Statistics and Computing (1997) **7**, 1–10

Inference in model-based cluster analysis

HALIMA BENSMAIL[1], GILLES CELEUX[2], ADRIAN E. RAFTERY[1] and CHRISTIAN P. ROBERT[3]

[1]*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA. E-mail: bensmail/raftery@stat.washington.edu*
[2]*INRIA Rhône-Alpes, ZIRST, 655 Avenue de l'Europe, 38330 Montbonnet Saint-Martin, France. E-mail: Gilles.celeux@imag.fr*
[3]*CREST, INSEE, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France. E-mail: robert@ensae.fr*

Received September 1995 and accepted October 1996

A new approach to cluster analysis has been introduced based on parsimonious geometric modelling of the within-group covariance matrices in a mixture of multivariate normal distributions, using hierarchical agglomeration and iterative relocation. It works well and is widely used via the MCLUST software available in S-PLUS and StatLib. However, it has several limitations: there is no assessment of the uncertainty about the classification, the partition can be suboptimal, parameter estimates are biased, the shape matrix has to be specified by the user, prior group probabilities are assumed to be equal, the method for choosing the number of groups is based on a crude approximation, and no formal way of choosing between the various possible models is included. Here, we propose a new approach which overcomes all these difficulties. It consists of exact Bayesian inference via Gibbs sampling, and the calculation of Bayes factors (for choosing the model and the number of groups) from the output using the Laplace–Metropolis estimator. It works well in several real and simulated examples.

Keywords: Bayes factor, eigenvalue decomposition, Gaussian mixture, Gibbs sampler

1. Introduction

Banfield and Raftery (1993)—hereafter BR—building on work of Murtagh and Raftery (1984), introduced a new approach to cluster analysis based on a mixture of multivariate normal distributions, where the covariance matrices are modelled parsimoniously in a geometrically interpretable way. The general finite normal mixture distribution for n data points $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in p dimensions with K groups is

$$p(\mathbf{x}|\pi, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i|\mu_k, \Sigma_k) \qquad (1)$$

where $\phi(\cdot|\mu, \Sigma)$ is the multivariate normal density with mean μ and covariance matrix Σ, $\pi = (\pi_1, \ldots, \pi_K)$ is a vector of group mixing proportions such that $\pi_k \geq 0$ and $\sum \pi_k = 1$, and $\theta = (\mu_1, \ldots, \mu_K; \Sigma_1, \ldots, \Sigma_K)$.

The BR approach is based on a variant of the standard spectral decomposition of Σ_k, namely

$$\Sigma_k = \lambda_k D_k A_k D_k^t \qquad (2)$$

where λ_k is a scalar, $A_k = \mathrm{diag}\{1, a_{k2}, \ldots, a_{kp}\}$ where $1 \geq a_{k2} \geq \ldots a_{kp} > 0$, and D_k is an orthogonal matrix for each $k = 1, \ldots, K$. Each factor in Equation 2 has a geometric interpretation: λ_k controls the *volume* of the kth group, A_k its *shape* and D_k its *orientation*. By imposing constraints such as $\lambda_k = \lambda$ and $A_k = A\ \forall k$ (i.e. each group has the same volume and shape, but different orientations), one obtains different models, which lead in turn to different clustering algorithms. The models considered here, including parsimonious spherically shaped ones, are listed in Table 1.

BR developed algorithms aimed at maximizing the *classification likelihood*

$$p(\mathbf{x}|\theta, \nu) = \prod_{i=1}^{n} \phi(\mathbf{x}_i|\mu_{\nu_i}, \Sigma_{\nu_i}) \qquad (3)$$

as a function of both θ and ν, where ν is the vector of group memberships, namely $\nu_i = k$ if $\mathbf{x}_i$ belongs to the kth group. These algorithms use hierarchical agglomeration and iterative relocation. They worked well on several real and simulated data sets and are now fairly widely used. They are implemented in the software MCLUST, which is both an S-PLUS function and a Fortran program available from StatLib. (The Fortran version is at *http://lib.stat.cmu.edu/*

Table 1. *Clustering models. The entries indicate whether the feature of interest (shape, orientation or volume) is the same for each group or not*

Model	Σ_k	Shape	Orientation	Volume
1.	λI	Spherical	None	Same
2.	$\lambda_k I$	Spherical	None	Different
3.	Σ	Same	Same	Same
4.	$\lambda_k \Sigma$	Same	Same	Different
5.	$\lambda D_k A D_k^t$	Same	Different	Same
6.	$\lambda_k D_k A D_k^t$	Same	Different	Different
7.	$\lambda_k D A_k D^t$	Different	Same	Different
8.	Σ_k	Different	Different	Different

general/mclust, and can also be obtained by e-mail by sending the message 'send mclust from general' to the address *statlib@lib.stat.cmu.edu*.)

However, the BR algorithms have limitations, several of which are common to all agglomerative hierarchical clustering methods:

1. They give only point classifications of each individual and produce no assessment of the associated uncertainty.
2. They tend to yield partitions that are suboptimal (even if often good). This is due to the use of hierarchical agglomeration.
3. Estimates of the model parameters θ based on the estimated partition tend to be biased (Marriott, 1975).
4. They assume the mixing proportions π_k in Equation 1 to be equal.
5. The algorithms based on models 5 and 6 in Table 1 require the shape matrix A to be specified in advance by the user. This can be useful but it limits the general usefulness of the model.
6. To choose K, the number of groups, BR proposed an approximation to the posterior probabilities based on a quantity called the AWE (Approximate Weight of Evidence). While this has worked fairly well in practice, it is quite crude.
7. BR proposed no formal way of choosing among the possible models; this must be done by the user.

We know of no way of fully overcoming limitations 1, 6 and 7 other than the fully Bayesian analysis that we develop here. Other possible ways of overcoming 2–5 are discussed in Section 4.

Here we present a new approach to clustering based on the models in Table 1; it consists of fully Bayesian inference for these models via Gibbs sampling. This overcomes all the limitations mentioned. A recently proposed way of calculating Bayes factors from posterior simulation output, the Laplace–Metropolis estimator (Lewis and Raftery, 1997; Raftery, 1996a), is used to choose the model and determine the number of groups simultaneously. In Section 2.1 we describe Bayesian estimation for the models of Table 1 using Markov chain Monte Carlo (MCMC) methods. In Section 2.2 we outline how Bayes factors can be calculated from the MCMC output and used to determine the appropriate model and the number of groups. In Section 3 we show the methods at work on real and simulated data sets.

2. Bayesian estimation of the Banfield–Raftery clustering models using the Gibbs sampler

2.1. *Estimation*

We assume that data $\mathbf{x}_i$, $i = 1, \ldots, n; \mathbf{x}_i \in \mathbb{R}^p$ to be classified arise from a random vector with density (1), and that the corresponding classification variables ν_i are unobserved. We are concerned with Bayesian inference about the model parameters θ, π and the classification indicators ν_i. MCMC methods provide a general recipe for Bayesian analysis of mixtures. For instance, Lavine and West (1992) and Soubiran *et al.* (1991) have used the Gibbs sampler for estimating the parameters of a multivariate Gaussian mixture assuming no specific characteristics for the component variance matrices; Diebolt and Robert (1994) have considered the Gibbs sampler and the Data Augmentation method of Tanner and Wong (1987) for general univariate Gaussian mixtures and proved that both algorithms converge in distribution to the true posterior distribution of the mixture parameters.

Like these authors, we use conjugate priors for the parameters π and θ of the mixture model. The prior distribution of the mixing proportions is a Dirichlet distribution $\pi \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K)$; and the prior distributions of the means μ_k of the mixture components conditionally on the variance matrices Σ_k are Gaussian: $\mu_k|\Sigma_k \sim \mathcal{N}(\xi_k, \frac{1}{\tau_k}\Sigma_k)$. The conjugate prior distribution of the variance matrices depends on the model, and will be given for each model in turn.

We estimate the models in Table 1 by simulating from the joint posterior distribution of π, θ and ν using the Gibbs sampler (Smith and Roberts, 1993). In our case, this consists of the following steps:

1. Simulate the classification variables ν_i according to their posterior probabilities $(t_{ik}, k = 1, \ldots, K)$ conditional on π and θ, namely

$$t_{ik} = \frac{\pi_k \phi(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_j \pi_j \phi(\mathbf{x}_i|\mu_j, \Sigma_j)}; \qquad i = 1, \ldots, n$$

2. Simulate the vector π of mixing proportions according to its posterior distribution conditional on the ν_i's.
3. Simulate the parameters θ of the model according to their posterior distributions conditional on the ν_i's. Details are given in the Appendix.

The validity of this procedure, namely the fact that the Markov chain associated with the algorithm converges in distribution to the true posterior distribution of θ, was shown by Diebolt and Robert (1994) in the context of

one-dimensional normal mixtures. Their proof is based on a *duality principle*, which uses the finite space $\{1, \ldots, K\}$, and is thus geometrically convergent and φ-mixing. These properties transfer automatically to the sequence of values of θ and π, and important properties such as the central limit theorem or the law of the iterated logarithm are then satisfied (Diebolt and Robert, 1994, Robert, 1993).

The same results apply here, the only difference being the more complex simulation structure imposed by the variance assumptions. Steps 1 and 2 do not depend on the model considered. Step 1 is straightforward, and Step 2 consists of simulating π from its conditional posterior distribution, namely $\pi \sim \mathcal{D}\ (\alpha_1 + \sum_{i=1}^{n} \mathbf{I}\{\nu_i = 1\}, \ldots, \alpha_K + \sum_{i=1}^{n} \mathbf{I}\{\nu_i = K\})$. Step 3 is not the same for the different models of Table 1, and is described in the Appendix for each model in turn.

2.2. *Choosing the number of groups and the model by Bayes factors*

BR left the choice of model to the user, and they based the choice of number of clusters on the AWE criterion, which is a crude approximation to twice the log Bayes factor for that number of clusters versus just one cluster.

Here we develop a way of choosing both the model and the number of groups in one step, using a more accurate approximation to the Bayes factor than that of BR. We compute approximate Bayes factors from the Gibbs sampler output using the *Laplace–Metropolis estimator* of Raftery (1996a). This was shown to give accurate results by Lewis and Raftery (1997).

In what follows, the word 'model' refers to a combination of one of the models in Table 1 with a specified number of clusters. The Bayes factor, B_{10} for a model M_1 against another model M_0 given data D is the ratio of posterior to prior odds, namely

$$B_{10} = \mathrm{pr}(D|M_1)/\mathrm{pr}(D|M_0) \tag{4}$$

the ratio of the integrated likelihoods. In Equation 4,

$$\mathrm{pr}(D|M_k) = \int \mathrm{pr}(D|\theta_k, M_k)\mathrm{pr}(\theta_k|M_k)\mathrm{d}\theta_k \tag{5}$$

where θ_k is the vector of parameters of M_k, and $\mathrm{pr}(\theta_k|M_k)$ is its prior density ($k = 0, 1$); this is called the *integrated likelihood* of model M_k. For a review of Bayes factors, their calculation and interpretation, see Kass and Raftery (1995). Bayesian model selection is based on Bayes factors, whose key ingredient is the integrated likelihood of a model. Our main computational challenge is thus to approximate the integrated likelihood using the Gibbs sampler output.

We do this using the Laplace–Metropolis estimator of the integrated likelihood (Raftery, 1996a; Lewis and Raftery, 1997). The Laplace method for integrals is based on a Taylor series expansion of the real-valued function $f(u)$ of the d-dimensional vector u, and yields the approximation

$$\int \mathrm{e}^{f(u)}\mathrm{d}u \approx (2\pi)^{d/2}|A|^{\frac{1}{2}}\exp\{f(u^*)\} \tag{6}$$

where u^* is the value of u at which f attains its maximum, and A is minus the inverse Hessian of f evaluated at u^*. When applied to Equation 5 it yields

$$p(D) \approx (2\pi)^{d/2}|\Psi|^{\frac{1}{2}}\,\mathrm{pr}(D|\tilde{\theta})\mathrm{pr}(\tilde{\theta}) \tag{7}$$

where d is the dimension of θ, $\tilde{\theta}$ is the posterior mode of θ, and Ψ is minus the inverse Hessian of $h(\theta) = \log\{\mathrm{pr}(D|\theta)\mathrm{pr}(\theta)\}$, evaluated at $\theta = \tilde{\theta}$. Arguments similar to those in the Appendix of Tierney and Kadane (1986) show that in regular statistical models the relative error in Equation 7, and hence in the resulting approximation to B_{10}, is $O(n^{-1})$, where n is sample size.

While the Laplace method is often very accurate, it is not directly applicable here because the derivatives it requires are not easily available. The idea of the Laplace–Metropolis estimator is to get around the limitations of the Laplace method by using posterior simulation to *estimate* the quantities it needs. The Laplace method requires the posterior mode, $\tilde{\theta}$, and $|\Psi|$. The Laplace–Metropolis estimator estimates these from the Gibbs sampler output using robust location and scale estimators. The likelihood at the approximate posterior mode is

$$\mathrm{pr}(D|\tilde{\theta}) = \prod_{i=1}^{n}\sum_{k=1}^{K}\tilde{\pi}_k\phi(x_i|\tilde{\mu}_k, \tilde{\Sigma}_k)$$

These quantities are then substituted into Equation 7 to obtain the integrated likelihood, and Bayes factors are computed by taking ratios of integrated likelihoods, as in Equation 4.

3. Examples

We now present four examples to illustrate the ability of our methods to overcome the limitations of other methods described in Section 1. The first and fourth examples use simulated data and the second and third examples are based on real data sets.

For each example, we consider only the models $[\lambda I]$, $[\lambda_k I]$, $[\Sigma]$, and $[\lambda_k \Sigma]$. Models $[\lambda I]$ and $[\Sigma]$ are probably the most used Gaussian mixture models for clustering data (e.g. McLachlan and Basford, 1988), and the generalizations of these, $[\lambda_k I]$ and $[\lambda_k \Sigma]$, to allow for different volumes have proved to be powerful in many practical situations (Celeux and Govaert, 1995).

Our priors are chosen from among the conjugate priors of Section 2.1 so as to be fairly flat in the region where the likelihood is substantial and not much greater elsewhere. Thus they satisfy the 'Principle of Stable Estimation' (Edwards *et al.*, 1963), and so it could be expected that the

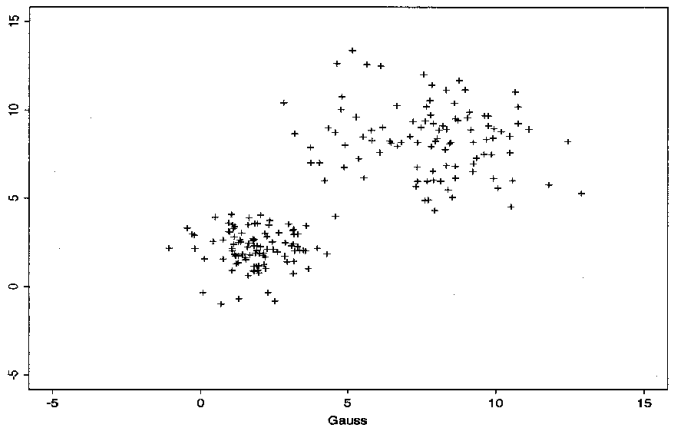

Fig. 1. *Example 1: simulated data*

results would be relatively insensitive to reasonable changes in the prior; we also checked this empirically for each example.

We used $\pi_k = 1/K$, $\xi_k = \bar{x}$, $\tau_k = 1$, $m_k = m_0 = 5$, $s_k^2 = s_0^2 = \hat{\sigma}^2$, and $\Psi_0 = S$, for $k = 1, \ldots, K$, where $\bar{x}$ and S are the empirical mean vector and variance matrix of the whole data set, and $\hat{\sigma}^2$ is the greatest eigenvalue of S. (The other notation used is defined in the Appendix.) The amount of information contained in this prior is similar to that contained in a typical single observation. Thus the prior may be viewed as comparable to the true prior of a person with some, but rather little, information. Similar priors have been used for generalized linear models by Raftery (1996a) and for linear regression models by Raftery *et al.* (1996). In each example we assessed the sensitivity of the results to changes in this prior and found it to be small; some of the sensitivity results for the first example are included below.

3.1. *Example 1: simulated data*

We simulated 200 points from a bivariate two-component Gaussian mixture with equal proportions, mean vectors $\mu_1^t = (8, 8)$, $\mu_2^t = (2, 2)$, and variance matrices $\Sigma_1 = 4I$, $\Sigma_2 = I$; the data are shown in Fig. 1. The first 600 iterations from the Gibbs sampler output for the model $[\lambda_k I]$ with two groups are shown in Fig. 2. Convergence was immediate and successive draws were almost independent; similar results were obtained for other starting values. We used 1500 iterations, estimated by the `gibbsit` program to be enough to estimate the cumulative distribution function at the 0.025 and 0.975 quantiles to within ± 0.01 for all the parameters (Raftery and Lewis, 1993, 1996).

The model comparison results are shown in Table 2. The correct model, $[\lambda_k I]$, and the correct number of groups, 2, are strongly favoured. The posterior means of the parameters for the preferred model are $\mu_1 = (7.8, 8.3)$, $\mu_2 = (1.9, 2.2)$, $\lambda_1 = 4.2$, $\lambda_2 = 1.1$, which are close to the true values. The marginal posterior distribution is summarized in Fig. 3, which shows the posterior distribution of the principal circles of the two groups.

Sensitivity to the prior distribution is investigated in Table 3. A new Gibbs sampler run was done for each choice

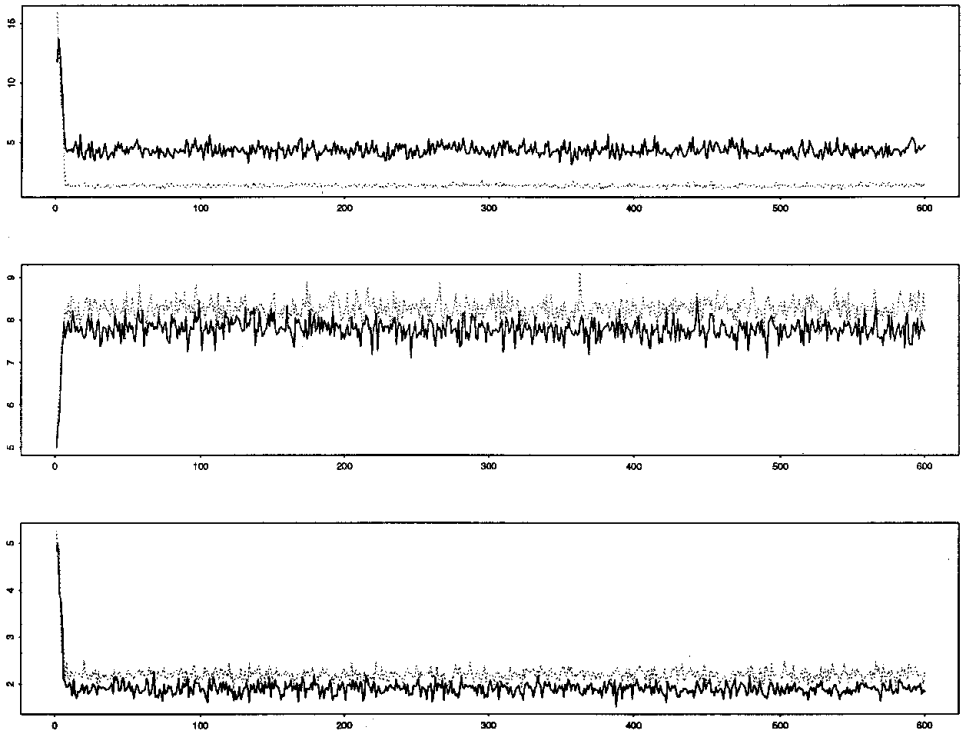

Fig. 2. *Example 1: Time series plot of the first 600 Gibbs sampler iterations: (a) volume parameters; (b) mean for group 1; (c) mean for group 2*

Table 2. *Example 1: approximate log integrated likelihoods*

No. of groups	$[\lambda I]$	$[\lambda_k I]$	$[\Sigma]$	$[\lambda_k \Sigma]$
1	−1064	−1067	−991	−991
2	−907	**−861**	−915	−869
3	−923	−883	−931	−894
4	−901	−875	−909	−880

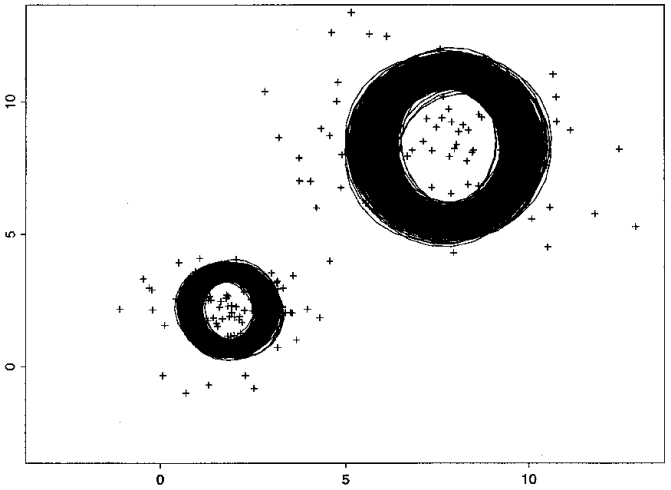

Fig. 3. *Example 1: posterior distribution of principal circles for the $[\lambda_k I]$ model with two groups. There is one circle for each Gibbs sampler iteration and each group, with centre μ_k and radius λ_k*

of prior parameters, and so the differences in Table 3 are due to both true sensitivity and Monte Carlo variation; the true sensitivity is thus likely to be smaller. The estimation results are quite insensitive. The testing results are somewhat more sensitive, which is to be expected (Kass and Raftery, 1995), but the overall conclusions remain the same over all combinations of prior parameters considered.

Perhaps the greatest advantage of the present approach is that it fully assesses uncertainty about group membership, rather than merely giving a single 'best' partition. In Fig. 4, this is summarized by showing the uncertainty for each point, measured by $U_i = \min_{k=1,\ldots,K} (1 - \widehat{\Pr}[\nu_i = k|D])$. When it is clear that x_i belongs to the kth group, then $(1 - \widehat{\Pr}[\nu_i = k|D])$ is small, and so U_i is also small. In these

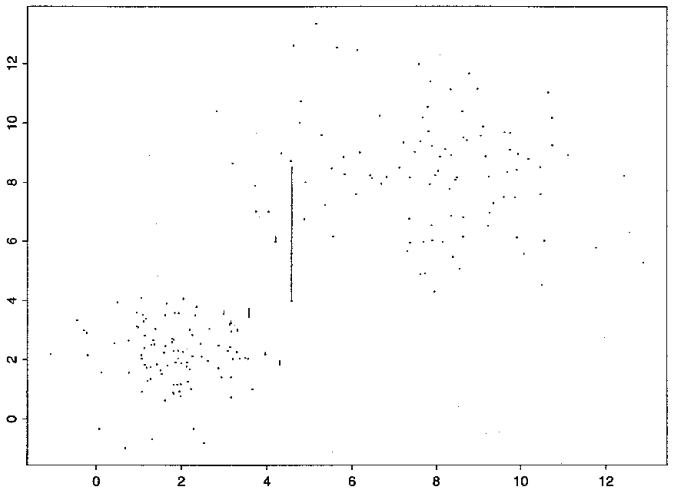

Fig. 4. *Example 1: uncertainty plot. At each point a vertical line of length proportional to $U_i = min_{k=1,2}(1-\widehat{\Pr}[\nu_i = k|D])$ is plotted. The longest line is of length 0.5*

data, U_i is large for only one point, no. 55, the one that lies on the boundary of the two groups, for which $\widehat{\Pr}[\nu_{55} = k|D] \approx 0.50$ $(k = 1, 2)$.

3.2. *Example 2: butterfly classification*

Figure 5 shows four wing measurements of a butterfly. Here we analyse data on two of these measurements, z_3 and z_4, for 23 butterflies, shown in Fig. 6, from Celeux and Robert (1993). The aim is to decide how many species are represented in this group of insects, and to classify them.

Table 4 shows that model $[\lambda_k \Sigma]$ with four groups is favoured quite strongly over the alternatives. The posterior means of the parameters are: $\mu_1 = (24.7, 19.7)$, $\mu_2 = (26.2, 16.2)$, $\mu_3 = (21.5, 21.5)$, $\mu_4 = (23.1, 31.7)$, $\sigma_{11} = 6.4$, $\sigma_{12} = 4.2$, $\sigma_{22} = 4.5$, $\lambda_1 = 1$, $\lambda_2 = 0.55$, $\lambda_3 = 0.55$, $\lambda_4 = 0.13$. The most likely group memberships a posteriori are shown in Fig. 7, along with the associated uncertainties.

All the butterflies are classified with confidence except numbers 4 and 15 which are close to the boundary between groups 1 and 3. Group 4 consists of just one butterfly, which is clearly out on its own. The correct classification

Table 3. *Example 1: sensitivity of selected results to changes in the prior hyperparameters for 1500 simulations. Log B_{23} is the log Bayes factor for two groups against three groups.* **x** *represents the data, $\bar{x}_o$ is the mean of the two optimal partitions, $\bar{x}_{o1} = (7.78, 8.28)$, $\bar{x}_{o2} = (1.85, 2.16)$, $\bar{x}$ is the classes global mean $\bar{x} = (4.9, 5.2)$, and $\sigma = 20.6$*

Perturbation	$\log B_{23}$	$\Pr[\nu_{55} = 1\|\mathbf{x}]$	$E[\lambda_1\|\mathbf{x}]$	$E[\mu_{11}\|\mathbf{x}]$
$\xi = \bar{x}$, $s_0^2 = \sigma^2$, $m_0 = 5$, $\tau = 1$	**22**	**0.50**	**4.20**	**7.80**
$\xi = \bar{x}$, $s_0^2 = \sigma^2$, $m_0 = 10$, $\tau = 1$	27	0.50	4.16	7.78
$\xi = \bar{x}$, $s_0^2 = \sigma^2/4$, $m_0 = 5$, $\tau = 1$	25	0.50	4.18	7.78
$\xi = \bar{x}$, $s_0^2 = 4\sigma^2$, $m_0 = 5$, $\tau = 1$	25	0.50	4.18	7.78
$\xi = \bar{x}$, $s_0^2 = \sigma^2$, $m_0 = 5$, $\tau = 2$	40	0.44	4.23	7.76
$\xi = \bar{x}_o$, $s_0^2 = \sigma^2$, $m_0 = 5$, $\tau = 1$	22	0.46	4.10	7.80

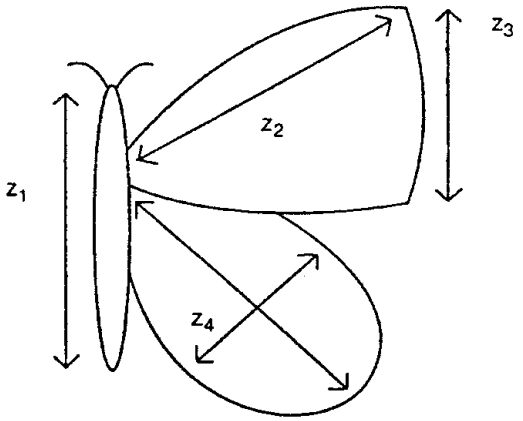

Fig. 5. *Example 2: butterfly measurements*

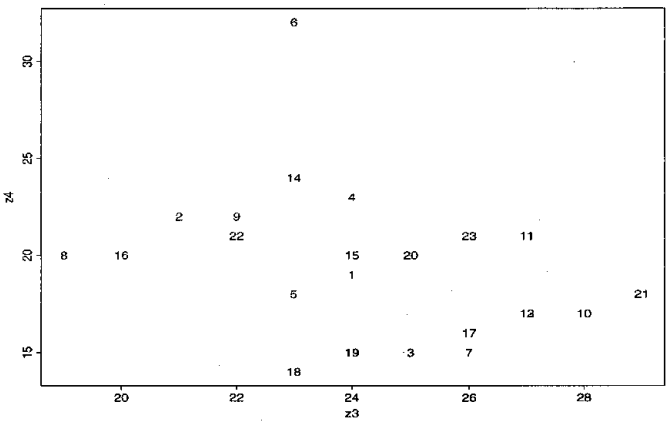

Fig. 6. *Example 2: butterfly data. Values of* (z_3, z_4) *for 23 butterflies*

Table 4. *Example 2: approximate log integrated likelihoods*

No. of groups	$[\lambda I]$	$[\lambda_k I]$	$[\Sigma]$	$[\lambda_k \Sigma]$
1	−124	−121	−122	−129
2	−123	−114	−121	−119
3	−122	−110	−120	−114
4	−119	−117	−107	**−104**

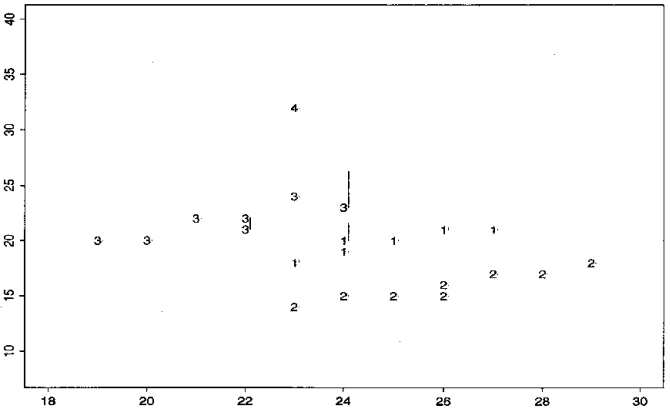

Fig. 7. *Example 2: estimated group memberships and uncertainty plot for the butterfly data*

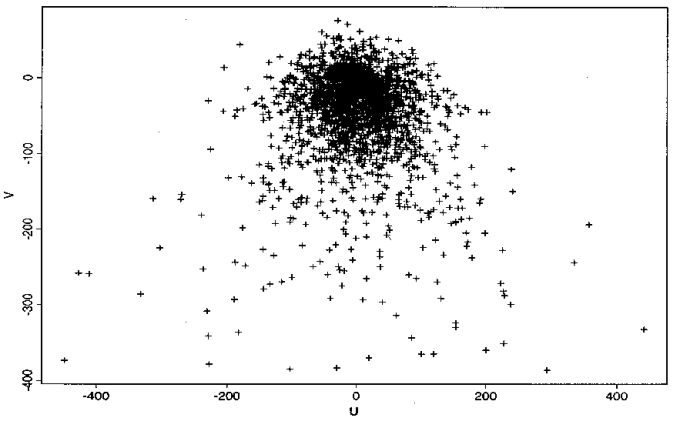

Fig. 8. *Example 3: kinematic stellar data. Radial (U) and rotational (V) velocities for 2370 stars in the Galaxy*

is known, and is exactly equal to the optimal classification found by our methods (Celeux and Robert, 1993).

3.3. *Example 3: kinematic stellar data*

Until fairly recently, it was believed that the Galaxy consists of two stellar populations, the disk and the halo. More recently, it has been hypothesized that there are in fact three stellar populations, the old (or thin) disk, the thick disk, and the halo, distinguished by their spatial distributions, their velocities, and their metallicities. These hypotheses have different implications for theories of the formation of the Galaxy. Some of the evidence for deciding whether there are two or three populations is shown in Fig. 8, which shows radial and rotational velocities for $n = 2370$ stars, from Soubiran (1993).

Table 5 shows that model $[\lambda_k \Sigma]$ is preferred and that there is strong evidence for three groups as against two. The balance of astronomical opinion has also tilted towards this conclusion, but based on much more information than just the velocities used here, including star positions and metallicities (Soubiran, 1993). It is impressive that such a strong conclusion can be reached with the present methods using only a relatively small part of the total available information.

The posterior means of the parameters for the preferred model are: $\lambda_1 = 1$, $\lambda_2 = 11.9$, $\lambda_3 = 1.8$; $\mu_1 = (-9.8, -10.1)$, $\mu_2 = (-2.4, -101.1)$, $\mu_3 = (15.8, -37.8)$; $\sigma_{11} = 1046$, $\sigma_{12} = -23$, $\sigma_{22} = 552$. The corresponding partition is shown in Fig. 9.

Table 5. Example 3: approximate log integrated likelihoods

No. of groups	$[\lambda I]$	$[\lambda_k I]$	$[\Sigma]$	$[\lambda_k \Sigma]$
1	−26397	−26416	−26421	−26418
2	−26566	−26283	−26424	−26212
3	−27120	−26010	−27162	**−25668**
4	−28440	−26244	−27161	−25705

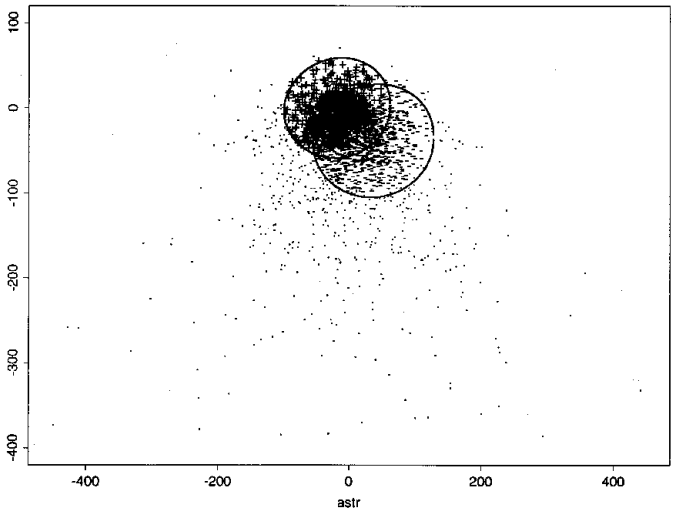

Fig. 9. *Example 3: optimal partition for Model 4 [$\lambda_k\Sigma$]. The three groups are represented by +, − and ·*

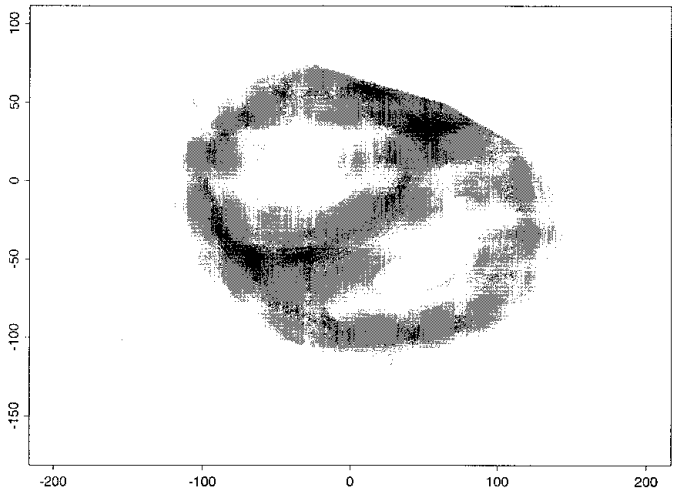

Fig. 10. *Example 3: uncertainty plot*

The uncertainty plot is shown in Fig. 10. The areas of high uncertainty are those on the boundaries between any two of the three groups. The greatest uncertainty is in the two small areas where all three groups intersect.

3.4. *Example 4: simulated data in 20 dimensions*

We simulated 200 points as in Example 1 but from a 20-dimensional two-component Gaussian mixture with equal proportions, mean vectors $\mu_1 = (8, 8, 0, \ldots, 0)$, $\mu_2 = (2, 2, 0, \ldots, 0)$, and variance matrices $\Sigma_1 = 4I$, $\Sigma_2 = I$. The first 600 iterations from the Gibbs sampler output for the model $[\lambda_k I]$ with two groups are shown in Fig. 11. Convergence was almost immediate and successive draws were almost independent; similar results were obtained from other starting values. We ran the Gibbs sampler for 1000 iterations in all; this was ample to achieve the required level of accuracy according to the `gibbsit` method of Raftery and Lewis (1993, 1996).

The model comparison results are shown in Table 6. The correct model, $[\lambda_k I]$, and the correct number of groups, 2,

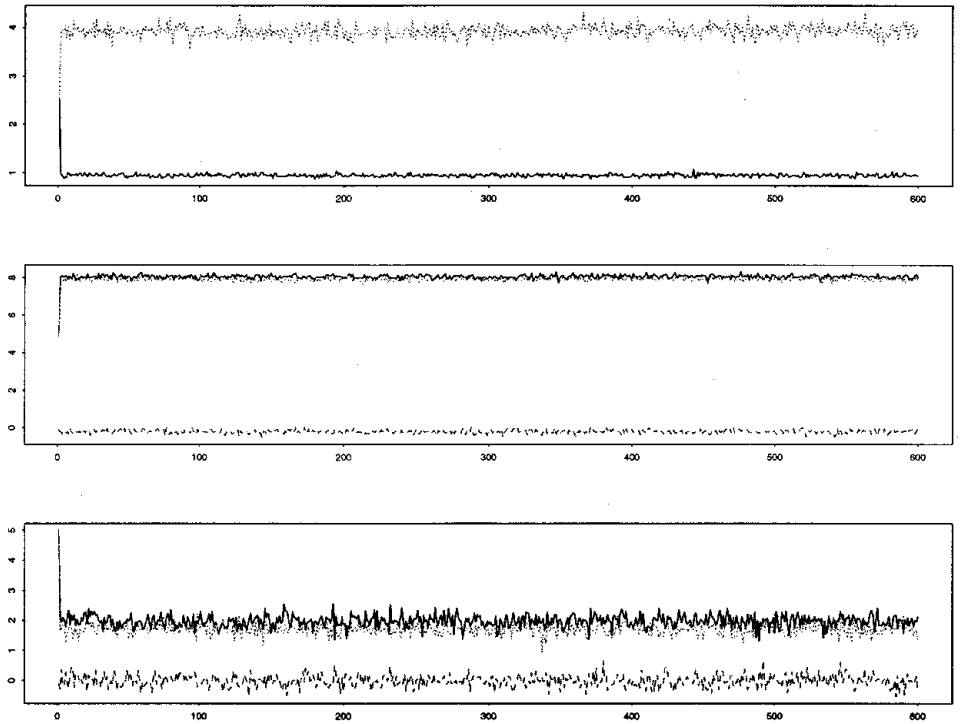

Fig. 11. *Example 4: time series plot of the first 600 Gibbs sampler iterations for model $[\lambda_k I]$ with two groups: (a) volume parameter, λ_1 and λ_2; (b) first three coordinates of the mean for group one; (c) first three coordinates of the mean for group two*

Table 6. *Example 4: approximate log integrated likelihoods*

No. of groups	$[\lambda I]$	$[\lambda_k I]$	$[\Sigma]$	$[\lambda_k \Sigma]$
1	−8954	−7525	−8172	−7518
2	−8322	**−6599**	−7129	−7126
3	−8474	−6762	−7491	−7277
4	−8861	−6933	−7762	−7523

are strongly favoured. The posterior means of the parameters for the preferred model are:

$\mu_1 = (8, 7.9, -0.2, -0.1, 0, -0.1, 0, -0.1, 0, 0, 0.1, 0.1, -0.1, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, -0.1)$,

$\mu_2 = (1.9, 1.8, 0, 0.2, 0, 0.1, -0.1, -0.1, 0, 0.2, 0.1, 0.1, 0.1, 0, 0.1, 0.2, 0.2, 0.2, 0, 0.1)$,

$\lambda_1 = 3.92$, $\lambda_2 = 0.94$, which are close to the true values.

4. Discussion

We have presented a fully Bayesian analysis of the model-based clustering methodology of Banfield and Raftery (1993), which overcomes many of the limitations of that approach. It appears to work well in several examples.

Alternative frequentist approaches, which might be easier to implement, consist of maximizing the full (mixture) likelihood using the EM algorithm or of maximizing the classification likelihood using the Classification EM (CEM) algorithm. Celeux and Govaert (1995) considered those approaches to the full range of clustering models derived from the eigenvalue decomposition of the group variance matrices including those considered here. They have shown in particular how it is possible to find the maximum likelihood estimate of the shape matrix A. Both approaches could overcome limitations 2, 4 and 5 of Section 1, and the mixture maximum likelihood approach could also overcome limitation 3. They would overcome difficulty 1 only partly: they do provide an estimate of the uncertainty about group membership, but this assessment is incomplete because it does not take account of uncertainty about π and θ. They do not overcome limitations 6 and 7.

In our examples, we explicitly considered only models 1–4 of Table 1; these were sufficient for the data we considered. However, more generally it would be useful to consider the other models, proceeding in the same way and using the results of Section 2.1.

Acknowledgements

This research was supported by the Office of Naval Research grants nos. N-00014-91-J-1074 and N-00014-96-1-0192 and by INRIA, France. The authors are grateful to Jean Diebolt and Christian Posse for helpful discussions, and to Caroline Soubiran for providing the data in Example 3.

References

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non Gaussian Clustering. *Biometrics*, **49**, 803–21.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–93.

Celeux, G. and Robert, C. (1993) Une histoire de discrétisation (avec commentaires). *La Revue de Modulad*, **11**, 7–44.

Diebolt, J. and Robert, C. P. (1994) Bayesian estimation of finite mixture distributions. *Journal of the Royal Statistical Society*, Series B, **56**, 363–75.

Edwards, W., Lindman, H. and Savage, L. J. (1963) Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–95.

Lavine, M. and West, M. (1992) A Bayesian method for classification and discrimination. *The Canadian Journal of Statistics*, **20**, 451–61.

Lewis, S. M. and Raftery, A. E. (1997) Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, to appear.

Marriott, F. H. C. (1975) Separating mixtures of normal distributions. *Biometrics*, **31**, 767–9.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models, Inference and Applications to Clustering*. New York, Marcel Dekker.

Murtagh, F. and Raftery, A. E. (1984) Fitting straight lines to point patterns. *Pattern Recognition*, **17**, 479–83.

Raftery, A. E. (1996a) Hypothesis testing and model selection via posterior simulation. In *Practical Markov Chain Monte Carlo* (W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds), London: Chapman and Hall, pp. 163–88.

Raftery, A. E. (1996b) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, **83**, 251–66.

Raftery, A. E. and Lewis, S. M. (1993) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds), Oxford University Press, pp. 763–73.

Raftery, A. E. and Lewis, S. M. (1996) Implementing MCMC. In *Practical Markov Chain Monte Carlo* (W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds), London: Chapman and Hall, pp. 115–30.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1996) Accounting for model uncertainty in linear regression. *Journal of the American Statistical Association*, **91**, to appear.

Robert, C. P. (1993) Convergence assessment of MCMC methods. Rapport Technique CREST, INSEE, Paris.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, Series B, **55**, 3–23.

Soubiran, C. (1993) Kinematics of the Galaxy's stellar population from a proper motion survey. *Astronomy and Astrophysics*, **274**, 181–8.

Soubiran, C., Celeux, G., Diebolt, J. and Robert, C. P. (1991) Analyse de mélanges gaussiens pour de petits échantillons: application à la cinématique stellaire. *Revue de Statistique Appliquée*, **39**, 3, 17–36.

Tanner, M. and Wong, W. (1987) The calculation of posterior distribution by data augmentation (with Discussion). *Journal of the American Statistical Association*, **82**, 528–50.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–6.

Appendix: Gibbs sampling for the clustering models

We now give details of Step 3 of Gibbs sampling for each of the clustering models used in the examples. Given a classification vector $\nu = (\nu_1, \ldots, \nu_n)$, we use the notation

$$\begin{aligned} n_k &= \sum_i \mathbf{I}\{\nu_i = k\}, \qquad \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i;\, \nu_i = k} \mathbf{x}_i, \\ W_k &= \sum_{i;\, \nu_i = k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t \end{aligned} \tag{8}$$

the componentwise statistics of location and scale $(k = 1, \ldots, K)$.

(a) **Model** $[\boldsymbol{\lambda I}]$ Here the scale parameter λ is common to all components of the mixture. We assume that the prior distribution on the parameters is conjugate, namely that

$$\begin{aligned} &\mu_k|\lambda \sim \mathcal{N}_p(\xi_k, \lambda I_p/\tau_k) \qquad (k = 1, \ldots, K) \\ &\lambda \sim \mathcal{I}g(m_0/2, s_0^2/2) \end{aligned} \tag{9}$$

where $\lambda \sim \mathcal{I}g(\frac{1}{2}\, r, \frac{1}{2}\, \rho)$ means that λ has the inverted gamma distribution

$$\mathrm{pr}(\lambda) = \frac{\lambda^{-(r/2)-1} \exp(-\rho/2\lambda)}{\Gamma(r/2)\rho^{-r/2}/2}$$

The posterior distribution on $(\mu_1, \ldots, \mu_K, \lambda)$ is therefore a convolution of normal distributions on the μ_ks and of an inverse gamma distribution on λ.

The Gibbs components of Step 3 are then:

3.1 For $k = 1, \ldots, K$, simulate

$$\mu_k|\lambda, \nu \sim \mathcal{N}_p\left(\bar{\xi}_k, \frac{\lambda}{n_k + \tau_k} I_p\right)$$

with $\bar{\xi}_k = (n_k \bar{\mathbf{x}}_k + \tau_k \xi_k)/(n_k + \tau_k)$

3.2 Simulate

$$\lambda|\nu \sim \mathcal{I}_g\left(\frac{m_0 + n}{2}, \right.$$

$$\left. \frac{1}{2}\left\{ s_0^2 + \sum_k \mathrm{tr}(W_k) + \sum_k \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \xi_k)^t (\bar{\mathbf{x}}_k - \xi_k) \right\}\right)$$

(b) **Model** $[\boldsymbol{\lambda_k I}]$ When the variance scales are different, the prior distributions are similar for all components

$$\begin{aligned} &\mu_k|\lambda_k \sim \mathcal{N}_p(\xi_k, \lambda_k I_p/\tau_k) \text{ and } \lambda_k \sim \mathcal{I}g(m_k/2, s_k^2/2) \\ &(k = 1, \ldots, K) \end{aligned} \tag{10}$$

We recover the case treated in Diebolt and Robert (1994), namely that in which the parameters (μ_k, λ_k) are generated separately:

3.1 For $k = 1, \ldots, K$, simulate

$$\mu_k|\lambda_k, \nu \sim \mathcal{N}_p\left(\bar{\xi}_k, \frac{\lambda_k}{n_k + \tau_k} I_p\right)$$

with $\bar{\xi}_k = (n_k \bar{\mathbf{x}}_k + \tau_k \xi_k)/(n_k + \tau_k)$.

3.2 Simulate

$$\lambda_k|\nu \sim \mathcal{I}g\left(\frac{m_k + n_k p}{2}, \right.$$

$$\left. \frac{1}{2}\left\{ s_k^2 + \mathrm{tr}(W_k) + \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \xi_k)^t (\bar{\mathbf{x}}_k - \xi_k) \right\}\right)$$

(c) **Model** $[\boldsymbol{\Sigma}]$ There is no need to consider the eigenvalue decomposition of the covariance matrix Σ, and the prior distribution is given by

$$\mu_k|\Sigma \sim \mathcal{N}_p(\xi_k, \Sigma/\tau_k) \quad (k = 1, \ldots, K), \quad \Sigma \sim \mathcal{W}_p^{-1}(m_0, \Psi_0) \tag{11}$$

where $\Sigma \sim \mathcal{W}_p^{-1}(m, \Psi)$ means that Σ has the inverse Wishart distribution

$$\mathrm{pr}(\Sigma) \propto |\Sigma|^{-(m+p+1)/2} \exp -\{\mathrm{tr}(\Psi\Sigma^{-1})/2\}$$

Step 3 of Gibbs sampling is then decomposed as follows:

3.1 For $k = 1, \ldots, K$, simulate

$$\mu_k|\Sigma, \nu \sim \mathcal{N}_p\left(\bar{\xi}_k, \frac{1}{n_k + \tau_k} \Sigma\right)$$

with $\bar{\xi}_k = (n_k \bar{\mathbf{x}}_k + \tau_k \xi_k)/(n_k + \tau_k)$.

3.2 Simulate

$$\Sigma|\nu \sim \mathcal{W}_p^{-1}\left(m_0 + n, \right.$$

$$\left. \Psi_0 + \sum_k \left\{ W_k + \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \right\}\right)$$

(d) **Model** $[\boldsymbol{\lambda_k \Sigma_0}]$ The prior distribution has three components

$$\begin{aligned} \mu_k|\lambda_k, \Sigma_0 &\sim \mathcal{N}_p(\xi_k, \lambda_k \Sigma_0/\tau_k), \qquad (k = 1, \ldots, k) \\ \lambda_k &\sim \mathcal{I}g(r_k/2, \rho_k/2), \qquad (k = 2, \ldots, K) \\ \Sigma_0 &\sim \mathcal{W}_p^{-1}(m_0, \Psi_0) \end{aligned}$$

We make the model identifiable by setting $\lambda_1 = 1$. Step 3 of Gibbs sampling is then decomposed as follows:

3.1 For $k = 1, \ldots, K$ simulate

$$\mu_k|\Sigma_0, \lambda_k, \nu \sim \mathcal{N}_p\left(\bar{\xi}_k, \frac{\lambda_k}{n_k + \tau_k}\Sigma_0\right)$$

with $\bar{\xi}_k = (n_k\bar{\mathbf{x}}_k + \tau_k\xi_k)/(n_k + \tau_k)$.

3.2 Simulate $(k = 2, \ldots, K)$

$$\lambda_k|\Sigma_0, \nu \sim \mathcal{I}g\left((r_k + n_k p)/2, \frac{1}{2}\{\rho_k + \mathrm{tr}(W_k\Sigma_0^{-1}) + \frac{n_k\tau_k}{n_k + \tau_k}(\bar{\mathbf{x}}_k - \xi_k)'\Sigma_0^{-1}(\bar{\mathbf{x}}_k - \xi_k)\right\}\right)$$

3.3 Simulate

$$\Sigma_0|\lambda_1, \ldots, \lambda_K, \nu \sim \mathcal{W}_p^{-1}\left(m_0 + n, \Psi_0 + \sum_k\left\{W_k/\lambda_k + \frac{n_k\tau_k}{\lambda_k(n_k + \tau_k)}(\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t\right\}\right)$$

(e) **Model** $[\boldsymbol{\Sigma_k}]$ This is the standard Gaussian mixture model considered by Lavine and West (1992) and by Soubiran *et al.* (1991). The prior distribution on (μ_k, Σ_k) is then

$$\mu_k|\Sigma_k \sim \mathcal{N}_p(\xi_k, \Sigma_k/\tau_k) \qquad (k = 1, \ldots, K)$$
$$\Sigma_k \sim \mathcal{W}_p^{-1}(m_k, \Psi_k)$$

and the corresponding Gibbs step is, for $k = 1, \ldots, K$, simulate

$$\mu_k|\Sigma_k \sim \mathcal{N}_p(\bar{\xi}_k, \Sigma_k/(\tau_k + n_k))$$
$$\Sigma_k|\nu \sim \mathcal{W}_p^{-1}\left(n_k + m_k, \Psi_k + W_k + \frac{n_k\tau_k}{n_k + \tau_k}(\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t\right)$$