

Model-based Clustering with Noise: Bayesian inference and estimation

H. Bensmail⁽¹⁾

University of Tennessee

J.J. Meulman⁽²⁾

Leiden University, The Netherlands

⁽¹⁾Department of Statistics, University of Tennessee, 334 Stokely Management Building, Knoxville, TN 37996-0532, USA. (e-mail: bensmail@utk.edu). This work was done while the first author was working at the Department of Education, Data Theory Group, Leiden University and it was supported by "The Netherlands Organization for Scientific Research" (NWO) by grant nr. 030-56403 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis' to the second author. During the review process, we have been acquainted with new independent work that touches upon a particular part of our work. First author's address: Halima Bensmail, Department of Statistics, University of Tennessee, 334 Stokely Management Building, Knoxville, TN 37996-0532, USA; e-mail: bensmail@utk.edu

Abstract:

Bensmail, Celeux, Raftery and Robert (1997) introduced a new approach to cluster analysis based on geometric modeling based on the within-group covariance in a mixture of multivariate normal distributions using a fully Bayesian framework. This is a model-based methodology, where the covariance matrix structure is involved. Previously, similar structures were used (using a maximum likelihood approach) by Banfield and Raftery (1993) for clustering data where they restricted some parameters of the covariance matrix structure to be known. In the same framework, Dasgupta and Raftery (1998) used the same reparameterization to detect the features in a spatial point process using maximum likelihood approach. These approaches work well, but they have some limitations. These limitations include the fact that not all covariance structures were considered and some parameters of the covariance structures were fixed.

This paper proposes a new way of overcoming the existing limitations. It generalizes the model used in the the previous approaches by introducing a more comprehensive portfolio of covariance matrix structures. Further, this paper proposes a Bayesian solution in the presence of the noise in clustering problems. The performance of the proposed method is first studied by simulation; the procedure is also applied to the analysis of data concerning species of butterflies and diabetes patients.

Keywords: Bayes factor; Eigenvalue decomposition; Gaussian mixture; Gibbs sampler; Canonical discriminant analysis; Markov chain Monte Carlo.

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 the last 25 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 (Binder 1978; Menzefricke 1981; Symons 1981; McLachlan 1982; McLachlan and Basford 1988; Banfield and Raftery 1993; Bock 1996; Fraley and Raftery 1998; and Fraley (1999)). One such probability model is that the population of interest consists of K different subpopulations G_1, \dots, G_K and that the density of a p -dimensional observation \mathbf{x} from the k th subpopulation is $f_k(\mathbf{x}, \boldsymbol{\theta}_k)$ for some unknown vector of parameters $\boldsymbol{\theta}_k$ ($k = 1, \dots, K$). Given observations $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, we let $\mathbf{v} = (v_1, \dots, v_n)^t$ denote the unknown identifying labels, where $v_i = k$ if \mathbf{x}_i comes from the k th subpopulation. In the so-called classification maximum likelihood procedure, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ and $\mathbf{v} = (v_1, \dots, v_n)^t$ are chosen to maximize the classification likelihood

$$p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K; v_1, \dots, v_n | \mathbf{x}) = \prod_{i=1}^n f_{v_i}(\mathbf{x}_i | \boldsymbol{\theta}_{v_i}). \quad (1)$$

In the theory of finite mixtures, the data to be classified are viewed as coming from a mixture of probability distributions, each representing a different cluster, so the likelihood is expressed as

$$p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K; \pi_1, \dots, \pi_K | \mathbf{x}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) \quad (2)$$

where π_k is the probability that an observation belongs to the k th component

$$(\pi_k \geq 0; \sum_{k=1}^K \pi_k = 1)$$

Methods based on this theory performed well in many cases and applications included character recognition (Murtagh and Raftery 1984), tissue segmentation (Banfield and Raftery 1993) , minefield and seismic fault detection (Dasgupta and Raftery 1998) , application to astronomical data (Bensmail et al. 1997; Roeder and Wasserman 1997; Mukherjee et al. 1998) and enzymatic activity in the blood (Richardson and Green 1997).

An advantage of Bayesian model-based clustering (Bensmail et al. 1997) is that there is an associated criterion (the Bayes factor) for assessing the model and the number of clusters and for providing a measure of uncertainty in the associated classification.

We describe here a clustering methodology, which is an extension of model-based clustering methods, based on multivariate normal mixtures using fully Bayesian calculation (and not an approximation of the posterior distribution of the parameters involved) for two specific models of interest (see Section 3). The Bayes factor will be used for comparison of the models and the number of components in the mixture.

This paper is organized as follows: In section 2, we give the necessary background for model-based cluster analysis, and describe different problems raised in the model-based cluster analysis approach. In section 2, we describe Bayesian calculation of the models we propose, and outline how the Bayes factor is approximated from the MCMC (Markov Chain Monte Carlo) output. In section 3 we show that the calculation can be extended to the case where there is noise. In section 4 we show the methods at work on real and simulated data sets.

2. Model-Based Cluster Analysis

In cluster analysis, we consider the problem of determining the structure of the data with respect to clusters when no information other than the observed values is available; from the extensive literature, we mention Hartigan (1975) , Gordon (1999), and Kaufman and

Rousseeuw (1990). Important references on the statistical aspects of cluster analysis include MacQueen (1967) , Wolfe(1978), Scott and Symons (1971), and Bock(1985). Various strategies for simultaneous determining of the number of clusters and the cluster membership have been proposed (e.g. Engelman and Hartigan 1969; Bozdogan 1993), for a review see Bock (1996). An alternative is described in this paper based on the reparameterization of the covariance matrices using a fully Bayesian framework.

Mixture models provide a useful statistical frame of reference for cluster analysis. The Bayesian approach is promising for a variety of mixture models, both Gaussian and non Gaussian (Binder, 1981; Banfield and Raftery, 1993; McLachlan and Peel, 2000, Ch. 4). Banfield and Raftery (1993) –hereafter BR– introduced a new approach to cluster analysis based on a mixture of multivariate normal distributions, where the covariance matrices Σ_k in the classes are modelled in a geometrically interpretable way. Their approach is based on a variant of the standard spectral decomposition of Σ_k , namely

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t \quad (3)$$

where λ_k is a scalar, $\mathbf{A}_k = \text{diag}(1, a_{k2}, \dots, a_{kp})$ where $1 \geq a_{k2} \geq \dots a_{kp} > 0$, and \mathbf{D}_k is an orthogonal matrix for each $k = 1, \dots, K$.

The approach proposed by BR has a number of limitations:

- (a) It only gives classifications of each individual, and produces no assessment of the associated uncertainty.
- (b) It tends to yield partitions that are suboptimal (even if often good).
- (c) Estimates of the model parameters θ based on the estimated partition tend to be biased.
- (d) It assumes the mixing proportions π_k to be equal.
- (e) The algorithms based on some of the models (see model 5 and model 6 in Table 1) require the shape matrix \mathbf{A} to be specified in advance by the user.

(f) To choose K , the number of groups, BR proposed an approximation to the posterior probabilities based on a quantity called the Approximate Weight of Evidence (AWE). While this has worked fairly well in practice, it is quite crude. Another criterion for assessing the number of clusters in a mixture model was used particularly ICOMP which uses the information complexity (Bozdogan 1993), Normalized Entropy of assessment which uses the components, classification maximum likelihood and entropy, of the maximum likelihood (Celeux and Soromenho 1996), and the Bayesian Information criterion or "BIC" introduced by Schwarz (1978). For the last criterion, although regularity conditions do not hold for mixture models, there is considerable theoretical and practical support for its use (Leroux 1992; Roeder and Wasserman 1997; Mukerjee et al. 1998; Dasgupta and Raftery 1998; and Fraley and Raftery 1998).

(g) BR proposed no formal way of choosing among the possible models, this must be done by the user.

To overcome the difficulties which arose in the BR approach, Bensmail, Celeux, Raftery and Robert (1997) proposed a Bayesian approach which overcomes the limitations mentioned above ((a),..., (g)). However, only four models for Σ_k were explicitly considered. These are the spherical models $[\lambda$

$]$ and $[\lambda_k \mathbf{I}]$ (in what follows, $[.]$ is used to indicate a particular model for Σ_k), the linear model $[\Sigma]$ and the proportional model $[\lambda_k \Sigma]$. Dasgupta and Raftery (1998) used the model $[\mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ to detect features in a spatial point process where the shape matrix \mathbf{A} was unknown but they constrained the diagonal terms of the shape matrix to be equal: $\mathbf{A} = \text{diag}\{1, \alpha, \dots, \alpha\}$ and to have a low value.

Our particular interest is to extend this previous work in two respects, using the fully Bayesian inference we develop here. First, to the family of models where the covariance matrix Σ_k is represented by $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$, $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$ and $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t$ respectively, where all the parameters

$\lambda, \lambda_k, \mathbf{A}, \mathbf{A}_k$ and \mathbf{D}_k ($k = 1, \dots, K$) are unknown. Second, to the case where the data contain outliers. The parameters involved in the parameterization of the covariance matrix Σ_k are unknown and not constrained as in Banfield and Raftery (1993), and Dasgupta and Raftery (1998). The models we are discussing here can be applied to minefields and seismic faults, earthquake and other particular problems as discussed by Dasgupta and Raftery (1998). Table 1 shows the geometric interpretation of the various parametrizations used. We estimate π, ν and the parameters of the models given in Table 1 using Gibbs sampling.

We use the Laplace-Metropolis approximation to calculate the Bayes factor (Bensmail et al. 1997 and Dasgupta and Raftery 1998); the latter is used to choose the model and determine the number of groups simultaneously.

3. Models and Estimation

3.1 Bayesian estimation of the models using the Gibbs sampler

We assume that the data are generated by a mixture of underlying probability distributions; each component of the mixture represents a different cluster so that the observations \mathbf{x}_i ($i = 1, \dots, n; \mathbf{x}_i \in R^p$) to be classified arise from a random vector X with likelihood density $p(\theta, \pi | X = \mathbf{x})$ as in (2), where $f_k(\cdot | (\theta_k = \mu_k, \Sigma_k))$ is the multivariate normal density function, μ_k is the mean and Σ_k is the covariance matrix for the k^{th} group. $\pi = (\pi_1, \dots, \pi_K)$ is the mixing proportion ($\pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$). We are concerned with Bayesian inference about the model parameters θ, π and the classification indicators ν . Markov Chain Monte Carlo (MCMC) methods (e.g. Gilks, Richardson and Spiegelhalter, 1996) provide an efficient and general recipe for Bayesian analysis of mixtures. In fact, as explained in Gelman, Carlin, Stern and Rubin (1995) the key to Markov chain simulation is to create a Markov process whose stationary distribution is a specified $p(\theta | \mathbf{x})$ and run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution. When, as in our case, the posterior conditional distribution of the parameters is a complicated function of the

parameters which in most cases are of high dimension, the MCMC algorithm is used to simulate a sample from the posterior distribution of each parameter and after convergence, the posterior mode of each sample is used as the Bayes estimate of the parameter considered. For instance, many authors have used the Gibbs sampler or the Data Augmentation method of Tanner and Wong (1987) (Wei and Tanner 1990 and Green 1995) for estimating parameters in univariate and multivariate Gaussian mixtures. One important consideration regarding the implementation of both algorithms is monitoring convergence. Tierney (1994) proved that both algorithms converge in probability to the true posterior distribution of the mixture parameters. The models we are investigating in this paper are described in Table 1.

Insert Table 1 about here

Given a classification vector $\mathbf{v} = (v_1, \dots, v_n)$, we use the notation $n_k = \#\{i : v_i = k\}$ for the number of observations in cluster k , $\bar{\mathbf{x}}_k = \sum_{i:v_i=k} \mathbf{x}_i/n_k$ for the sample mean vector of all observations in cluster k , and $\mathbf{W}_k = \sum_{i:v_i=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'$ for the sample covariance matrix. We use conjugate priors for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ of the mixture model. The prior distribution of the mixing proportions is a Dirichlet distribution

$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K),$$

$$\left(\text{with joint distribution } p(\boldsymbol{\pi}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_K)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \pi_1^{\alpha_1 - 1} \dots \pi_K^{\alpha_K - 1} \right)$$

The prior distributions of the means $\boldsymbol{\mu}_k$ of the mixture components conditionally on the covariance matrices $\boldsymbol{\Sigma}_k$ are Gaussian

$$(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k) \sim N_p(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k / \tau_k). \tag{4}$$

with known scale factors $\tau_1, \dots, \tau_K > 0$ and locations $\xi_1, \dots, \xi_K \in R^p$, and in addition

$\boldsymbol{\pi}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ are independent

$\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K | \boldsymbol{\pi}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ are independent under different models

The conjugate prior distribution of the covariance matrices $\boldsymbol{\Sigma}_k$ depends on the model, and will be given for each model in turn.

We estimate the parameters of the models in Table 1 by determining the configurations $\boldsymbol{\pi}, \boldsymbol{\theta}$ that maximize the posterior density of $\boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{x}$ (posterior mode values). This posterior density is calculated (approximated) by the Gibbs sampler by simulating from the joint posterior

distribution of $\boldsymbol{\pi}, \boldsymbol{\theta}$ and \mathbf{v} . At iteration $(t + 1)$, the Gibbs sampler steps go as follows:

1. Simulate the classification variables $v_i^{(t+1)}, i = 1, \dots, n$, independently according to the posterior probabilities $p_{ik}(\boldsymbol{\pi}, \boldsymbol{\theta}) = P(v_i^{(t)} = k | \boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{x}_i)$ ($\mathbf{k} = \mathbf{1}, \dots, \mathbf{K}$) conditional on the current values for $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ such that

$$p_{ik}^{(t+1)} = \pi_k f_k(\mathbf{x}_i | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) / \sum_{k=1}^K \pi_k^{(t)} f_k(\mathbf{x}_i | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) \quad i = 1, \dots, n, \quad k = 1, \dots, K.$$

There might be classes k which are empty. To solve this problem, we assign the observation which is closest to $\boldsymbol{\mu}_k^{(t)}$ to this empty class.

2. Simulate the vector $\boldsymbol{\pi}^{(t+1)} = (\pi_1^{(t+1)}, \dots, \pi_K^{(t+1)})$ of mixing proportions from its posterior distribution given $\mathbf{v}^{(t+1)}$, in particular from

$$\boldsymbol{\pi}^{(t+1)} \sim \text{Dirichlet}(\alpha_1 + n_1^{(t+1)}, \dots, \alpha_K + n_K^{(t+1)})$$

with α_k the known parameters of the prior Dirichlet distribution.

3. Simulate the parameter $\boldsymbol{\theta}^{(t+1)}$ of the model from the posterior distribution $\boldsymbol{\theta} | \mathbf{v}^{(t+1)}, \boldsymbol{\pi}$.
4. Iterate the steps 1 to 3.
(Details on the simulation of the parameters $\lambda, \lambda_k, \mathbf{A}, \mathbf{A}_k$ and \mathbf{D}_k are discussed in the paragraphs of 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 $\boldsymbol{\theta}$, was demonstrated 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 nature of the chain associated with the v_i 's. This chain is ergodic with state space $\{1, \dots, K\}$, and is thus geometrically convergent. These properties transfer automatically to the sequence of simulated values of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, and important properties as the central limit theorem or the law of the iterated logarithm are then

satisfied (Diebolt and Robert 1994).

For the models 1, 2, 3 and 4 of Table 1 the calculations are given in Bensmail et al. (1997), so we proceed here with the models 5-7. In what follows, we will describe the simulation steps in step 3 of the algorithm for the parameters to be estimated which are μ_k , λ , \mathbf{A} and \mathbf{D}_k

($k = 1, \dots, K$) for the model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$, μ_k , λ_k , \mathbf{A} and \mathbf{D}_k ($k = 1, \dots, K$) for the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$, and μ_k , Σ_k ($k = 1, \dots, K$) for the general model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$.

(a) Model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$

If the prior distribution of the parameter μ_k is as given in (4) and if the prior distribution of λ is assumed to be an inverse gamma distribution

$$\lambda \sim IG\left(\frac{m_0}{2}, \frac{s_0}{2}\right) \text{ with density } p(\lambda) = \frac{(s_0/2)^{\frac{m_0}{2}}}{\Gamma(\frac{m_0}{2})} \lambda^{-(\frac{m_0}{2}+1)} e^{-\frac{s_0}{2\lambda}} \quad (5)$$

with m_0 and s_0 hyperparameters chosen by the user (see Section 5), then the posterior distribution of $(\mu_k | \Sigma_k, \mathbf{v})$ is a multivariate normal distribution with mean

$\bar{\xi}_k = (n_k \bar{\mathbf{x}}_k + \tau_k \xi_k) / (n_k + \tau_k)$ and covariance matrix $\Sigma_k / (n_k + \tau_k)$. The posterior distribution of

$\lambda | \mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}$ is then given by

$$IG\left(\frac{m_0 + np}{2}, \frac{1}{2} \left\{ s_0 + \sum_k \text{tr} \left\{ \mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \xi_k) (\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0 \right) \right\} \right\} \right). \quad (6)$$

For the other parameters, we assume that

$$\Sigma_k \sim W_p^{-1}(m_0, \Psi_0)$$

has the random spectral decomposition $\Sigma_k = \mathbf{D}_k \mathbf{Q}_k \mathbf{D}_k^t = \lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$ with random eigenvalues

$q_{k1} \geq q_{k2} \geq \dots \geq q_{kp} \geq 0$, $\mathbf{Q}_k = \text{diag}(q_{k1}, \dots, q_{kp})$ and we define

$\lambda_k := q_{k1}$, $\mathbf{A} := \text{diag}(1, q_{k2}/q_{k1}, \dots, q_{kp}/q_{k1})$. In particular, we assume that $\mathbf{A} = \text{diag}(1, a_2, \dots, a_p)$

and \mathbf{D}_k are the shape and direction components of an inverse Wishart random variable

$W_p^{-1}(m_0, \Psi_0)$ (for the choice of m_0 , Ψ_0 and other priors, see again Section 5). If we assume that

\mathbf{A} and \mathbf{D}_k are *a priori independent* (Anderson 1984), the corresponding Gibbs sampler step is to simulate $a_j | \mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v}$, for $j = 1, \dots, p$, independently from the inverse gamma distribution

$$IG\left(\frac{1}{2}(n + K(m_0 + p) - 1), \frac{1}{2} \left\{ \sum_k \lambda^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0 \right) \mathbf{D}_k \right\}_{jj}\right).$$

Moreover, the \mathbf{D}_k 's are the principal direction vectors from the following inverse Wishart distribution

$$W_p^{-1}\left(n_k + m_0, \Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t\right). \quad (7)$$

(b) Model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$.

Again, \mathbf{D}_k and \mathbf{A} considered here are unknown. In addition, the k different groups are allowed to have different volumes λ_k . The prior distribution of λ_k is assumed to be an inverse gamma distribution

$$\lambda_k \sim IG(m_k/2, s_k/2)$$

independently for $k = 1, \dots, K$, and the corresponding Gibbs sampler step 3 is to simulate $a_j | \mathbf{D}_1, \dots, \mathbf{D}_K, \lambda_1, \dots, \lambda_K, \mathbf{v}$, for $j = 1, \dots, p$, independently from

$$IG\left(\frac{1}{2}(n + K(m_0 + p) - 1), \frac{1}{2} \left\{ \sum_k \lambda_k^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0 \right) \mathbf{D}_k \right\}_{jj}\right), \quad (8)$$

and $\lambda_k | \mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}$, for $k = 1, \dots, K$, independently from

$$IG\left(\frac{1}{2}(m_k + n_k p), \frac{1}{2} \left\{ s_k + \text{tr}[(\mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^t) \left((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0 \right)] \right\}\right). \quad (9)$$

The \mathbf{D}_k 's are simulated in the same way as in model (a).

(c) General model $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t]$

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 distributions on μ_k and Σ_k are assumed:

$$\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim N_p(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k / \tau_k), \quad \text{and} \quad \boldsymbol{\Sigma}_k \sim W_p^{-1}(m_k, \boldsymbol{\Psi}_k), \quad (k = 1, \dots, K),$$

and the corresponding Gibbs sampler step 3 is to simulate, $\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k, \mathbf{v}$, for $k = 1, \dots, K$, independently from

$$N_p(\bar{\boldsymbol{\xi}}_k, \boldsymbol{\Sigma}_k / (\tau_k + n_k))$$

where $\bar{\boldsymbol{\xi}}_k = (n_k \bar{\mathbf{x}}_k + \tau_k \boldsymbol{\xi}_k) / (n_k + \tau_k)$, and $\boldsymbol{\Sigma}_k | \mathbf{v}$, for $k = 1, \dots, K$, from

$$W_p^{-1}\left(n_k + m_k, \boldsymbol{\Psi}_k + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t\right).$$

3.2 Bayesian Model selection

So far we described the models of interest and working in a mixture-model framework, we will use the approximate Bayes factors to compare the models. For a review of Bayes factors, their calculation and their interpretation, see Kass and Raftery (1995). Here, we have to select not only the parametrization of the model but also the number of clusters K .

For simultaneously choosing between two models M_1, M_2 and deciding on the number of groups, we compute the approximate Bayes factor

$$BF_{1,2} = p(\mathbf{x} | M_2) / p(\mathbf{x} | M_1)$$

with

$$p(\mathbf{x} | M_h) = \int p(\mathbf{x} | \boldsymbol{\theta}_h) p(\boldsymbol{\theta}_h | M_h) d\boldsymbol{\theta}_h, \quad (10)$$

where $\boldsymbol{\theta}_h$ is the vector of parameters under the model M_h , and $p(\boldsymbol{\theta}_h | M_h)$ is its prior density ($h = 1, 2$). The quantity defined in (10) is called the *integrated likelihood* of model M_h . Bayesian model selection is based on Bayes factors, whose key ingredient is the integrated likelihood of a model. By convention, $\log(BF_{1,2}) < 2$ represents weak evidence for the model M_2 , differences between 2 and 6 represent positive evidence, differences between 6 to 10 represent strong

evidence, and differences > 10 represents very strong evidence (Jeffreys 1961). We approximate the integrated likelihood from the Gibbs sampler output using the *Laplace-Metropolis estimator* (Raftery 1996), which is very simple to calculate and was shown to give sufficiently accurate results by Lewis and Raftery (1997) and Bensmail et al. (1997). In the sequel, the word "model" refers to a combination of one of the models in Table 1 with a particular number of clusters K . Using the Laplace-Metropolis estimator, the Bayes factor becomes

$$BF_{1,2} = \frac{p(\mathbf{x}|M_2)}{p(\mathbf{x}|M_1)} = \frac{|\Psi^{(2)}|^{1/2} p(\mathbf{x}|\tilde{\theta}^{(2)}) p(\tilde{\theta}^{(2)})}{|\Psi^{(1)}|^{1/2} p(\mathbf{x}|\tilde{\theta}^{(1)}) p(\tilde{\theta}^{(1)})}, \quad (11)$$

where $\tilde{\theta}^{(h)}$, ($h = 1, 2$) is the posterior mode of $\theta^{(h)}$, denoting the parameters μ and Σ of the model M_h , and $\Psi^{(h)}$ is minus the inverse Hessian of $g(\theta) = \log p(\mathbf{x}|\theta)p(\theta)$ under the model h , evaluated at $\theta = \tilde{\theta}^{(h)}$. The Laplace method requires us to know the posterior mode, $\tilde{\theta}$, and $|\Psi|$. The Laplace-Metropolis estimator estimates these parameters from the Gibbs sampler output $\tilde{\pi}_k$, $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$. The likelihood at the approximate posterior mode is

$$\prod_{i=1}^n \sum_{k=1}^K \tilde{\pi}_k f_k(\mathbf{x}_i | \tilde{\mu}_k, \tilde{\Sigma}_k) \quad (12)$$

which is then substituted into equation (12) to obtain the Bayes factor.

For choosing the appropriate model, we calculate the Bayes factor for each pair of different combinations for a number of different clusters with the number of the components varying from $1, \dots, M$ for all models. This procedure is exemplified in Section 5.

4. A further extension by adding noise to the data

We have assumed that each observation belongs to a cluster. However, there may be some observations that do not follow that rule. Therefore, we consider the possibility of extending our models to include such irrelevant background or noise observations. Dasgupta and Raftery

(1998) 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 (optional) addition of a component described by a homogeneous spatial Poisson process, to represent “noise”.

They developed an EM algorithm to estimate the shape parameter of a particular model and a particular number of cluster, using maximum likelihood estimates of the parameter obtained by the EM algorithm of the maximized mixture likelihood.

Here, we assume a mixture of a Gaussian distribution satisfying (2), and of noise distributed as a homogeneous spatial Poisson process with constant rate π_0 so the general finite mixture distribution likelihood is

$$p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K; \pi_0, \pi_1, \dots, \pi_K | \mathbf{x}) = \prod_{i=1}^n \left[\frac{\pi_0}{\Lambda} + \sum_{k=1}^K \pi_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) \right] \quad (13)$$

where $f_k(\cdot | \boldsymbol{\theta}_k)$ is a $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density for $k = 1, \dots, K$, and $\Lambda > 0$ is the volume of the finite domain Ξ in which the data are assumed to be randomly distributed in R^p .

Operationally, the definition of the volume used here is

$\Lambda = \prod_{j=1}^p (\max_{i=1, \dots, n} \{x_{ij}\} - \min_{i=1, \dots, n} \{x_{ij}\})$, which is the volume of the smallest hyperrectangle with one side parallel to the coordinate axes containing the data. Other definitions could also be used, such as the volume enclosed by the convex hull of the data. Thus an observation contributes $1/\Lambda$ (a uniform density over the region Ξ of interest) if it belongs to the noise; otherwise it contributes a Gaussian term. The basic procedure for data including noise is realized by applying the augmented model to the entire data set, with the Gaussian components initialized by the empirical mean vector $\bar{\mathbf{x}}$ and variance matrix \mathbf{W} of the whole data set, and the noise component initialized with the volume of the whole data set (with volume defined as before). The Bayes factor (BF) is then used to select the best model representing the data when the noise has been removed.

The program “bayes-model-based” (Fortran version) is available . It may be downloaded from

<http://web.utk.edu/~hbensmai>. The SPLUS version of software is ready and will be available soon on the same address.

5. Examples

We present three examples to illustrate the ability of our approach to overcome the limitations described in Section 2. The first example uses simulated data. We simulated 200 points from a bivariate two-component Gaussian mixture and we add 5% and 10% noise which was generated by a Poisson process. The second and the third example are based on real data. For each example, we consider the models in Table 1. Our priors are chosen among conjugate priors 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, Lindman, and Savage 1963), and so it could be expected that the results would be relatively insensitive to reasonable changes in the prior.

We used $\xi_k = \bar{\mathbf{x}}$, $\tau_k = 1$, $m_k = m_0 = 5$, $s_k = s = \sigma^2$, and $\Psi_0 = \Psi_k = \mathbf{S}$, for $k = 1, \dots, K$, where $\bar{\mathbf{x}}$ and \mathbf{S} are the empirical mean vector and variance matrix of the whole data set, and σ^2 is the greatest eigenvalue of \mathbf{S} .

5.1 Example 1: Simulated Data

We simulated 200 points from a bivariate two-component Gaussian mixture ($K = 2, p = 2$) with equal proportions $\pi_1 = \pi_2 = 1/2$, mean vectors $\mu_1^t = \mu_2^t = (0, 0)$ and covariance matrices

$$\Sigma_1 = \text{diag}\left(\frac{1}{4}, \frac{(\alpha + 1)^2}{4}\right), \quad \Sigma_2 = \text{diag}\left(\frac{(\alpha + 1)^2}{4}, \frac{1}{4}\right)$$

The model used here is $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$, where $\mathbf{D}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{D}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

When $\alpha = 9$, we used $\lambda = 25$ and $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & .01 \end{pmatrix}$; when $\alpha = 3$, then $\lambda = 4$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & .625 \end{pmatrix}.$$

For both examples, 10 ($\pi_0 = .05$) and 20 ($\pi_0 = .10$) points are simulated from a homogeneous Poisson distribution in the hyperrectangle occupied by the data (see Figure 1: In the upper panels, the data are simulated using $\alpha = 9$ with 5% of noise (Figure 1a), and 10% of noise (Figure 1b). In the lower panels, the data are simulated using $\alpha = 3$, with 5% of noise (Figure 1c) and 10% of noise (Figure 1d).)

Insert Figure 1 about here.

Insert Figure 2 about here

Figure 2 shows for each data point its class assignment resulting from the maximum a posteriori rule with two groups after 500 iterations of the Gibbs sampler for the model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$. For Example (1a), convergence was immediate (see Figure 3a); similar results were obtained for Examples (1b), (1c) and (1d). Two noise points were classified as non-noise points for the Example (1a) and (1b). One noise point was classified as non-noise point for the Example (1c) and three noise points were classified as non-noise points for the Example (1d). For the other way around, there were two non-noise points which were classified as noise points for the Example (1b) and one non-noise point which was classified as noise point for the Example (1c).

Figure 3 shows the time series plot of the first 500 Gibbs sampler iterations for the mean of the group 1 (μ_1) and group 2 (μ_2) and for the shape parameters a_1 and a_2 .

Insert Figure 3 about here.

The model comparison results for Example (1a) are shown in Figure 4. The correct model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ (model 5) and the correct number of groups (2 clusters), for Example (1a) and (1b) are strongly favored. For Example (1a), the posterior modes of the parameters for the preferred model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ are $\hat{\mu}_1 = (.05, .003)$, $\hat{\mu}_2 = (.06, .01)$, $\hat{\lambda} = 24.8$, $\hat{\mathbf{A}} = \text{diag}(.89, .01)$, $\hat{\mathbf{D}}_1 = \begin{pmatrix} .001 & .9 \\ .9 & .001 \end{pmatrix}$ and $\hat{\mathbf{D}}_2 = \begin{pmatrix} .9 & 0. \\ 0. & .9 \end{pmatrix}$, which are close to the true values.

For Example (1c) or (1d), the correct model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ (model 5) and the correct number of groups (2 clusters) are strongly favored. For Example (1d), the posterior modes of the parameters for the preferred model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ are $\hat{\mu}_1 = (.00, .001)$, $\hat{\mu}_2 = (.03, .01)$, $\hat{\lambda} = 3.9$, $\hat{\mathbf{A}} = \text{diag}(.99, .61)$, $\hat{\mathbf{D}}_1 = \begin{pmatrix} .00 & .98 \\ .98 & .00 \end{pmatrix}$ and $\hat{\mathbf{D}}_2 = \begin{pmatrix} .99 & .001 \\ .001 & .99 \end{pmatrix}$ which are close to the true values.

Insert Figure 4 about here.

Perhaps one of the greatest advantages of the present approach is that it fully assesses uncertainty about group membership, rather than merely giving a single "best" partition.

5.2 Example 2: Butterfly classification

Figure 5 shows four shape measurements of a butterfly. Data on two of these measurements, \mathbf{z}_3 and \mathbf{z}_4 , for 23 butterflies are shown in Figure 6. The example is taken from Celeux and Robert (1993) and was analyzed by Bensmail et al. (1997) using only the two measurements \mathbf{z}_3 and \mathbf{z}_4 . Here we are using the four wing measurements \mathbf{z}_1 , \mathbf{z}_2 , \mathbf{z}_3 and \mathbf{z}_4 to decide how many species are represented in this group of insects, and how to classify them.

Insert Figure 5, 6, 7, 8 about here.

Figure 7 shows that the best fit is obtained for model 5 ($[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$), with three groups of butterflies. Model 3 ($[\lambda \mathbf{D} \mathbf{A} \mathbf{D}^t]$) comes very close, followed by model 6 ($[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$) and the general model 7 ($[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t]$). The models 1, 2 and 4 do not fit at all. Bensmail et al. (1997) obtained 4 groups with the model $[\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^t]$ where the fourth group contains only the observation 23 considered as an aberrant butterfly. The present approach considers butterfly 23 as noise, confirming the previous result and, moreover, it gives more information on the clusters by estimating the features shape, volume and orientation of each. To investigate the obtained classification with respect to the original measurements, we performed a canonical discriminant analysis on the classified individuals. The results are given in Figure 8. The butterflies are displayed in a two-dimensional canonical space, labeled with their group symbols. The clouds of points for the different groups are clearly separated. To investigate the relation with the variables, these latter ones are depicted as vectors, projected in the canonical space (Meulman, Zeppa, Boon, and Rietveld 1992). 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 (Meulman 1986). The variables seem to consist of three subsets; subset 1 contains the two wing measurements \mathbf{z}_2 and \mathbf{z}_3 and separates groups 1 and 2 from group 3 (and the outlying butterfly 23). The second subset, consisting of the wing measurement \mathbf{z}_4 , separates group 2 and group 3 from group 1 and the third subset, of the wing measurement \mathbf{z}_1 , finally, separates the groups 1 and 3 from group 2.

5.3 Example 3: Diabetes Data

Reaven and Miller (1979) described and analyzed data for 145 subjects, these variables are: the area under a plasma glucose curve (glucose area), the area under a plasma insulin curve (insulin area), and the steady-state plasma glucose response (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. The data show the three-dimensional shape of a boomerang with two wings and a fat middle where Insulin Area is plotted against Glucose Area. Figure 9 shows three two-dimensional projections of the three-dimensional diabetes data of Reaven and Miller (1979). Figure 9d shows the results of the present Bayesian model-based classification.

Insert Figure 9 about here.

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) used the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$, where $\mathbf{A} = \text{diag}\{1, \alpha, \alpha\}$. The result reported for their algorithm is $\hat{\alpha} = 0.2$. We used our approach developed in the previous section to classify the data. Figure 10 shows that the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ with three groups is favored quite strongly over the alternatives. Especially the model 1 and 2 ($[\lambda \mathbf{I}]$ and $[\lambda_k \mathbf{I}]$) fit very badly.

Insert Figure 10 about here.

The estimated values of the shape parameters \hat{a}_2 and \hat{a}_3 for the three groups are .19 and .15 ($\hat{a}_1 = 1$) and the values of the volume parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are .035, .42 and .14. This

result shows that the three clusters, with different sizes (λ_k) and orientations (\mathbf{D}_k), but the same "tubular shape", are clustered circularly in a 2-dimensional subspace in R^3 .

The optimal classification resulted in only 9% of the points being misclassified, as indicated in Table 2. The misclassified points are mostly situated at the border of the clusters. A 10% error rate was obtained by the Banfield and Raftery algorithm, but the misclassified points were dispersed over the three clusters. The present results compare favorably our approach to the procedures of Symons (1981) where the misclassification rates vary from 13 to 26 percent.

Insert Table 2 about here.

Figure 11 shows the plot of the canonical discriminant analysis results for the diabetes data. The vectors are again projections of the original variables in the canonical space obtained by the discriminant analysis. The length of the line drawn from the origin to each variable arrow head represents the 'importance' of that variable. The canonical variables are represented by horizontal and vertical lines drawn through the origin. The diabetes subjects are represented with their group symbols. The clouds of points for the different groups are not well separated.

Insert Figure 11 about here.

The variable insulin arranges the groups from chemical diabetes (Type 1), normal (nondiabetic) to overt diabetes (Type 2). The variable glucose arranges 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.

6. 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.

Our particular interest has been to extend Bensmail et al. (1997) work in two respects, using the fully Bayesian inference and Markov chain Monte Carlo sampling to estimate the parameters involved in each model. First, the range of models has been extended to a family of models where the covariance matrix Σ_k is represented by $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$, $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$ and $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t$ respectively, where all the parameters $\lambda, \lambda_k, \mathbf{A}, \mathbf{A}_k$ and \mathbf{D}_k ($k = 1, \dots, K$) are unknown. These models, especially the first one, often turn out to be selected by the Bayes factor as being preferable to previous, more constrained models. Second, the range of applications has been extended to the case where the data contain outliers. The parameters involved in the parameterization of the covariance matrix Σ_k are unknown and not constrained, as in Banfield and Raftery (1993), and Dasgupta and Raftery (1998).

Alternative approaches consist of maximizing the mixture likelihood using the EM algorithm. Dasgupta and Raftery (1998) considered the model $[\mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ in which the features of interest are presented by multivariate normal densities with high linearity, specified by the *shape* \mathbf{A} being the same across features, what 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.

Moreover, we have proposed a way of generalizing and overcoming the limitations of Bensmail et al. (1997), Dasgupta and Raftery (1998) related procedures for classification. These are the inability to specify some but not all features to be constant across clusters (Dasgupta and Raftery (1998), Banfield and Raftery (1993); the limitation of Bensmail et al. (1997) by considering a Bayesian clustering method based only on simple models and the failure to extend it 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 the context of Gaussian clustering, we applied canonical discriminant analysis, on the classified individuals, as a tool after 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, verify the importance of the variables in the canonical space, and interpret the position of the different clusters.

References

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

BANFIELD, J. D. and RAFTERY, A. E., (1993), "Model-Based Gaussian and Non Gaussian Clustering", *Biometrics*, 49, 803-821.

BENSMAIL, H., CELEUX, G., RAFTERY, A. E. and ROBERT, C. (1997), "Inference in Model-Based Cluster Analysis", *Computing and Statistics*, 7, 1-10.

BINDER, D. H. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65 (1), 31-38.

BINDER, D. H. (1981), "Approximations to Bayesian Clustering Rules," *Biometrika*, 68 (1), 275-285.

BOCK, H. H. (1985), "On Some Significance Tests in Cluster Analysis," *Journal of Classification*, 2, 77-108.

BOCK, H. H. (1996), "Probability Models in Partitional Cluster Analysis," *Computational Statistics and Data Analysis*, 23, 5-28.

BOZDOGAN, H. (1993), "Choosing the Number of Component Clusters in the Mixture Model Using a New Informational Complexity Criterion of the Inverse Fisher Information Matrix," O. Opitz, B. Lausen, and R. Klar, Eds., *Information and Classification*, Springer-Verlag, 40-54.

CELEUX, G., and ROBERT, C. (1993), "Une Histoire de Discretisation (avec Commentaires),"

La Revue de Modulad, 11, 7-44.

CELEUX, G., and SOROMENHO, G. (1996), "An entropy Criterion for Assessing the Number of Clusters in a Mixture Model", *Journal of Classification*, 13(1), 1996.

DASGUPTA, A., and RAFTERY, A. E. (1998), "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 294-302.

DIEBOLT, J., and ROBERT, C. P.(1994), "Bayesian Estimation of Finite Mixture Distributions," *Journal of of the Royal Statistical Society, Series B*, 56, 363-375.

EDWARDS, W., LINDMAN, H., and SAVAGE, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193-242.

ENGELMAN, L., and HARTIGAN, J. A. (1969), "Percentage Points of a Test for Clusters," *Journal of the American Statistical Association*, 64, 1647-1648.

FRALEY, C. (1999), "Algorithms for Model-Based Gaussian Hierarchical Clustering," *SIAM Journal on Scientific Computing*, 20, 270-281.

FRALEY, C., and RAFTERY, A.(1998), "How Many Clusters? Which Clustering Method? - Answers via Model-Based Cluster Analysis", *Computer Journal*, 41, 578-588.

GELMAN, A., CARLIN, J. B., STERN, H. S., and RUBIN, D. B. (1995), *Bayesian data analysis*. Chapman and Hall.

GILKS, W. R., RICHARDSON, S., SPIEGELHALTER, D., J. (1996), "*Markov Chain Monte Carlo Methods in Practice*". New York: Chapman and Hall.

GREEN, P. G., (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination", *Biometrika*, 82, 711-732 (1995).

GORDON, A. D. (1999), *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, 2nd Eds. New York.

- HARTIGAN, J. A. (1975), *Clustering Algorithms*, Wiley, New York.
- JEFFREYS, H. (1961) , *Theory of Probability*, Clarendon.
- KASS, R. E., and RAFTERY, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.
- KAUFMAN, L., and ROUSSEEUW, P. J. (1990), *Finding Groups in Data*, Wiley, New York.
- LAVINE, M., and WEST, M. (1992) , "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics*, 20, 451-461.
- LEROUX, B. G., (1992), "Consistent Estimation of a Mixing Distribution", *The Annals of Statistics*, 20, 1350-1360.
- 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*, 92, 438, 648-655.
- MACQUEEN, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281-297.
- MCLACHLAN, G. (1982), *The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis*, (Vol. 2), Amsterdam: Handbook of Statistics, in P. R. Krishnaiah and L. N. Kanal, 199-208.
- MCLACHLAN, G., and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.
- MCLACHLAN, G., and Peel, D. (2000). *Finite Mixture Models*. New York; Wiley.
- MENZEFRICKE, U. (1981), "Bayesian Clustering of Data Sets," *Communication in Statistics-Theory and Methods A.*, 10, 65-77.
- MEULMAN, J. J. (1986), *A Distance Approach to Nonlinear Multivariate Analysis*, DSWO Press,

Leiden.

MEULMAN, J. J., ZEPPA, P., BOON, M. E., and 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.

MUKHERJEE, M., and FEIGELSON, E. D., BABU, G. J., MURTAGH, F., FRALEY, C., and RAFTERY, A. (1998), "Three Types of Gamma Ray Bursts," *Astrophysical Journal*, 508, 314-327.

MURTAGH, F., and RAFTERY, A. (1984), "Fitting Straight Lines to Point Patterns", *Pattern Recognition*, 17, 479-483.

RAFTERY, A. E. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, 83, 251-266.

REAVEN, G. M., and MILLER, R. G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis," *Diabetologia*, 16, 17-24.

RICHARDSON, S., and GREEN, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society, Series B*, 59", 731-792.

ROEDER, K., and WASSERMAN, L. (1997), "Practical Bayesian Density Estimation Using Mixture of Normals," *Journal of the American Statistical Association*, 92, 439, 894-902.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461-464.

SCOTT, A. J., and SYMONS, M. J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics*, 27, 387-397.

SYMONS, M (1981), "Clustering Criteria and Multivariate Normal Mixtures," *Biometrics*, 37,

35-43.

TANNER, M., and WONG, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation (with Discussion)," *Journal of the American Statistical Association*, 82, 528-550.

TIERNEY, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22, 1701-1762.

WEI, G. C. G, and TANNER M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms", *Journal of the American Statistical Association*, 85, 699-704.

WOLFE, J. H. (1978), "Comparative Cluster Analysis of Patterns of Vocational Interest," *Multivariate Behavioral Research* ,13, 33-44.

Appendix: Gibbs Sampling for the Clustering Models

(a) Model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$.

The prior distribution of $(\mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$ is determined by the prior distribution of $(\mu_1, \dots, \mu_K, \lambda, \mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K)$ and we assume that $\lambda, \mu_1, \dots, \mu_K | (\Sigma_1, \dots, \Sigma_K)$ are independent with

$$\lambda | \Sigma_1, \dots, \Sigma_K \sim IG(m_0/2, s_0/2)$$

$$\mu_k | \Sigma_1, \dots, \Sigma_K \sim N_p(\xi_k, \Sigma_k / \tau_k)$$

$(\mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K)$ are independent and distributed as the following:

We suppose that:

$$\Sigma_k \sim W_p^{-1}(m_0, \Psi_0)$$

has the random spectral decomposition $\Sigma_k = \mathbf{D}_k \mathbf{Q}_k \mathbf{D}_k^t = \lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$ with random eigenvalues

$q_{k1} \geq q_{k2} \geq \dots \geq q_{kp} \geq 0$, $\mathbf{Q}_k = \text{diag}(q_{k1}, \dots, q_{kp})$, and we define we define

$\lambda_k := q_{k1}$, $\mathbf{A} := \text{diag}(1, q_{k2}/q_{k1}, \dots, q_{kp}/q_{k1}) = \text{diag}(1, a_2, \dots, a_p)$. This is true only asymptotically

(i.e. as the number of degrees of freedom $(n - K)$ goes to infinity; Anderson 1984) but it considerably simplifies the simulation, with moderate effects (if any) on the resulting posterior distribution. We derive then from the posterior distribution density of $\mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v}$

$$\prod_k |\mathbf{A}|^{-(n_k-1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left(\lambda^{-1} \mathbf{A}^{-1} \sum_k \mathbf{D}_k^t ((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k \tau_k}{n_k + \tau_k} + \mathbf{W}_k) \mathbf{D}_k \right) \right. \\ \left. |\mathbf{A}|^{-(m_0+p+1)/2} \exp \left\{ -\text{tr}(\lambda^{-1} \mathbf{D}_k^t \mathbf{A}^{-1} \mathbf{D}_k \Psi_0) / 2 \right\} \right. \quad (14)$$

the distribution of $\mathbf{A} | \mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v}$. Thus we have

$$\mathbf{A} | \mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v} \sim \pi(\mathbf{A}) \propto |\mathbf{A}|^{-(n-K+K(m_0+p+1))/2} \times \quad (15)$$

$$\exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{A}^{-1}\left[\sum_k \lambda^{-1}\mathbf{D}_k^t((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k\tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0)\mathbf{D}_k\right]\right)\right\},$$

which is a condensed way of saying that the diagonal elements of \mathbf{A} are independently distributed according to some inverted gamma distributions, i.e., that

$$a_j|\mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v} \sim \text{IG}\left(\frac{1}{2}(n + K(m_0 + p) - 1), \frac{1}{2}\left\{\lambda \sum_k \mathbf{D}_k^t((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k\tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0)\mathbf{D}_k\right\}_{jj}\right). \quad (16)$$

The \mathbf{D}_k 's are then independently distributed *a posteriori* as the principal direction vectors from the following inverse Wishart distribution,

$$W_p^{-1}\left(n_k + m_0, \Psi_0 + \mathbf{W}_k + \frac{n_k\tau_k}{n_k + \tau_k}(\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t\right). \quad (17)$$

In our algorithm, the four Gibbs components of Step 3 are the following:

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

$$(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \mathbf{v}) \sim N_p(\bar{\xi}_k, \boldsymbol{\Sigma}_k/(n_k + \tau_k)); \quad \bar{\xi}_k = (n_k\bar{\mathbf{x}}_k + \tau_k\xi_k)/(n_k + \tau_k). \quad (18)$$

3.2 Simulate

$$(\lambda|\mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}) \sim \text{IG}\left(\frac{m_0 + np}{2}, \frac{1}{2}\left\{s_0 + \sum_k \text{tr}\{\mathbf{D}_k\mathbf{A}^{-1}\mathbf{D}_k^t((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k\tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0)\}\right\}\right); \quad (19)$$

3.3 For $j = 1, \dots, p$, simulate

$$a_j|\mathbf{D}_1, \dots, \mathbf{D}_K, \lambda, \mathbf{v} \sim \text{IG}\left(\frac{1}{2}(n + K(m_0 + p) - 1), \frac{1}{2}\left\{\sum_k \lambda^{-1}\mathbf{D}_k^t((\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^t \frac{n_k\tau_k}{n_k + \tau_k} + \mathbf{W}_k + \Psi_0)\mathbf{D}_k\right\}_{jj}\right). \quad (20)$$

3.4 For $k = 1, \dots, K$, calculate the principal directions vectors from

$$\mathbf{W}_p^{-1} \left(n_k + m_0, \boldsymbol{\Psi}_0 + \mathbf{W}_k + \frac{n_k \boldsymbol{\tau}_k}{n_k + \boldsymbol{\tau}_k} (\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t \right). \quad (21)$$

(b) Model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$.

The prior distributions are similar for all components:

$$(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k) \sim N_p(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k / \boldsymbol{\tau}_k), \text{ and } \lambda_k \sim IG(m_k/2, s_k/2), \quad k = 1, \dots, K. \quad (22)$$

The simulation of \mathbf{A} is also quite close to the previous version since

$$\begin{aligned} (\mathbf{A} | \lambda_1, \dots, \lambda_K, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}) &\sim \pi(\mathbf{A}) \propto |\mathbf{A}|^{-(n-K+K(m_0+p+1))/2} \times \\ &\exp - \frac{\text{tr}}{2} \left(\mathbf{A}^{-1} \sum_k \lambda_k^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t \frac{n_k \boldsymbol{\tau}_k}{n_k + \boldsymbol{\tau}_k} + \mathbf{W}_k + \boldsymbol{\Psi}_0 \right) \mathbf{D}_k \right). \end{aligned} \quad (23)$$

In our algorithm, the four Gibbs components of Step 3 are then:

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

$$(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k, \mathbf{v}) \sim N_p(\bar{\boldsymbol{\xi}}_k, \boldsymbol{\Sigma}_k / (n_k + \boldsymbol{\tau}_k)); \quad \bar{\boldsymbol{\xi}}_k = (n_k \bar{\mathbf{x}}_k + \boldsymbol{\tau}_k \boldsymbol{\xi}_k) / (n_k + \boldsymbol{\tau}_k). \quad (24)$$

3.2 For $k = 1, \dots, K$, simulate

$$\begin{aligned} (\lambda_k | \mathbf{A}, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}) &\sim \\ IG \left(\frac{m_k + n_k p}{2}, \frac{1}{2} \left\{ s_k + \text{tr} \left\{ \mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t \frac{n_k \boldsymbol{\tau}_k}{n_k + \boldsymbol{\tau}_k} + \mathbf{W}_k + \boldsymbol{\Psi}_0 \right) \right\} \right\} \right); \end{aligned} \quad (25)$$

3.3 For $j = 1, \dots, p$, simulate

$$\begin{aligned} (\mathbf{a}_j | \lambda_1, \dots, \lambda_K, \mathbf{D}_1, \dots, \mathbf{D}_K, \mathbf{v}) &\sim \\ IG \left(\frac{n + K(m_0 + p) - 1}{2}, \frac{1}{2} \left\{ \sum_k \lambda_k^{-1} \mathbf{D}_k^t \left((\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t \frac{n_k \boldsymbol{\tau}_k}{n_k + \boldsymbol{\tau}_k} + \mathbf{W}_k + \boldsymbol{\Psi}_0 \right) \mathbf{D}_k \right\}_{jj} \right). \end{aligned} \quad (26)$$

3.4 For $k = 1, \dots, K$, calculate the principal directions vectors from

$$\mathbf{W}_p^{-1} \left(n_k + m_0, \boldsymbol{\Psi}_0 + \mathbf{W}_k + \frac{n_k \boldsymbol{\tau}_k}{n_k + \boldsymbol{\tau}_k} (\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t \right). \quad (27)$$

Table 1: Parametrization of the covariance matrix Σ_k in the Gaussian model and its geometric interpretation. The entries indicates whether the feature of interest (shape, orientation or volume) is the same for each group or not.

Model	Σ_k	Shape(A_k)	Direction (D_k)	Volume (λ_k)
1.	λI	Spherical (same)	N/A	same
2.	$\lambda_k I$	Spherical (same)	N/A	different
3.	λDAD^t	Elliptical (same)	same	same
4.	$\lambda_k DAD^t$	Elliptical (same)	same	different
5.	$\lambda D_k A D_k^t$	Elliptical (same)	different	same
6.	$\lambda_k D_k A D_k^t$	Elliptical (same)	different	different
7.	$\lambda_k D_k A_k D_k^t$	Elliptical (different)	different	different

data. Table 2: Cross-classification table giving the clustering results for the diabetes

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

Figure Captions

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

Figure 2: Results of the classification of the four simulated data sets, produced by the best model ($[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$) indicated by the Bayes factor (BF). The black squares represent noise. There were two noise points which were classified as non-noise points for the Example (1a) and (1b), one noise point which was classified as non-noise point for the Example (1c) and three noise points which were classified as non-noise points for the Example (1d). For the other way around, there were two non-noise points which were classified as noise points for the Example (1b) and one non-noise point which was classified as noise point for the Example (1c).

Figure 3: Time series plot of the first 500 Gibbs sampler iterations for the mean of the group 1 and group 2 (two first plots) and for the shape parameters a_1 and a_2 (third plot).

Figure 4: Bayes factors for the model-based methods applied to the simulated data (Figure 1b). The first local (also global) maximum occurs for the model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ (model 5; constant shape, equal volume) with two clusters.

Figure 5: Four butterfly measurements (z_1, z_2, z_3 and z_4).

Figure 6: This plot shows the values (z_3, z_4) for the 23 butterflies.

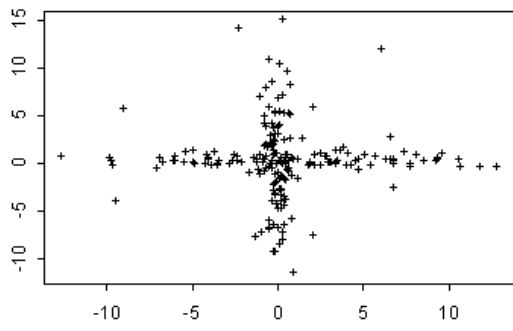
Figure 7: Bayes factors for model-based methods. The first local (also global) maximum occurs for the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$ (model 6; constant shape, varying volume) with three clusters.

Figure 8: Classification and projection of the four measurements of the 23 butterflies, onto the canonical space, for $K = 3$ classes.

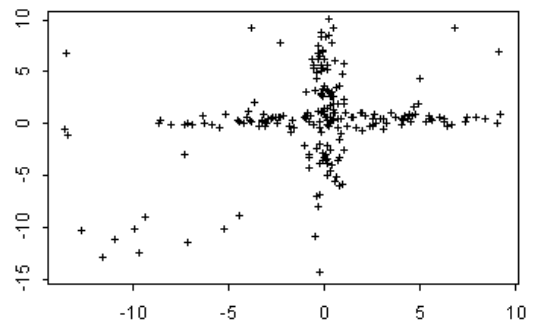
Figure 9: Three two-dimensional projections of the three-dimensional diabetes data. The symbols indicate the clinical classification of subjects as having chemical diabetes, overt diabetes, or being normal. Lower right panel shows the three clusters in the diabetes data found by our approach using the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$ (model 6; constant shape, varying volume).

Figure 10: Bayes factors for model-based methods, the model $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$ with three clusters is chosen.

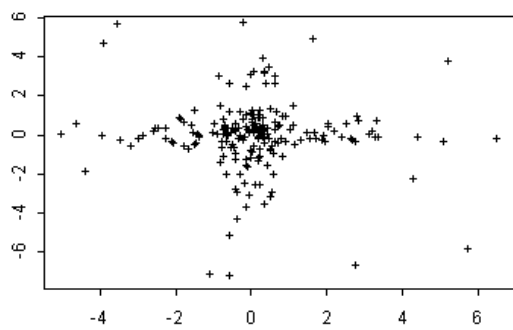
Figure 11: Classification, canonical projection of diabetes data and representation of the variables in the canonical space.



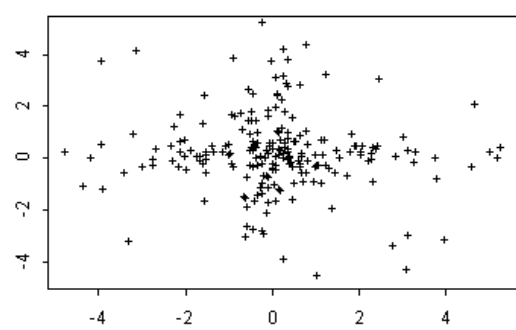
(1a)



(1b)

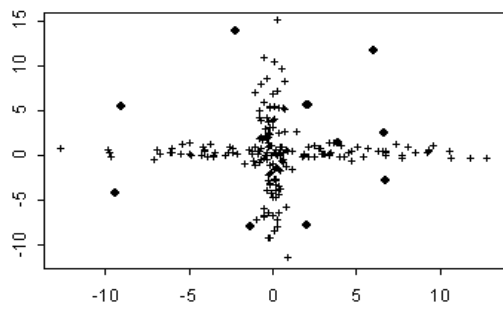


(1c)

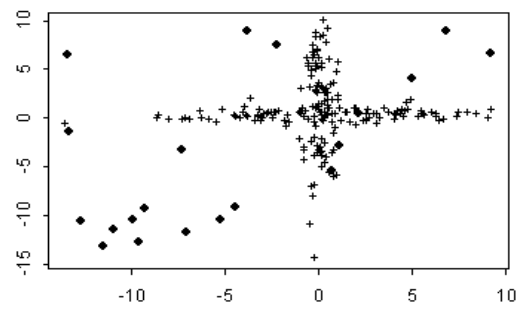


(1d)

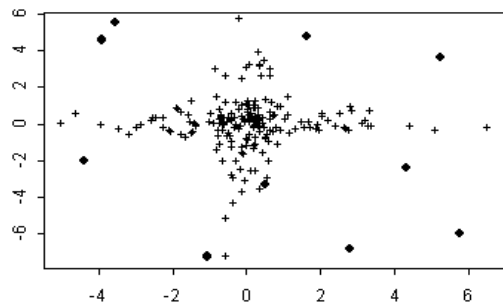
Figure 1



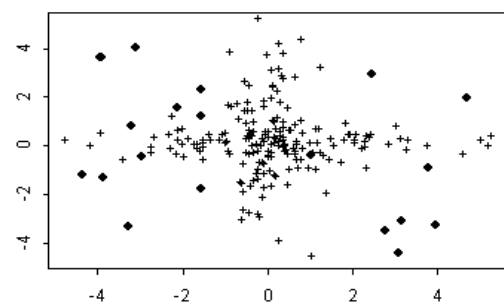
(2a)



(2b)



(2c)



(2d)

Figure 2

TIME SERIES OF THE PARAMETERS: vector means and shape matrix

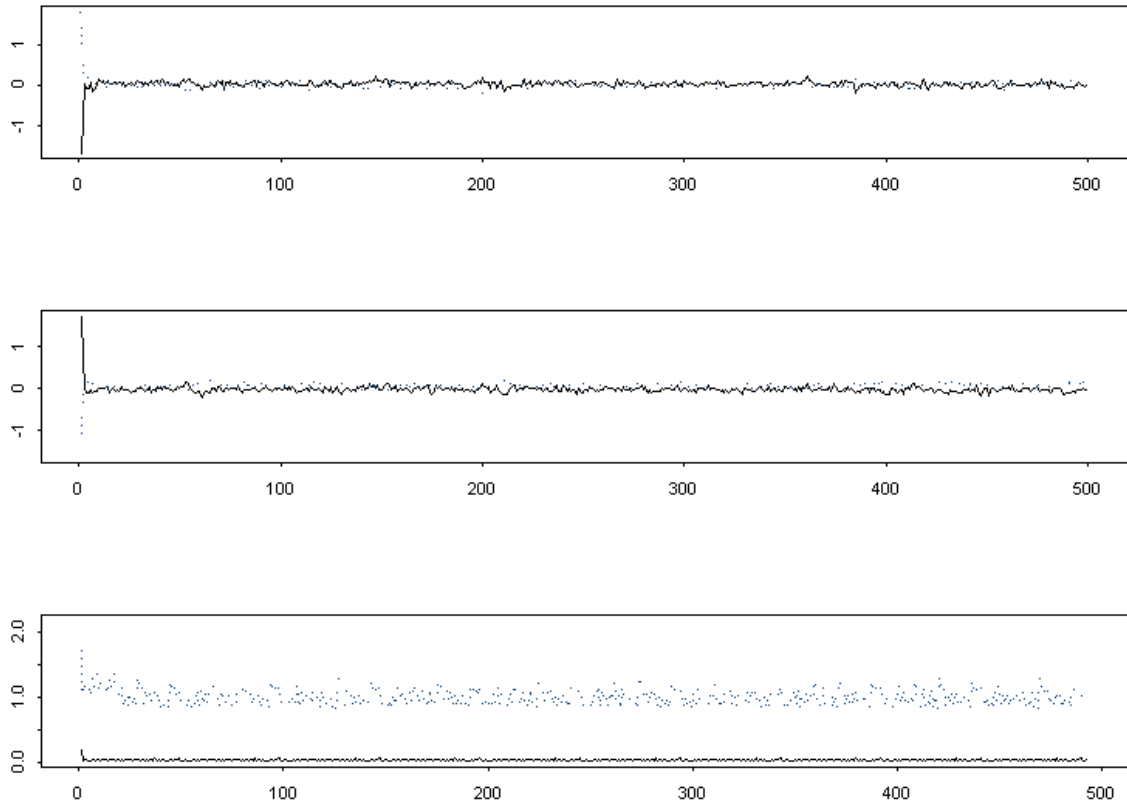


Figure 3

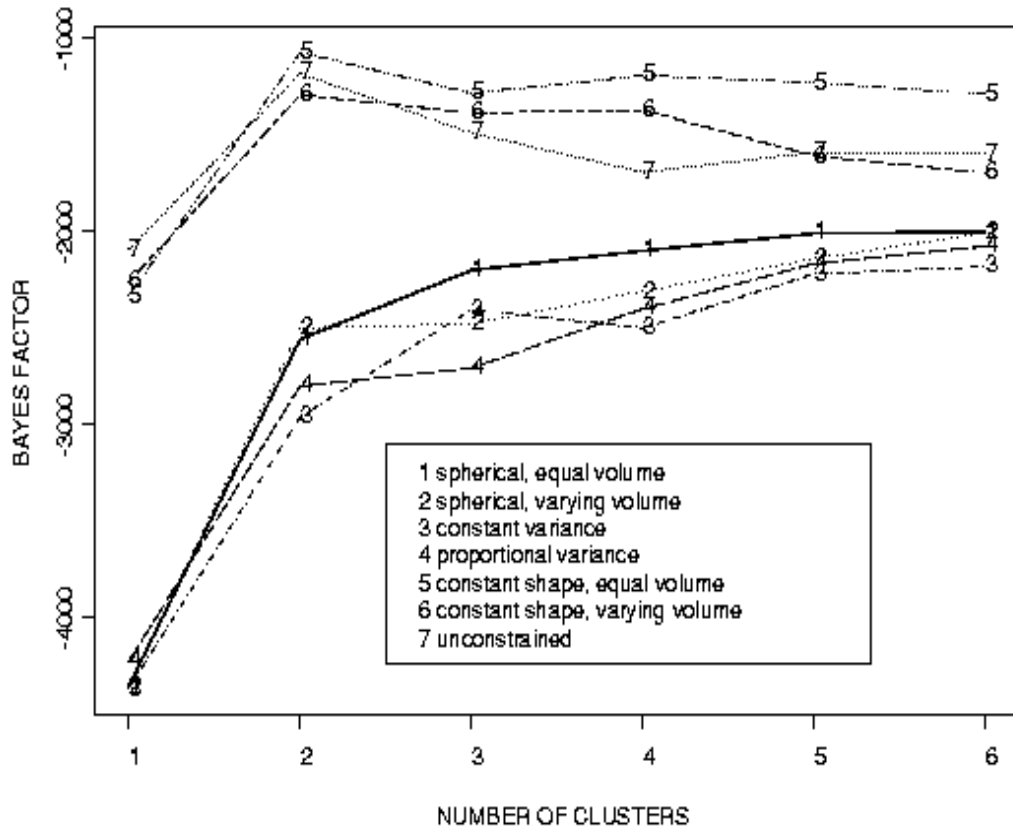


Figure 4.

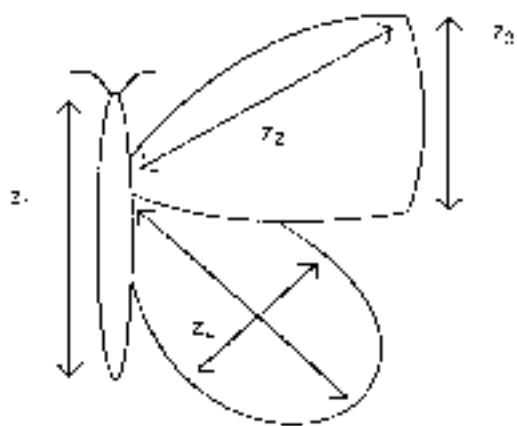


Figure 5

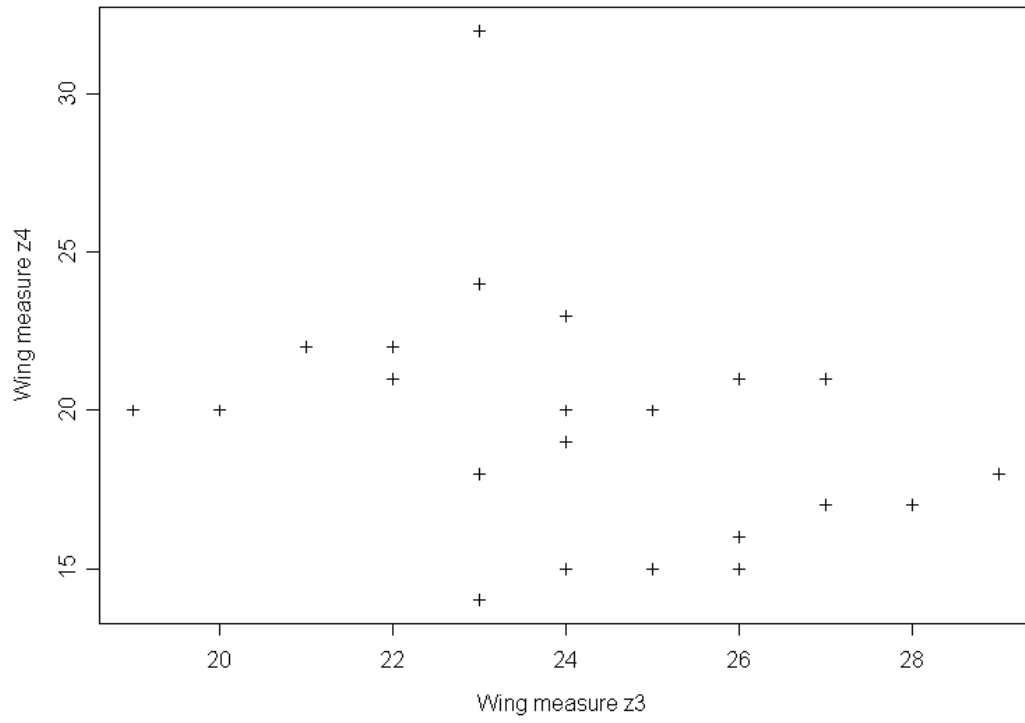


Figure 6

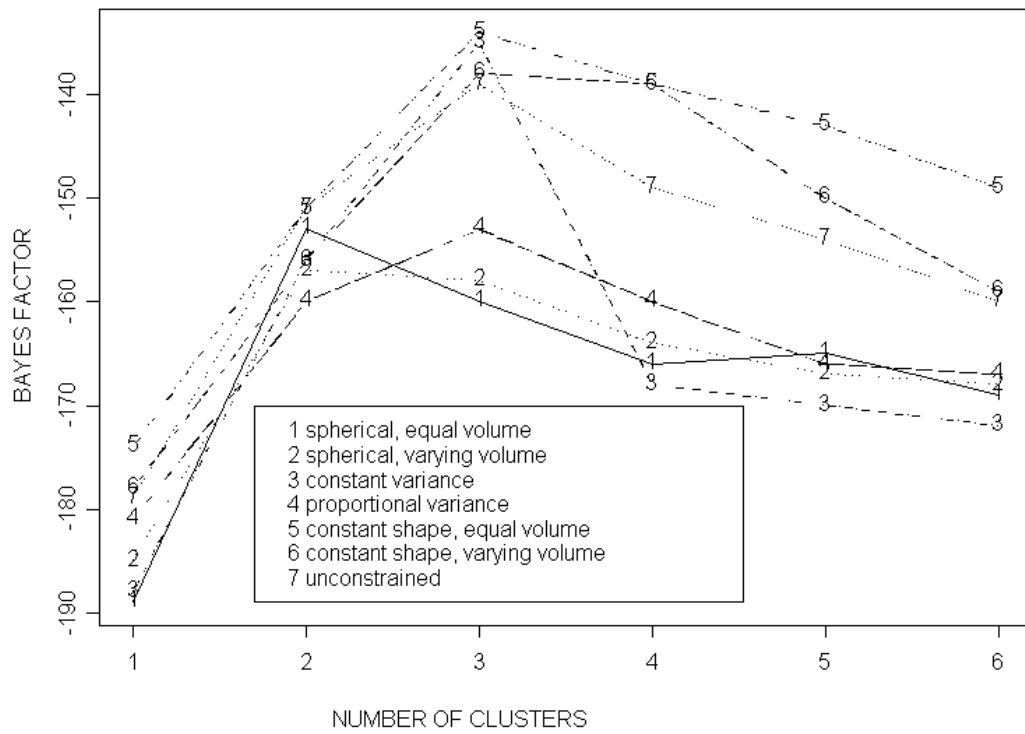


Figure 7

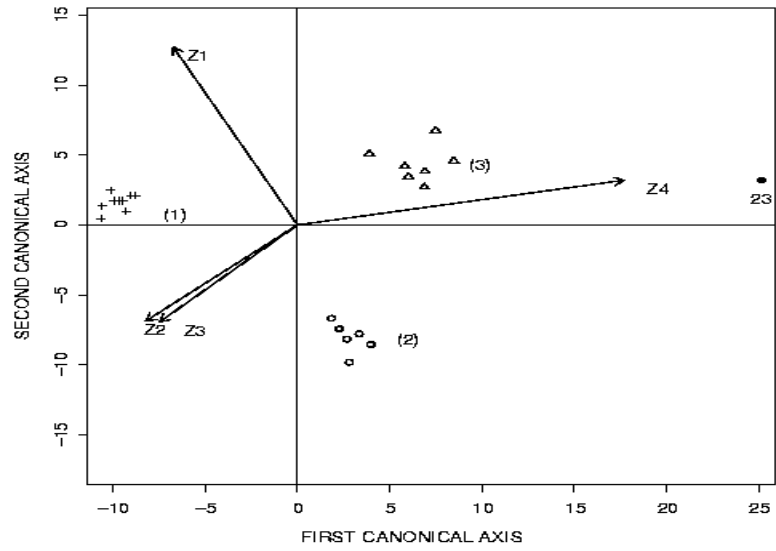


Figure 8

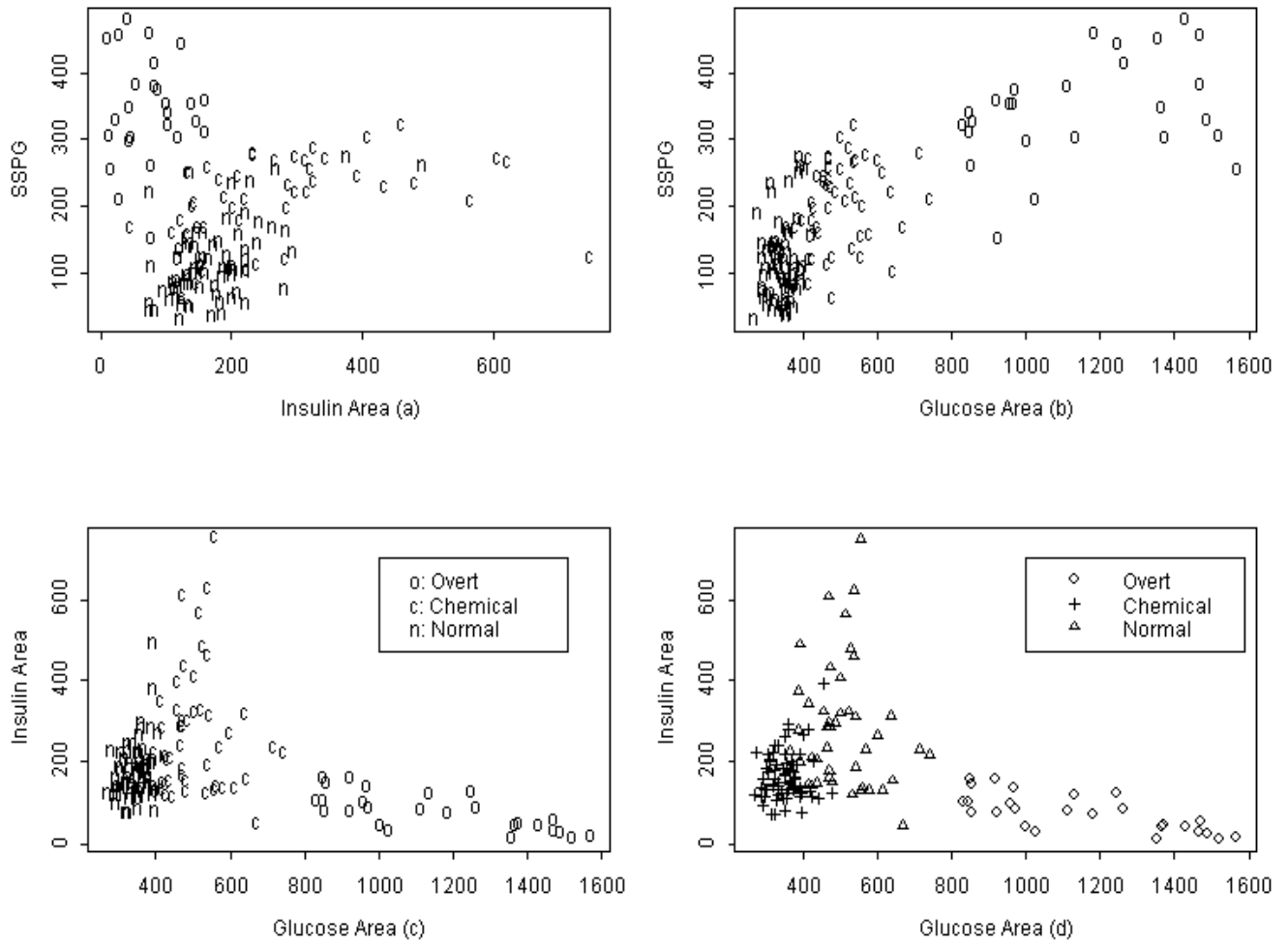


Figure 9

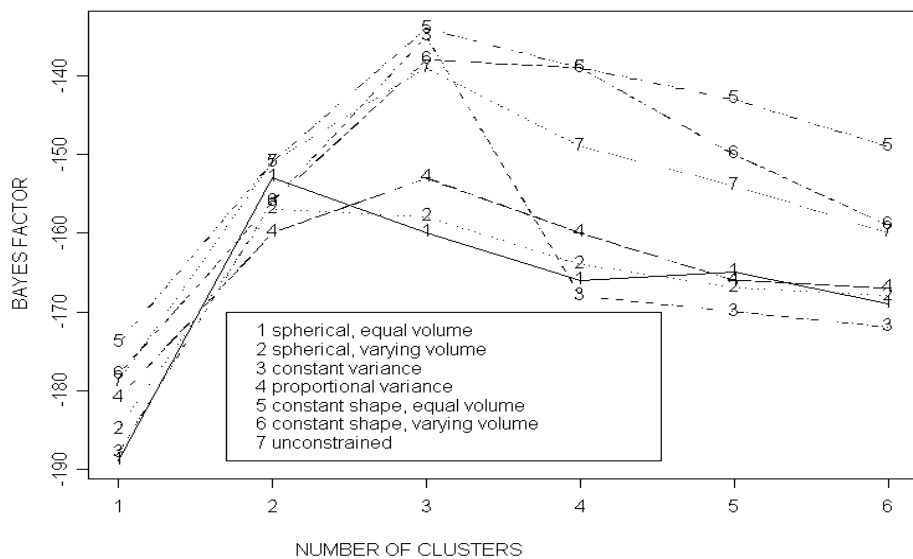


Figure 10

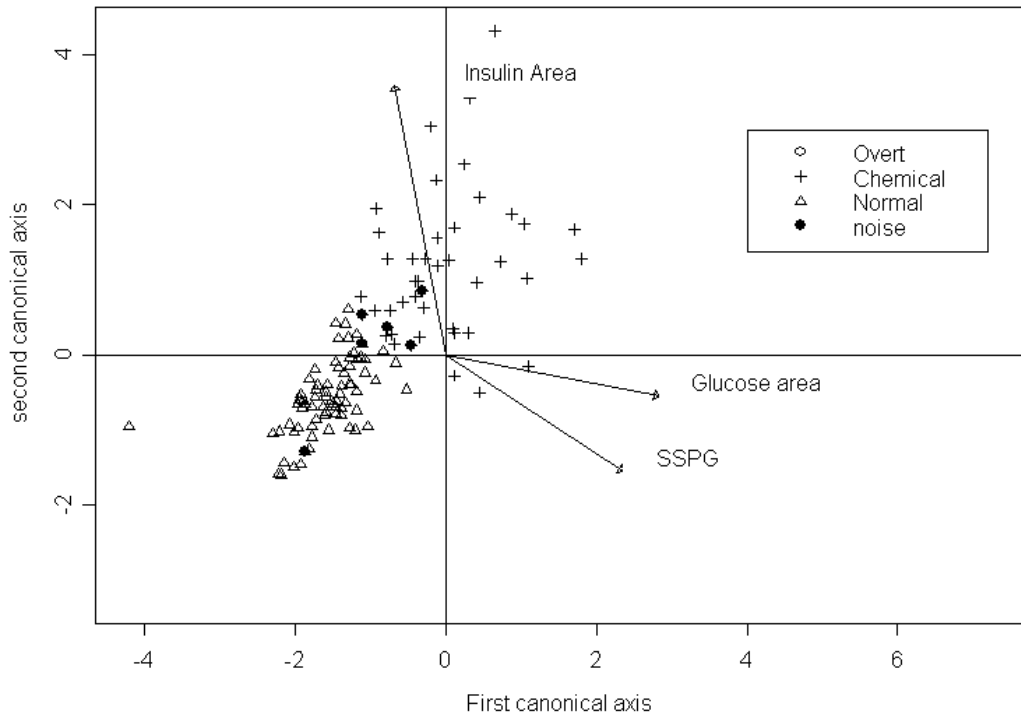


Figure 11