hand side)

A common reaction to the label switching problem is to impose some formal identifiability constraint as in Subsection 1.3.3 within sampling-based Bayesian estimation (Albert and Chib, 1993; Richardson and Green, 1997). It has been realized only rather recently that an arbitrary formal identifiability constraint does not necessarily generate a unique labeling and that a poorly chosen constraint introduces a bias (Celeux, 1998; Celeux et al., 2000; Stephens, 2000b; Frühwirth-Schnatter, 2001b); recall also the discussion at the end of Subsection 2.4.3.

To identify sensible identifiability constraints, Frühwirth-Schnatter (2001b) explored the point process representation of the MCMC draws, introduced

earlier in Subsection 3.7.1. Note that the constraint is only an indirect device to describe the differences between the components, and therefore is not necessarily unique. Various case studies where it is useful to explore the point process representation of the MCMC draws in this way may be found throughout the book; see also Frühwirth-Schnatter (2001a, 2001b), Kaufmann and Frühwirth-Schnatter (2002), and Frühwirth-Schnatter et al. (2004).

A straightforward method to impose a constraint on the posterior draws is to postprocess the MCMC draws that were generated from the mixture posterior. Whenever a draw does not satisfy the constraint, one permutes the labeling of the components such that the constraint is fulfilled (Richardson and Green, 1997; Stephens, 1997b; Frühwirth-Schnatter, 2001b). Frühwirth-Schnatter (2001b) also provides a formal proof that this method actually delivers a sample from the constrained posterior.

For illustration, we return to the synthetic data set 3 introduced at the end of Subsection 2.4.3. Figure 3.10 shows a sampling representation of the bivariate marginal distribution $p(\mu_k, \eta_k|\mathbf{y})$ for these data. From this scatter plot it is obvious that the component parameters differ mainly in the mean, whereas the weights are rather equal. The constraint $\mu_1 < \mu_2$ is actually able to impose a unique labeling.

Model Identification Through Unsupervised Clustering of the Posterior Draws

For higher-dimensional problems, in particular for multivariate mixtures, it is possible, but somewhat time-consuming to search for identifiability constraints in the MCMC output (Frühwirth-Schnatter et al., 2004). As a more automatic procedure, Celeux (1998) suggested permuting the MCMC draws obtained from unconstrained sampling by using a clustering procedure. His algorithm is an on-line k-means type algorithm with $K!$ clusters, which is initialized from the first 100 draws after reaching burn-in, by defining $K!$ reference centers from these draws. For each MCMC draw $\boldsymbol{\vartheta}^{(m)}$ the distance to each of these $K!$ centers is computed, which is then used to permute the labels.

Model Identification Through Clustering in the Point Process Representation

A related method is to search for clusters in the point process representation of the MCMC draws, introduced earlier in Subsection 3.7.1, which additionally provides some control over the important question of whether the model is overfitting the number of components.

Each of the MCMC simulations $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ corresponds to a certain point process representation that will cluster around the point process representation of the underlying true finite mixture distribution. If the heterogeneity between the underlying points is large enough, K simulation clusters will be present in the point process representation of the MCMC draws.

Permuting the labels of $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ does not change the point representation; it only changes the one-to-one correspondence between the component-specific draws and the simulation clusters. A unique labeling is achieved if all draws $\boldsymbol{\theta}_k^{(m)}$ are associated with the same simulation cluster for all $m = 1, \ldots, M$. This is achieved by applying a standard k-means clustering algorithm with K clusters to a sample of size MK, formed from the MCMC draws $\{(\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}), m = 1, \ldots, M\}$, with the posterior mode estimator $(\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star)$ serving as a starting value for the cluster means. The clustering algorithm delivers a classification sequence $\{(\rho_m(1), \ldots, \rho_m(K)), m = 1, \ldots, M\}$, where $\rho_m(k)$ determines to which simulation cluster the MCMC draw $\boldsymbol{\theta}_k^{(m)}$ belongs.

If the simulation clusters are well separated, then the classification sequence $\{\rho_m(1), \ldots, \rho_m(K)\}$ is a permutation of $\{1, \ldots, K\}$; that is,

$$\sum_{k=1}^{K} \rho_m(k) = \frac{K(K+1)}{2}. \tag{3.57}$$

In this case it is possible to relabel the MCMC draw $\boldsymbol{\vartheta}^{(m)}$ through the permutation $\{\rho_m(1), \ldots, \rho_m(K)\}$; that is

$$\boldsymbol{\vartheta}^{\mathcal{L},(m)} = (\boldsymbol{\theta}_{\rho_m(1)}^{\mathcal{L},(m)}, \ldots, \boldsymbol{\theta}_{\rho_m(K)}^{\mathcal{L},(m)}, \eta_{\rho_m(1)}^{\mathcal{L},(m)}, \ldots, \eta_{\rho_m(K)}^{\mathcal{L},(m)}) \tag{3.58}$$

defined by:

$$\boldsymbol{\theta}_{\rho_m(k)}^{\mathcal{L},(m)} = \boldsymbol{\theta}_k^{(m)}, \qquad \eta_{\rho_m(k)}^{\mathcal{L},(m)} = \eta_k^{(m)}, \qquad k = 1, \ldots, K.$$

As all component-specific draws are associated with the same simulation cluster, the draws defined in (3.58) may be regarded as coming from an identified mixture model.

If in addition to $\boldsymbol{\vartheta}^{(m)}$, allocation variables $\mathbf{S}^{(m)} = (S_1^{(m)}, \ldots, S_N^{(m)})$ have been stored, then the same permutation could be used on them to define allocation under a unique labeling for each $i = 1, \ldots, N$:

$$S_i^{\mathcal{L},(m)} = \rho_m(S_i^{(m)}). \tag{3.59}$$

If $\{\rho_m(1), \ldots, \rho_m(K)\}$ is not a permutation of $\{1, \ldots, K\}$ (i.e., if (3.57) is violated for a considerable fraction of the MCMC draws), this is an indication that the mixture is overfitting the number of components, a problem that is discussed in Subsection 4.2.2.

Further Approaches Toward Relabeling the MCMC Draws

Various authors found other ways of relabeling the MCMC draws useful. Stephens (1997b) suggested relabeling the MCMC output so that the estimated marginal posterior distributions of the parameters of interest are as

close to unimodality as possible. Stephens (2000b) tackles the whole relabeling problem from a decision-theoretic viewpoint and shows that the relabeling strategies studied in Stephens (1997b) and Celeux (1998) may be viewed as an attempt to minimize the posterior expectation of a certain loss function.