MCMC Inference for Model-based Clustering

Halima Bensmail
Jacqueline J. Meulman

Data Theory Group
Leiden University

MCMC Inference for Model-based Clustering [1]

Halima Bensmail, Jacqueline J. Meulman
Data Theory Group
Department of Education,
Faculty of Social and Behavioral Sciences, Leiden University,
The Netherlands

september 10 , 1998

[1]This research was supported by The Netherlands Organization for Scientific Research (NWO) by grant no 575-67-053 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis' to the second author.

Abstract

Bensmail, Celeux, Raftery and Robert (1997) introduced a new approach to cluster analysis based on geometric modeling of the within-group covariance in a mixture of multivariate normal distributions, using a fully Bayesian framework. Using a maximum Likelihood approach, Banfield and Raftery (1993) proposed a method of clustering based on the reparameterization of the covariance matrices but restricted some parameters to be known and to have a low value (parameter shape). Dasgupta and Raftery(1995) used the same reparameterization to detect features in a spatial point process in the presence of substantial noise. All approaches work well but have some limitations. The first approach proposed parsimonious models in Bayesian clustering but didn't include the case where there is noise. The two other procedures used a specified model with a low value of the shape parameter.

The present work overcomes those limitations: it generalizes the model used in the three approaches and proposes a Bayesian study in the case there is noise. Canonical discriminant analysis was used to project the data in the canonical space and predict the results obtained by our method of clustering.

The performance of the proposed method is studied by simulation; the procedure is also applied to the analysis of data concerning species of butterflies and diabetes patients. The results are very promising.

Contents

List of Figures

List of Tables

1 Introduction

Cluster analysis has been developed mainly through the invention of empirical, and lately Bayesian study of ad hoc methods, in isolation from more formal statistical procedures. In recent years it has been found that basing cluster analysis on a probability model can be useful both for understanding when existing methods are likely to be successful and for suggesting new methods (Symons, 1981: McLachlan 1982; McLachlan and Basford 1988; Banfield and Raftery 1993). One such probability model is that the population of interest consists of K different subpopulations, and that the density of a p-dimensional observation $\mathbf{x}$ from the kth subpopulation is $f_k(\mathbf{x}, \theta)$ for some unknown vector of parameters θ. Given observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we let $\nu = (\nu_1, \ldots, \nu_n)^t$ denote the identifiying labels, where $\nu_i = k$ if $\mathbf{x}_i$ comes from the kth subpopulation. In the so-called classification maximum likelihood procedure, θ and ν are chosen to maximize the likelihood

$$\prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{\nu_i} f_{\nu_i}(\mathbf{x}_i, \theta_k). \tag{1}$$

In the Bayesian classification, we assume that the data to be classified $\mathbf{x}_i, (i = 1, \ldots, n, \mathbf{x}_i \in \mathbf{R}^p)$ arise from a random vector with density

$$p(\mathbf{x}, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{\nu_i} f_{\nu_i}(\mathbf{x}_i, \theta_k) \tag{2}$$

and that the corresponding classification variables ν_i are unobserved. We are concerned with Bayesian inference about the model parameter θ, π and the classification indicators ν_i.
Bensmail, Celeux, Raftery and Robert (1997) — hereafter BCRR — proposed a Bayesian approach to overcome the difficulties which arose in Banfield and Raftery's (1993) work, hereafter BR. In particular, with the BR algorithm, the user must specify the shape matrix A when fitting the covariance matrix Σ_k by the model

$$D_k A D_k^t,$$

where $A = diag\{1, a_2, \ldots, a_p\}$ with $1 \geq a_2 \geq \ldots a_p > 0$, and D_k is an orthogonal matrix for each $k = 1, \ldots, K$. Dasgupta and Raftery (1995) used the same model to detect features in a spatial point process where the shape matrix was unknown but they constrained the diagonal terms of the matrix shape to be equal: $A = diag\{1, \alpha, \ldots, \alpha\}$ and to have a low value (α is called the shape parameter). In our approach, the shape matrix is completely, unknown and is estimated using a fully Bayesian calculation. In BCRR, four parsimonious

models were explicitely considered to classify the data. Those are the spherical models $[\lambda I]$, $[\lambda_k I]$, the linear model $[\Sigma]$ and the proportional model $[\lambda_k \Sigma]$. We propose to extend this work to the family of models where the covariance matrix is represented by $[\lambda D_k A D_k^t]$ and $[\lambda_k D_k A D_k^t]$ and to the case where the data contain outliers using a fully Bayesian inference via Gibbs sampling. These steps overcome all the limitation mentioned earlier. We use the Laplace-Metropolis approximation to calculate the Bayes factor (Lewis and Raftery 1997; Raftery 1996, BCRR 1997); this approximation is used to choose the model and determine the number of groups simultaneously. In section 2, we describe Bayesian calculation of the models, and outline how the Bayes factor is approximated from the MCMC output. In section 3 we show that the calculation can be extended without great difficulty to the case where there is noise. In section 4 we show the methods at work on real and simulated data sets.

2 Models and Estimation

We assume that the data $\mathbf{x}_i$, $(i = 1, \ldots, n; \mathbf{x}_i \in \mathbf{R}^p)$ to be classified arise from a random vector with density $p(\mathbf{x}, \theta)$ as in (2) where $f_k(., \mu_k, \Sigma_k)$ is the multivariate normal density function, μ_k is the mean and Σ_k is the covariance matrix of the group G_k. $\pi = (\pi_1, \ldots, \pi_K)$ is the mixing proportion $(\pi_k \geq 0, \quad \sum_{k=1}^K \pi_k = 1)$. We are concerned with Bayesian inference about the model parameters θ, π and the classification indicators ν_i. MCMC methods provide an efficient and general recipe for Bayesian analysis of mixtures. For instance, many authors have used the Gibbs sampler or the Data Augmentation method of Tanner and Wong (1987) for estimating parameters in univariate and multivariate Gaussian mixtures and proved that both algorithms converge in distribution to the true posterior distribution of the mixture parameters.

The models we are investigating in this paper are described in Table 1.

Given a classification vector $\nu = (\nu_1, \ldots, \nu_n)$, we use the notation

$$n_k = \sum_i \mathbf{I}\{\nu_i = k\}, \; \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i;\nu_i=k} \mathbf{x}_i, \; W_k = \sum_{i:\nu_i=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t$$

for the component-wise statistics of location x_k and scale W_k, $(k = 1, \ldots, K)$.

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

$$(\pi_1, \ldots, \pi_K) \sim D(\alpha_1, \ldots, \alpha_K)$$

Table 1: Description of the models

Model	Σ_k	Shape	Orientation	Volume
1.	$\lambda D_k A D_k^t$	Same	Different	Same
2.	$\lambda_k D_k A D_k^t$	Same	Different	Different
3.	Σ_k	Different	Different	Different
4.	$D_k A D_k^t$+noise	Same+ Noise	Different	Same
5.	$\lambda_k D_k A D_k^t$+noise	Same+ Noise	Different	Different

and the prior distributions of the means μ_k of the mixture components conditionally on the covariance matrices Σ_k are Gaussian

$$(\mu_k|\Sigma_k) \sim \mathcal{N}_p(\xi_k, \Sigma_k/\tau_k). \tag{3}$$

The conjugate prior distribution of the covariance 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 and following the same steps as in BCRR. So the Gibbs sampler steps go as follows:

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

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

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's. Details are given in the appendix.

(a) Model $[D_k A D_k^t]$

If the prior distribution of the parameters μ_k and A of the model is

$$\mu_k|\Sigma_k \sim \mathcal{N}_p(\xi_k, \Sigma_k/\tau_k) \quad (k = 1, \ldots, K), \quad a_j \sim \mathcal{IG}(\frac{r_j}{2}, \frac{\rho_j}{2}) \quad (j = 1, \ldots, p),$$

then the posterior distribution of $\mu_k|\Sigma_k, A, \nu$ is a MVN with mean $\bar{\xi}_k = \frac{n_k \bar{\mathbf{x}}_k + \tau_k \xi_k}{n_k + \tau_k}$ and covariance matrix $\frac{1}{n_k+\tau_k}\Sigma_k$.

For the other parameters, we assume that A and D_k are the shape and direction components of an inverse Wishart random variable $\mathcal{W}_p^{-1}(m_0, \Psi_0)$ and *that they are independent*

(Anderson 1984, Theorem 13.5.1), so

$$a_t|\nu \sim \mathcal{IG}\left(\frac{n+K(m_0+p)-1}{2}, \frac{\left\{\sum_k D_k^t\left(\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0\right)D_k\right\}_{tt}}{2}\right).$$

The D_k's are then distributed *a posteriori* as the principal direction vectors from the following inverse Wishart distribution,

$$\mathcal{W}_p^{-1}\left(n_k+m_0, \Psi_0+W_k+\frac{n_k\tau_k}{n_k+\tau_k}(\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\right) \qquad (k=1,\ldots,K). \tag{4}$$

(b) Model $[\lambda_k D_k A D_k^t]$.

In this case, the volume is a parameter of interest because the groups are assumed to have differents volumes λ_k. The prior distribution of λ_k is then an inverse gamma distribution

$$\lambda_k \sim \mathcal{IG}(\frac{l_k}{2}, \frac{s_k}{2})$$

and the corresponding Gibbs sampler step is to simulate

$$a_j|\lambda_1,\ldots,\lambda_k,\nu \sim \mathcal{IG}\left(\frac{n+K(m_0+p)-1}{2},\right.$$

$$\left.\frac{1}{2}\left\{\sum_k \lambda_k^{-1}D_k^t((\bar{x}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0)D_k\right\}_{jj}\right), \; j=1,\ldots,p;$$

$$\lambda_k|A,D_1,\ldots,D_K,\nu \sim \mathcal{IG}\left(\frac{l_k+n_kp}{2},\right.$$

$$\left.\frac{1}{2}\left\{s_k+tr(D_kA^{-1}D_k^t)\left((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0\right)\right\}\right).$$

(c) Model $[\Sigma_k]$

This is the standard Gaussian mixture model considered by Lavine and West (1992). In this case, there is no need to use the eigenvalue decomposition of Σ_k. The prior distribution on (μ_k, Σ_k) is then

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

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

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

For choosing both a "model M_2 against a model M_1" and the number of groups in one step, we compute the approximate Bayes factor from the Gibbs sampler output using the *Laplace-Metropolis estimator* of Raftery (1996) and BCRR (1997) given by the formulas:

$$B_{12} = \frac{p(\mathbf{x}|M_2)}{p(\mathbf{x}|M_1)} = \frac{|\Psi_2|^{1/2} p(\mathbf{x}|\tilde{\theta}^{(2)}) p(\tilde{\theta}^{(2)})}{|\Psi_1|^{1/2} p(\mathbf{x}|\tilde{\theta}^{(1)}) p(\tilde{\theta}^{(1)})}, \tag{5}$$

where $\tilde{\theta}^{(i)}, (i = 1, 2)$ is the posterior mode of $\tilde{\theta}^{(i)}$ ($\theta^{(i)}$ is the parameter of the Model M_i), and $\Psi^{(i)}$ is minus the inverse Hessian of $h(\theta) = \log p(\mathbf{x}|\theta)p(\theta)$, evaluated at $\theta = \tilde{\theta}^{(i)}$.
The Laplace method requires the posterior mode, $\tilde{\theta}$, and $|\Psi|$. The Laplace-Metropolis estimator estimates these parameters from the Gibbs sampler output. The Likelihood at the approximate posterior mode is

$$p(\mathbf{x}|\tilde{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tilde{\pi}_k f(\mathbf{x}_i|\tilde{\mu}_k, \tilde{\Sigma}_k)$$

which is then substituted into equation 5 to obtain the Bayes factor.

3 Adding noise to the data

So far, we have assumed that each observation belongs to a cluster. However, there may be some observations that don't follow the rule. For this, we consider the possibility of extending our models to include such observations.
Dasgupta and Raftery(1995) proposed a method based on the EM algorithm for a model-based clustering of p-dimensional data based on a mixture of Gaussian distributions, with an addition (optional) of a component consisting of a homogeneous spatial Poisson process, to represent "noise". They developed an EM algorithm to estimate the shape parameter using maximum likelihood estimates of the parameter obtained by the EM algorithm of the maximized mixture likelihood.
Here, we suppose we have a mixture of a Gaussian distribution, and of noise. The observations representing noise are assumed to arise from a Poisson process with intensity γ. Let n_0 be the number of noise points, so the general finite mixture distribution is

$$p(\mathbf{x}, \theta) = \frac{(\gamma A)^{n_0} e^{-\gamma A}}{n_0!} \prod_{i=1}^{n-n_0} \sum_{k=1}^{K} \pi_{\nu_i} f_{\nu_i}(\mathbf{x}_i, \theta_k) \tag{6}$$

where $f_k(\mathbf{x}, \theta)$ is a $MVN(\mu_k, \Sigma_k)$ density for $k = 1, \ldots, K$ and A is the volume occupied by the data in $\mathbf{R}^p$. Operationally, the definition used here is $A = \prod_{j=1}^{p} (\max_{i=1,\ldots,n}\{\mathbf{x}_{ij}\} -$

$\min_{i=1,\ldots,n}\{x_{ij}\}$), which is the volume of the smallest hyperrectangle with side parallel to the coordinate axes containing the data. Other, perhaps more satisfactory, definitions could also be used, such as the volume enclosed by the convex hull of the data.

4 Examples

We present three examples to illustrate the ability of our model to overcome the limitations described in the Section 1. The first example uses simulated data. We simulate 200 points from a variant bivariate two-component Gaussian mixture and we add 5% and 10% noise which was generated by a Poisson process. The second and the third examples are based on real data.

4.1 Example 1: Simulated Data

We simulated 200 points from a bivariate two-component Gaussian mixture with equal proportions, mean vectors $\mu_1^t = \mu_2^t = (0,0)$ and covariance matrices $\Sigma_1 = diag\{e_{11}, e_{12}\}$ and $\Sigma_2 = diag\{e_{21}, e_{22}\}$ where

$$e_{1j} = \left\{\frac{\alpha(j-1)+p-1}{p}\right\}^2 \quad e_{2j} = \left\{\frac{\alpha(p-j)+p-1}{p}\right\}^2; \quad j = 1,2.$$

When $\alpha = 9$, we get two thin and well separated ellipses (c) and (d) and when $\alpha = 3$ we get two fat ellipses (a) and (b) as we can see in Figure 1. For both examples, 10 (5%) and 20(10%) points are simulated from a Poisson distribution in the hyperrectangle occupied by the data.

The most likely memberships a posteriori of the example (a), (b), (c) and (d) with 500 iterations from Gibbs sampler output with two groups are shown in Figure 2. Convergence was immediate; similar results were obtained for other starting values.

For (c), for example, the model comparison results are shown in Table 2. The correct model $[D_k A D_k^t]$ and the correct number of groups (2) are strongly favored. The posterior means of the parameters for the preferred model are $\mu_1 = (0.05, 0.003)$, $\mu_2 = (0.06, 0.01)$, $A = diag(24.6, 0.24)$ and $d_{11}^1 = d_{22}^1 = 0.001$, $d_{12}^1 = d_{21}^1 = 0.999$, $d_{11}^2 = d_{22}^2 = 0.999$, $d_{12}^2 = d_{21}^2 = 0.001$.

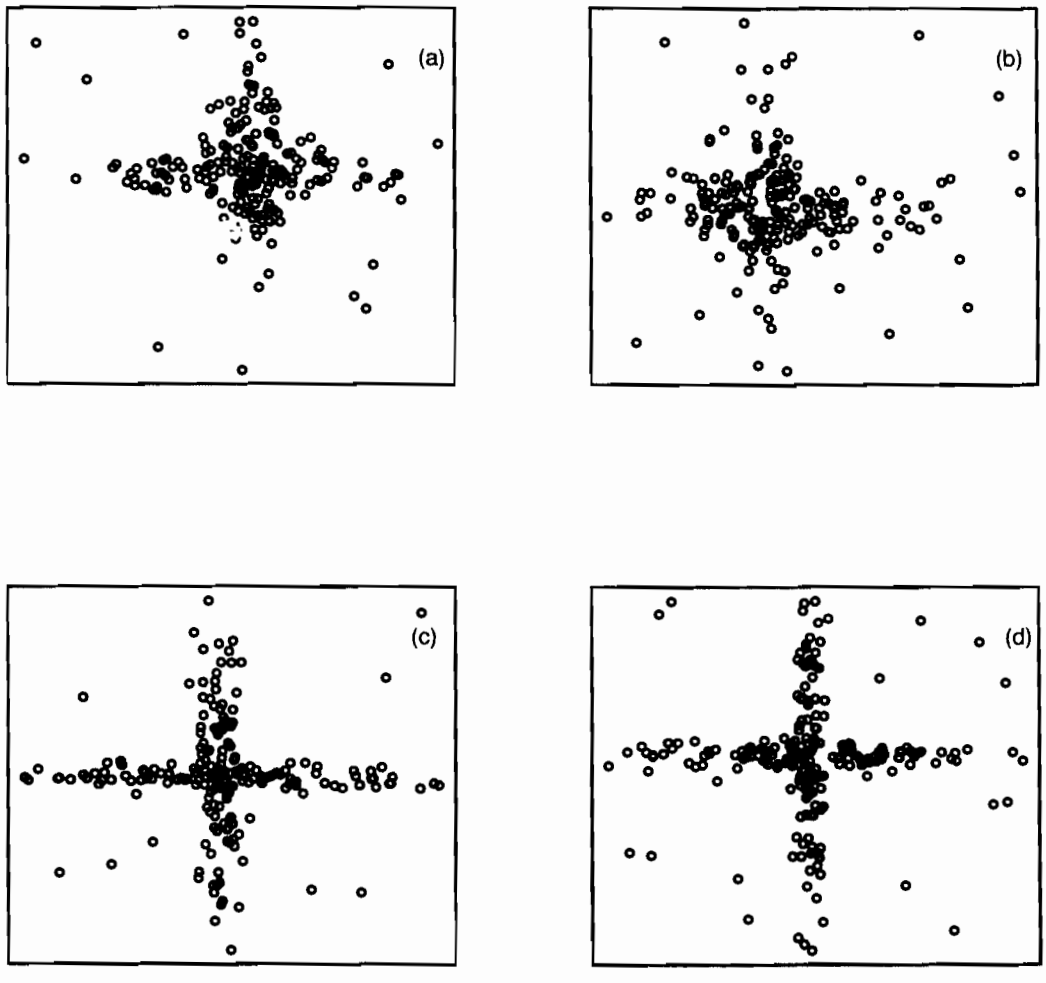


Figure 1: Example 1: Four simulated data sets. In the upper panels, the data are simulated using $\alpha = 3$, with 5% (a) and 10% (b) of noise. In the lower panels, the data are simulated using $\alpha = 9$, with 5% (c) and 10% (d) of noise.

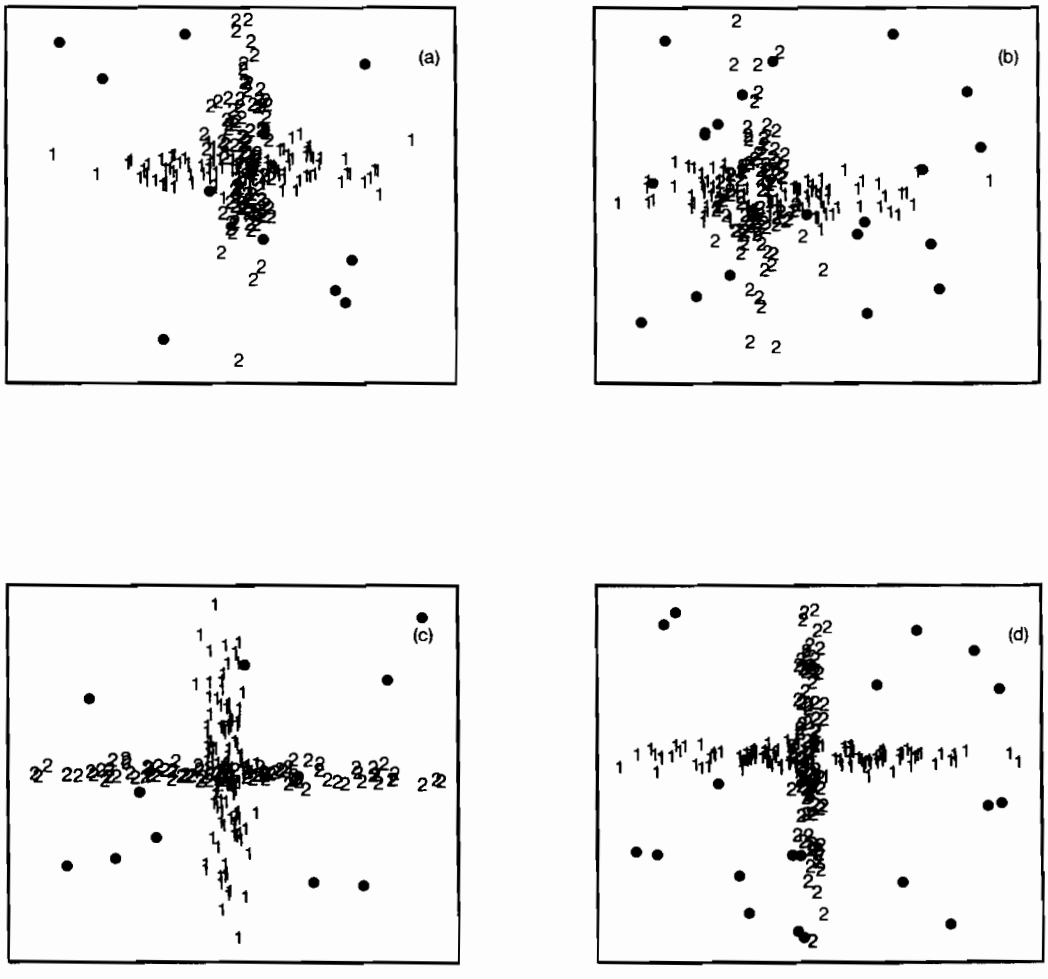


Figure 2: Example 1: Results of the classification of the four simulated data sets from Figure 1. The black dots represent noise.

Table 2: Example 1: Approximate Log Integrated Likelihoods for simulated data (c).

No. of groups	$[D_k A D_k^t]$	$[\lambda_k D_k A D_k^t]$	$[\Sigma_k]$
1	-1336	-2262	-3094
2	**-1074**	-1294	-1081
3	-1286	-9086	-5091
4	-1194	-4172	-7794

4.2 Example 2: Butterfly classification

Figure 3 shows two variables, z_3 and z_4, of a set with four measurements on 23 butterflies wings. The example is taken from Celeux and Robert (1993), it was also analyzed by BCRR (1997). The aim is to decide how many species are represented in this group of insects, and to classify them.

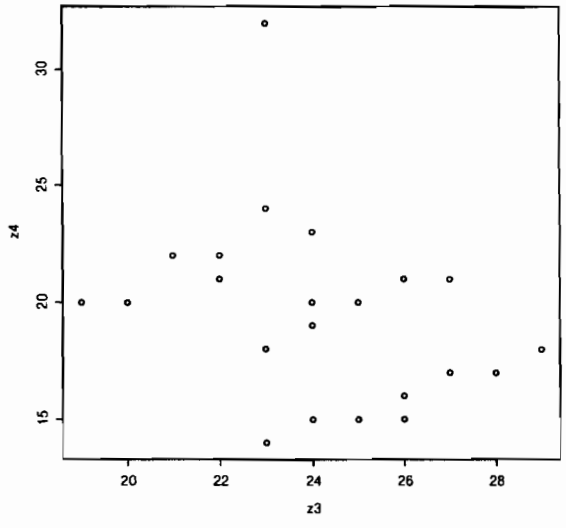


Figure 3: Example: Butterfly Data

The 22 butterflies were classified into 3 groups by the model $[D_k A D_k^t]$, as Table 3 shows, and butterfly 23 was considered as noise as we can see in Figure 4. BCRR (1997) obtained 4 groups where the fourth group contains only the observation 23 considered as an aberrant butterfly. Our approach confirms this result, moreover, it gives more information by estimating the features shape, volume and orientation of each cluster.

Table 3: Example 2: Approximate Log Integrated Likelihoods.

No. of groups	$[D_k A D_k^t]$	$[\lambda_k D_k A D_k^t]$	$[\Sigma_k]$
1	-164	-160	-161
2	-163	-176	-175
3	**-134**	-138	-139

Figure 4: Example 2: Classification of the Butterfly data by the model $[D_k A D_k^t]$.

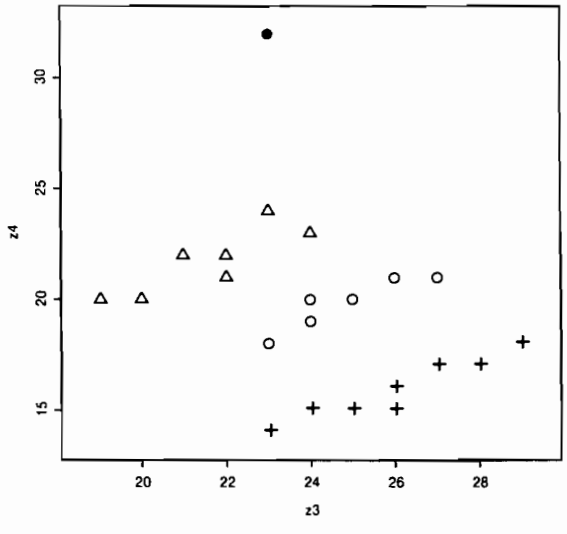
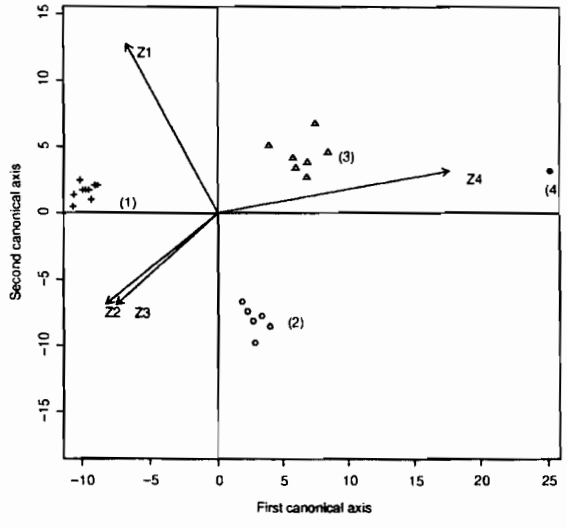


Figure 5: Example 2: Classification and projection of the Butterfly data in the canonical space using canonical discriminant analysis. The lines drawn represent the variables in the canonical space. The canonical variables are represented by the horizontal and the vertical lines drawn through the origin.



In Figure 5; the butterflies are displayed in the canonical space, represented with their group symbols. The clouds of points for the different groups are clearly separated. To investigate the set of variables, they are depicted as vectors in the canonical space. The orthogonal projection of the butterfly points onto the variable vectors gives an approximation of the data; the length of the vectors is proportional to their goodness-of-fit, which is the correlation of the approximation and the original variable. The variables seem to consist of three subsets; subset 1 contains z_2, z_3 ordering the groups from 1 to 4. The second subset, consisting of z_4, reverse the order from 4 to 1. The third subset finally reverse the order of the groups 2 and 3 and put the aberrant butterfly into the group 2.

4.3 Example 3: Diabetes Data

Reaven and Miller (1979) described and analyzed data for 145 subjects, consisting of the area under a plasma glucose curve (glucose area), the area under a plasma insulin curve (insulin area) and steady-state plasma glucose reponse (SSPG). The subjects were clinically classified into three groups, chemical diabetes (Type 1), overt diabetes (Type 2), and normal (nondiabetic). Symons (1981) reanalyzed the data using seven different clustering criteria.

Table 4: Example 3: Approximate Log Integrated Likelihoods for Diabetes data.

No. of groups	$[D_k AD_k^t]$	$[\lambda_k D_k AD_k^t]$	$[\Sigma_k]$
1	-556	-427	-500
2	-466	-679	-440
3	-380	**-350**	-445
4	-997	-456	-460

The data have the three-dimensional shape of a boomerang with two wings and a fat middle as we can see in Figure 6 (c). One of the wings corresponds to patients with overt diabetes, the other wing is composed primarily of patients with chemical diabetes, and the "fat middle" is composed of normal patients.

Banfield and Raftery(1993) evaluated the criterion based on the model $[\lambda_k D_k AD_k^t]$, where $A = diag\{1, \alpha, \alpha\}$, which means that the clusters have different sizes and orientations but have the same "tubular" shape. Based on the previous information, we used our approach developed in the previous section to classify the data. We write $\Lambda_k = \lambda_k A$. Table 4 shows that the model $[\lambda_k D_k AD_k^t]$ with three groups is favored quite strongly over the alternatives. The estimated values of $\Lambda_1 = diag\{0.039, 0.038, 0.035\}$, $\Lambda_2 = diag\{0.47, 0.46, 0.42\}$ and $\Lambda_3 = diag\{0.17, 0.15, 0.014\}$. The result reported for the BR algorithm is $\alpha = 0.2$ which limitates the variant characteristics of the shape for the three clusters.

The optimal classification, as we can see in Figure 6(d), 7(a) resulted in only 9% of the points being misclassified, given in Table 5, which are mostly situated at the border of the clusters. Ten percent error rate was obtained by the BR algorithm, but the misclassified points were dispersed over the three clusters. The present results compare our approach favorably to the procedures of Symons (1981) and BR (1993).

		Predicted classification			
		Normal	Chemical	Overt	Total
Clinical classification	Normal	69	7	0	76
	Chemical	1	35	0	36
	Overt	0	5	28	33
	% correct	0.91	0.97	0.85	0.91

Table 5: Example 3: Classification table giving the results of clustering the diabetes patients.

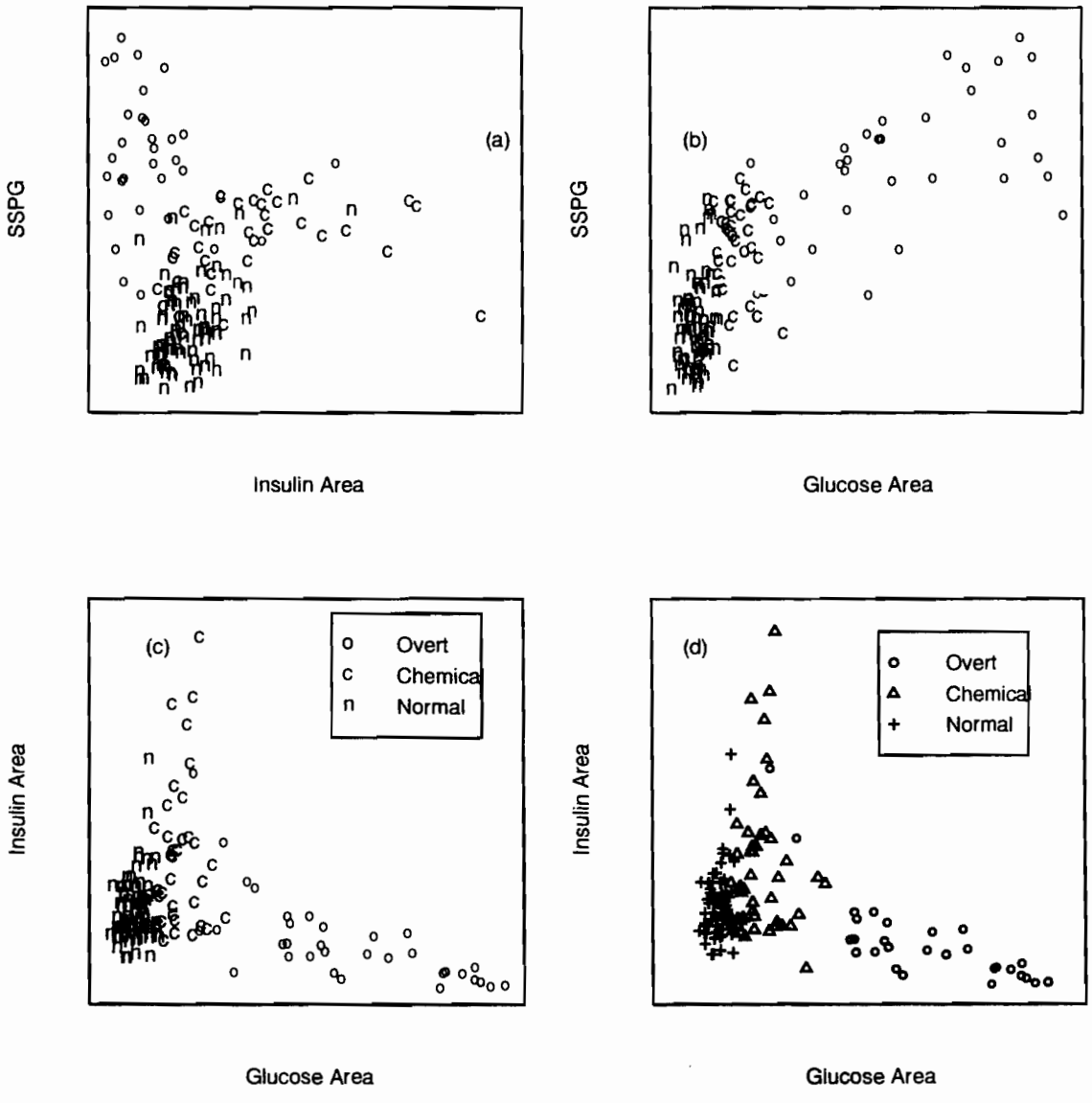


Figure 6: Example 3: Three two-dimensional projections of the three-dimensional diabetes data. The symbols indicate the clinical classification of subjects as having chemical diabetes, overt diabetes, or being normal. Lower right panel shows the three clusters in the diabetes data found by our approach using the model $[\lambda_k D_k A D_k^t]$.

Figure 7(b) shows the plot of component loadings for the diabetes data. The component loadings (Meulman 1986, Meulman et al 1992) are equivalent to the Pearson correlations between the variables and the canonical scores obtained by discriminant analysis. The length of the line drawn from the origin to each variable point approximates the importance of that variable. The canonical variables are not plotted but can be represented by horizontal and vertical lines drawn through the origin.

In Figure 7(a); the diabetes subjects are displayed in the canonical space, represented with their groups symbols. The clouds of points for the different groups are mildly separated. To investigate the structure of the set of variables, they are depicted as vectors in the space (the Figures 7(a) and 7(b) are represented separately for the sake of clarity, but they show the very same space).

The variable insulin orders the groups from chemical diabetes (Type 1), normal (nondiabetic) to overt diabetes (Type 2). The variable glucose orders the groups from overt diabetes, chemical diabetes to normal (nondiabetic) and the variable SSPG distinguishes predominantly chemical and normal on one side, and overt diabetes on the other.

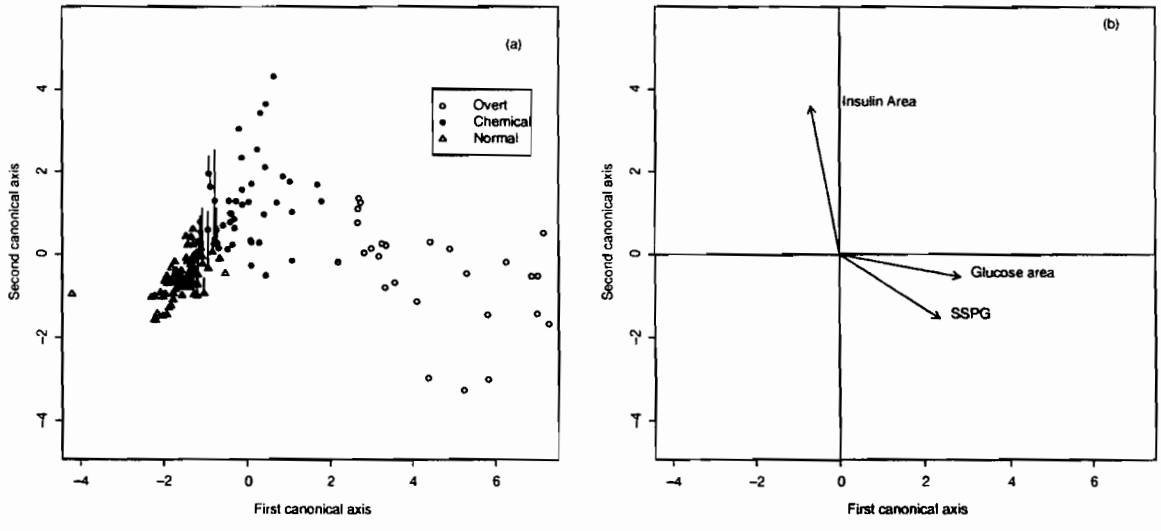


Figure 7: Example 3: (a) shows the estimated group memberships and uncertainty plot for diabetes data. (b) shows the representation of the variables in the canonical space.

5 Discussion

We have presented a fully Bayesian analysis using model-based clustering, with a mixture model in which the features of interest are presented by multivariate normal densities, specified by the shape, the features and orientations across clusters.

Alternative, frequentist, approaches, consist of maximizing the mixture likelihood using the EM algorithm. Dasgupta and Raftery (1995) considered the model $[D_k A D_k]$ in which the features of interest are presented by multivariate normal densities with high linearity, specified by the *shape* being the same across features, which make the model restricted. Here, we considered the case where shape, orientation and volume are totally unknown and no restriction is made on the shape parameter. We also included the case where there is noise. Moreover, we have proposed a way of generalizing and overcoming the limitations of the previous related classification procedures for cluster analysis. These are the inability to specify some but not all features to be constant across clusters; the limitation of BCRR by considering a Bayesian clustering with only parsimonious models and the failure to do a Bayesian approach to account for "noise". We have also used an approximate Bayesian solution to the problem of choosing the number of clusters and the optimal model.

In context of Gaussian clustering, we introduced the canonical discriminant analysis as a tool in Bayesian clustering for multidimensional data. The calculation of the canonical scores and the projection of the data in the canonical space, is useful to view the clusters and verify the importance of the variables in the canonical space.

Appendix: Gibbs Sampling for the Clustering Models

(a) Model $[\lambda D_k A D_k^t]$.

Here, the distinction between λ and A is entirely geometric (volume versus shape), and we will therefore consider a single parameter A, where the first term of the diagonal a_1 is no longer constrained to be equal to 1.

If the a's are numbered in descending order of magnitude and if $d_{j1} \geq 0$, then (with probability 1) the transformation from Σ_k to A and D_k is unique. Let the matrix D_k be given the coordinates $d_1, \ldots, d_{p(p-1)/2}$, and let the Jacobian of the transformation be $J(A, D_k)$. Then the joint density of A and D_k is $g(a_1, \ldots, a_p)J(A, D_k)$, where g is the density of Σ_k. We can calculate from the marginal distribution of an inverse wishart $W_p^{-1}(m_k, I_p)$, the marginal distribution of A, but the marginal distribution of D_k is not easy to calculate since

$$\begin{aligned} p(A) &= \int \ldots \int g(a_1, \ldots, a_p) J(A, D_k) dD_{k,1} \ldots dD_{k,p(p-1)/2} \\ &\propto \prod_{k=1}^{K} e^{-\frac{1}{2}\sum_{i=1}^{p} \frac{1}{a_i}} \prod_{i=1}^{p} a_i^{-(m_k+p+2)/2} \prod_{i<j} (a_i - a_j) \end{aligned}$$

$$p(D_k) \propto \int \ldots \int e^{-\frac{1}{2}\sum_{i=1}^{p} \frac{1}{a_i}} \prod_{i=1}^{p} a_i^{-(m_k+p+2)/2} \prod_{i<j} (a_i - a_j) da_1, \ldots, da_p.$$

An elegant approach is to use as a prior distribution on D_k, a normalized measure called *Haar* distribution (Muirhead 1982, Anderson 1956, page 321), which is a uniform distribution with the constant $\frac{\Gamma_p[\frac{p}{2}]}{2^p \pi^{\frac{p^2}{2}}}$. If we consider a conjugate distribution on $a_1, \ldots, a_p$, then the prior distribution of the parameters μ_k and A of the model is

$$\mu_k | \Sigma_k \sim \mathcal{N}_p(\xi_k, \Sigma_k/\tau_k) \qquad (k = 1, \ldots, K), \qquad a_t \sim \mathcal{I}g(r_t/2, \rho_t/2) \qquad (t = 1, \ldots, p). \quad (7)$$

The Gibbs sampler step is, for $k = 1, \ldots, K$, simulate

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

We assume that D_k and A are the direction and shape components of an inverse Wishart random variable and *that they are independent.* This is true only asymptotically (i.e. as the number of degrees of freedom goes to infinity) (Anderson 1984, Theorem 13.5.1) but it considerably simplifies the simulation, with moderate effects (if any) on the resulting

posterior distribution. If we thus assume that the couples (D_k, A) are all distributed as $\mathcal{W}_p^{-1}(m_0, \Psi_0)$, we derive from the posterior distribution

$$\prod_k |A|^{-(n_k-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(A^{-1}\left[\sum_k D_k^t\left((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k\right)D_k\right]\right)\right\}\times$$

$$|A|^{-(m_0+p+1)/2}\exp\{-\mathrm{tr}(D_k^tA_0^{-1}D_k\Psi_0)/2\}$$

so

$$A|D_1,\ldots,D_K,\nu\sim\pi(A)\propto|A|^{-(n-K+K(m_0+p+1))/2}\times$$

$$\exp\left\{-\frac{1}{2}\mathrm{tr}\left(A^{-1}\left[\sum_k D_k^t\left((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0\right)D_k\right]\right)\right\}.$$

This is a condensed way of saying that the diagonal elements of A are distributed according to the inverted gamma distributions,

$$a_t|\nu\sim\mathcal{I}g\left(\frac{n+K(m_0+p)-1}{2},\frac{\left\{\sum_k D_k^t\left((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0\right)D_k\right\}_{tt}}{2}\right).$$

The D_k's are then distributed *a posteriori* as the principal direction vectors from the following inverse Wishart distribution,

$$\mathcal{W}_p^{-1}\left(n_k+m_0,\Psi_0+W_k+\frac{n_k\tau_k}{n_k+\tau_k}(\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\right)\qquad(k=1,\ldots,K).\tag{8}$$

(b) Model $[\lambda_k D_k A D_k^t]$.

The Gibbs simulations of the μ_k's and D_k's are similar to those of Model $[\lambda D_k A D_k^t]$, with $\lambda_k A$ replacing A in the previews equations.

The simulation of A is also quite close to the previous version since

$$A|D_1,\lambda_1,\ldots,D_K,\lambda_K,\nu\sim\pi(A)\propto|A|^{-(n-K+K(m_0+p+1))/2}\times$$

$$\exp\left\{-\mathrm{tr}\left(A^{-1}\left[\sum_k\lambda_k^{-1}D_k^t\left(\frac{n_k\tau_k}{n_k+\tau_k}(\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t+W_k+\Psi_0\right)D_k^t\right]/2\right)\right\}.$$

If we assume $\lambda_k\sim\mathcal{I}g(l_k/2,\rho_k/2)$, the posterior distribution of λ_k in the Gibbs sampler is for $(k=1,\ldots,K)$

$$\lambda_k|A,D_1,\ldots,D_K,\nu\sim\mathcal{I}g\left(\frac{l_k+n_kp}{2},\frac{\rho_k+\mathrm{tr}\{D_kA^{-1}D_k^t((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+W_k+\Psi_0)\}}{2}\right).$$

and for $(j=1,\ldots,p)$

$$\mathbf{a}_j|\lambda_1,\ldots,\lambda_k,\nu\sim\mathcal{I}G\left(\frac{n+K(m_0+p)-1}{2},\right.$$

$$\frac{1}{2}\left\{\sum_k \lambda_k{}^{-1} D_k^t((\bar{x}_k - \xi_k)(\bar{x}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + W_k + \Psi_0) D_k \right\}_{jj}\Bigg) .$$

References

Anderson, T. W. (1956,1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Banfield, J. D., Raftery, A. E. (1993). "Model-based Gaussian and non Gaussian Clustering" *Biometrics* **49**, 803-821.

Bensmail, H., Celeux, G., Raftery, A. & Robert, C.(1997). "Inference in Model-Based Cluster Analysis." *Computing and Statistics*, **7**, 1-10.

Celeux, G. & Govaert, G. (1995). "Gaussian parsimonious clustering models." *Pattern Recognition*, **28**, 781-793.

Dasgupta, A. & Raftery, A. E. (1995). " Detecting Features in Spatial Point Processes with clutter via Model-Based Clustering", *Technical Report No. 295,* Department of Statistics, University of Washington.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Wiley, New York.

Lavine, M. & West, M. (1992). " A Bayesian method for classification and discrimination." *The Canadian Journal of Statistics* **20**, 451-461

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

McLachlan, G. (1982) "The classification and mixture maximum likelihood approaches to cluster analysis." In *Handbook of Statistics*, (Vol 2), P. R . Krishnaiah and L. N. Kanal (eds), 199-208. Amsterdam: North-Holland.

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

Meulman, J. J. (1986). *A Distance Approach to Nonlinear Multivariate Analysis.* DSWO Press, Holland.

Meulman, J. J., Zeppa, P., Boon, M. E., & Rietveld, W. J. (1992). "Prediction of various grades of cervical preneoplasia and neoplasia on plastic embedded cytobrush samples: discriminant analysis with qualitative and quantitative predictors." *Analytical and Quantitative Cytology and Histology*, **14**, 60-72.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.* Wiley, New York.

Raftery, A. E. (1996). "Approximate Bayes factors and accounting for model uncertainty in generalized linear models." *Biometrika* **83**, 251-266.

Reaven, G. M. & Miller, R. G. (1979). " An attempt to define the nature of chemical diabetes using a multidimensional analysis." *Diabetologia* **16**, 17-24.

Symons, M. (1981). "Clustering Criteria and multivariate normal mixtures." *Biometrics* **37**, 35-43.

Tanner, M. & Wong, W. (1987). "The calculation of posterior distributions by data augmentation(with discussion)". *Journal of the American Statistical Association* **82**, 528-550.