Stat Comput (2011) 21:511–522
DOI 10.1007/s11222-010-9188-x

Finite mixtures of matrix normal distributions for classifying three-way data

Cinzia Viroli

Received: 21 December 2009 / Accepted: 17 May 2010 / Published online: 26 May 2010

Abstract Matrix-variate distributions represent a natural way for modeling random matrices. Realizations from random matrices are generated by the simultaneous observation of variables in different situations or locations, and are commonly arranged in three-way data structures. Among the matrix-variate distributions, the matrix normal density plays the same pivotal role as the multivariate normal distribution in the family of multivariate distributions. In this work we define and explore finite mixtures of matrix normals. An EM algorithm for the model estimation is developed and some useful properties are demonstrated. We finally show that the proposed mixture model can be a powerful tool for classifying three-way data both in supervised and unsupervised problems. A simulation study and some real examples are presented.

Keywords Model based clustering · Random matrix · Three-way data · EM-algorithm

1 Introduction

Finite mixture models provide a particularly suitable method for clustering data coming from different sub-populations (Wolfe 1970; Ganesalingam and McLachlan 1979; McLachlan 1982; McLachlan and Basford 1988; McLachlan and Peel 2000). They are generally applied to data expressed in a two-way structure, i.e. units observed in different situations or with respect to some characteristics. For continuous observed variables, finite mixtures of normals have been widely investigated and applied in many situations, under the framework of model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 2002a), and thanks to the availability of an efficient software (Fraley and Raftery 1999, 2002b, 2003).

In this paper we define, develop and discuss finite mixtures of matrix normals, and show their potentiality as a tool for classifying three-way data.

Three-way data sets can occur from the observation of various attributes measured on a set of units in different situations or occasions. Longitudinal data on multiple response variables or spatial multivariate data represent some typical examples. Alternatively, they can be the result of one measurement on some units in different time points and locations, thus leading to spatio-temporal data. Other examples arise when some objects are rated on multiple attributes by multiple experts or from experiments in which individuals provide multiple ratings for multiple objects (Vermunt 2007). Symbolic data represent another example as well, provided that the complex information can be structured as multiple values for each variable (Billard and Diday 2003). All these data examples can be arranged in a three-way structure. More specifically, suppose we observe a p-variate response in r occasions, or a univariate variable in p locations and r times. Both situations yield an $r \times p$ observed matrix, Y_j, for each statistical unit, with $j = 1, \dots, n$. We assume that a random sampling of n individuals provides n independent and identically distributed random matrices $Y_1, \dots, Y_n$, which can be arranged in a three-way array. A three-way data set is thus characterized by three class of entities or *modes* (Carroll and Arabie 1980): units (rows), variables (columns) and occasions (layers).

Suppose we are interested in clustering the n observed matrices in some k groups or classes, with $k < n$, using the full information of the other two modes. Clustering a three-way data set is a complex problem, since correlations

C. Viroli (✉)
Department of Statistics, University of Bologna, Bologna, Italy
e-mail: cinzia.viroli@unibo.it

between variables could change across occasions and, vice versa, correlations across the different occasions or situations can be quite different for each response. This problem has been variously addressed in the statistical literature. A very simple solution consists of applying some dimension reduction techniques, such as principal component analysis, to one of the modes, so as to convert the three-way data set to a two-way data set, and thereby to apply conventional clustering techniques. However, the first principal component, being the direction which explains the major part of total variance, could not necessarily preserve all the clustering structure of the data (see, for a deeper discussion, Chang 1983; Jones and Sibson 1987).

Some different solutions for clustering three-way data are based on a least-square approach (Gordon and Vichi 1998; Vichi 1999); they have been more recently extended in order to combine clustering and data reduction (Vichi et al. 2007). These methodologies are based on least square approaches which do not require explicit distributional assumption on the clusters. In a model-based perspective one way of developing a solution for the three-way clustering problem is to adapt the mixture likelihood approach to dealing with three-way data (see Basford and McLachlan 1985). In this approach it is assumed that each unit j belongs to one of k possible groups in proportions $\pi_1, \ldots, \pi_k$ respectively, so that in a given occasion l, with $l = 1, \ldots, r$,

$$Y_{jl} \sim \phi_i^{(p)}(\mu_{il}, \Sigma_i) \quad (i = 1, \ldots, k),$$

with probability π_i. In the previous expression Y_{jl} is a vector of length p and ϕ^p is the p-variate normal. The mean vectors, μ_{il} vary between groups and occasions, while the within component covariance matrices, Σ_i, are taken not to depend on the occasion. The mixture model takes the form:

$$f(Y_j) = \sum_{i=1}^{k} \pi_i \prod_{l=1}^{r} \phi_i^{(p)}(\mu_{il}, \Sigma_i) \quad (i = 1, \ldots, k). \tag{1}$$

This approach has been further extended by Hunt and Basford (1999) for dealing with mixed observed variables and by Vermunt (2007) for allowing units to belong to different classes in different situations by using a hierarchical approach similar to the one proposed for the multilevel latent class model (Vermunt 2003). A first drawback of this mixture model based approach for three-way data is that it does not explicitly estimate the correlations between occasions (they are implicitly taken to be zero). Moreover, as previously observed, correlations between variables are assumed to be constant across the third mode.

In some sense, all the likelihood based methods so far proposed (and also those based on a least-squares approach mentioned above), perform clustering after collapsing the three-way structure into a two-way matrix in different ways: either by dealing with sub-matrices separately for occasions, or by matricizing Y along one of the modes.

In this work we aim at taking into account the full information on the two modes, variables and situations, simultaneously. This can be achieved by modeling the distribution of observed matrices instead of units. More precisely, for continuous variables, we model each mixture component according to a matrix-variate normal distribution (Nel 1977; Dutilleul 1999). This approach enters a very general framework which includes, as special cases, both mixtures of multivariate normals and the variant proposed by Basford and McLachlan (1985) for the analysis of three-way data.

The paper is organized as follows. The next section provides a short description of the matrix normal distribution. In Sect. 3, mixtures of matrix normals are defined and some properties are presented. In Sect. 4, maximum likelihood estimation of model parameters by means of the EM algorithm is developed. Section 5 presents a simulation study in order to evaluate the goodness of fit of the proposed mixture model. If no restriction is imposed on the mixture parameters, the proposed mixture model represents a very flexible solution. However, the number of parameters to be estimated rapidly increases as the number of components increases, due to the estimation of the covariance matrices of variables and situations. In Sect. 6 this issue is discussed and some model extensions are presented in order to deal with more parsimonious models. Connection with other mixture based proposals is also presented. Moreover the application of matrix normal mixtures as a discriminant analysis tool is discussed. Section 7 is devoted to some real applications.

2 Preliminaries: matrix normal distribution

Matrix-variate distributions play an important role in the theory of multivariate analysis as a tool for modeling random matrices in different contexts (see, among the others, Dawid 1981; De Wall 1988; Rowe 2003). Among these, the matrix-variate normal distribution plays the same pivotal role as the multivariate normal distribution in the family of multivariate distributions. Reasons are its mathematical tractability which still holds in the matrix-variate context, its several properties and its role as reference model for most multivariate phenomenons which is guaranteed by the central limit theorem.

Suppose we observe n independent and identically distributed random matrices $Y_1, \ldots, Y_n$ of dimension $r \times p$, where r represents the number of occasions and p the number of attributes. Let M be a $r \times p$ matrix of means; Φ a $r \times r$ covariance matrix containing the variances and covariances between the r occasions or times; and Ω is a $p \times p$ covariance matrix containing the variance and covariances of the p variables or locations. The matrices Φ and Ω are commonly referred to as the *between* and the *within* covariance matrices, respectively. The $r \times p$ matrix normal distribution is defined as

$$f(Y|M,\Phi,\Omega) = (2\pi)^{-\frac{rp}{2}}|\Phi|^{-\frac{p}{2}}|\Omega|^{-\frac{r}{2}} \times \exp\left\{-\frac{1}{2}\operatorname{tr}\Phi^{-1}(Y-M)\Omega^{-1}(Y-M)^\top\right\}, \tag{2}$$

or in compact notation

$$Y \sim \phi^{(r\times p)}(M,\Phi,\Omega). \tag{3}$$

An equivalent definition specifies the $r \times p$ matrix normal distribution as a special case of the rp-dimensional normal distribution when its covariance matrix, Σ, is separable in the form $\Sigma = \Phi \otimes \Omega$ (where $\otimes$ is the Kronecker product). Denote by y the rp-dimensional random normal variable with vector mean μ and covariance matrix Σ of dimension $rp \times rp$, then

$$\begin{aligned} f(y|\mu,\Sigma) &= (2\pi)^{-\frac{rp}{2}}|\Sigma|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu)\right\} \\ &= (2\pi)^{-\frac{rp}{2}}|\Phi|^{-\frac{p}{2}}|\Omega|^{-\frac{r}{2}} \times \exp\left\{-\frac{1}{2}(y-\mu)^\top(\Phi\otimes\Omega)^{-1}(y-\mu)\right\} \end{aligned} \tag{4}$$

which becomes equivalent to the density in (2) by using the identities $y = \operatorname{vec}(Y)$ and $\mu = \operatorname{vec}(M)$. Therefore expression (3) holds if and only if $\operatorname{vec}(Y) \sim \phi^{(rp)}(\operatorname{vec}(M), \Phi \otimes \Omega)$. The mean and the variance of the matrix normal distribution are:

$$E(\operatorname{vec}(Y)|M,\Phi,\Omega) = \operatorname{vec}(M), \tag{5}$$

$$\operatorname{Var}(\operatorname{vec}(Y)|M,\Phi,\Omega) = \Phi\otimes\Omega. \tag{6}$$

Being a special case of the rp-dimensional normal distribution, the matrix normal distribution shares the same various properties, like, for instance, closure under marginalization, conditioning and linear transformations. Gupta and Nagar (2000) provides an exhaustive description of all these properties.

3 Finite mixtures of matrix normals

3.1 Definition

Suppose we have unobserved heterogeneity in the data, so that we can assume observed matrices belong to different sub-populations k of sizes $\pi_1,\dots,\pi_k$ (with $\sum_i \pi_i = 1$). For three-way continuous data, we can assume the density of the $r \times p$ matrix of observations, Y_j, is a matrix normal distribution of parameters M_i, Φ_i and Ω_i, with $i = 1,\dots,k$ and $j = 1,\dots,n$. In this perspective, the problem is to attempt a classification of a random sample of n observed matrices $Y_1, Y_2, \dots, Y_n$ into the sub-populations from which they come. The density of the generic observed matrix is defined as

$$f(Y_j|\pi_1,\dots,\pi_k,\mathbf{\Theta}_1,\dots,\mathbf{\Theta}_k) = \sum_{i=1}^k \pi_i\phi_i^{(r\times p)}(Y_j; M_i,\Phi_i,\Omega_i), \tag{7}$$

where $\mathbf{\Theta}_i = \{M_i,\Phi_i,\Omega_i\}$ collectively denotes the set of matrix normal parameters. The weights π_i with $i = 1,\dots,k$ represent the prior probabilities of belonging to each sub-population corresponding to a mixture component. The posterior probability $\tau_{ij}(Y_j|\pi_1,\dots,\pi_k,\mathbf{\Theta}_1,\dots,\mathbf{\Theta}_k)$ that the observed matrix Y_j belongs to the ith component of the mixture can be expressed by Bayes's theorem as

$$\begin{aligned} \tau_{ij}(Y_j|\pi_1,\dots,\pi_k,\mathbf{\Theta}_1,\dots,\mathbf{\Theta}_k) &= \frac{\pi_i\phi_i^{(r\times p)}(Y_j; M_i,\Phi_i,\Omega_i)}{\sum_{h=1}^k \pi_h\phi_i^{(r\times p)}(Y_j; M_h,\Phi_h,\Omega_h)} \\ &= \frac{\pi_i\phi_{ij}}{\sum_{h=1}^k \pi_h\phi_{hj}}. \end{aligned} \tag{8}$$

3.2 Properties

A random matrix Y, distributed according to a mixture of k matrix normal components, as in (7), has the following useful properties.

Proposition 1 *The mean and variance of* vec(Y) *are*:

$$E(\operatorname{vec}(Y)) = \sum_{i=1}^k \pi_i \operatorname{vec}(M_i), \tag{9}$$

$$\begin{aligned} \operatorname{Var}(\operatorname{vec}(Y)) &= \sum_{i=1}^k \pi_i \operatorname{vec}(M_i)\operatorname{vec}(M_i)^\top + \sum_{i=1}^k \pi_i(\Phi_i\otimes\Omega_i) \\ &\quad - \left(\sum_{i=1}^k \pi_i \operatorname{vec}(M_i)\right)\left(\sum_{i=1}^k \pi_i \operatorname{vec}(M_i)\right)^\top. \end{aligned} \tag{10}$$

Proof is given in Appendix A.

Theorem 1 *Let A be a $m \times r$ matrix whose columns are a set of orthonormal vectors, with $m \le r$. The linear transformation $X = AY$ of dimension $m \times p$ is distributed according* *to a mixture of k matrix normals having the form*:

$$f(X|A,\pi_1,\dots,\pi_k,\boldsymbol{\Theta}_1,\dots,\boldsymbol{\Theta}_k)$$
$$=\sum_{i=1}^{k}\pi_i\phi_i^{(m\times p)}(X;AM_i,(A\Phi_iA^\top),\Omega_i). \qquad (11)$$

Proof is derived in Appendix A. This theorem can be easily extended to a double linear transformation related to both modes of the random matrix Y.

Corollary 1 *Let A be a $m\times r$ matrix whose columns are a set of orthonormal vectors, with $m\le r$ and let B be a $p\times d$ matrix with orthonormal columns, $d\le p$. The double linear transformation $X=AYB$ of dimension $m\times d$ is distributed according to a mixture of k matrix normals having the form*:

$$f(X|A,B,\pi_1,\dots,\pi_k,\boldsymbol{\Theta}_1,\dots,\boldsymbol{\Theta}_k)$$
$$=\sum_{i=1}^{k}\pi_i\phi_i^{(m\times d)}(X;AM_iB,A\Phi_iA^\top,B^\top\Omega_iB). \qquad (12)$$

Theorem 1 and its corollary are important since linear transformations could be applied for dimension reduction or visualization purposes. As a special case, when A or (and) B are vectors the previous properties show the connection between mixtures of matrix normals and mixtures of multivariate (univariate) normals.

4 Maximum likelihood estimation

Given a set of independent random matrices $Y_1,\dots,Y_n$, the log-likelihood function can be written

$$\ell(\boldsymbol{\pi},\boldsymbol{\Theta}|Y_1,\dots,Y_n)$$
$$=\sum_{j=1}^{n}\log\left\{\sum_{i=1}^{k}\pi_i\phi_i^{(r\times p)}(Y_j;M_i,\Phi_i,\Omega_i)\right\}, \qquad (13)$$

where $\boldsymbol{\pi}=\{\pi_1,\dots,\pi_k\}$ and $\boldsymbol{\Theta}=\{\boldsymbol{\Theta}_1,\dots,\boldsymbol{\Theta}_k\}$.

Parameters in (13) can be efficiently estimated through the EM algorithm (Dempster et al. 1977) which alternates between the expectation and the maximization steps until convergence, with the aim of maximizing the conditional expectation of the so-called complete density given the observable data. Let z be the allocation variable of the mixture model defined in (7). More precisely, z is a vector of dimension k denoting the component membership of each matrix sample. Evidently, z follows a multinomial distribution

$$f(z|\boldsymbol{\pi},\boldsymbol{\Theta})=\prod_{i=1}^{k}\pi_i^{z_i}, \qquad (14)$$

from which $f(z_i=1|\boldsymbol{\pi},\boldsymbol{\Theta})=\pi_i$.

The conditional density of the random matrix, Y, given the allocation variable is the matrix normal distribution in the form:

$$f(Y|z_i=1;\boldsymbol{\pi},\boldsymbol{\Theta})=\phi_i^{(r\times p)}(Y;M_i,\Phi_i,\Omega_i). \qquad (15)$$

Given the allocation variable the complete density, defined as $f(Y,z|\boldsymbol{\pi},\boldsymbol{\Theta})$, can be decomposed into the product of the two densities in (14) and (15). Then the parameter function to be maximized is the conditional expectation of the complete density given the observable data, using a fixed set of parameters $\boldsymbol{\pi}'$ and $\boldsymbol{\Theta}'$:

$$\arg\max_{\boldsymbol{\pi},\boldsymbol{\Theta}}E_{z|Y;\boldsymbol{\pi}',\boldsymbol{\Theta}'}\left[\log f\,(Y,z|\boldsymbol{\pi},\boldsymbol{\Theta})\right]$$
$$=\arg\max_{\boldsymbol{\pi},\boldsymbol{\Theta}}E_{z|Y;\boldsymbol{\pi}',\boldsymbol{\Theta}'}\left[\log f\,(Y|z;\boldsymbol{\pi},\boldsymbol{\Theta})\right.$$
$$\left.+\log f\,(z|\boldsymbol{\pi},\boldsymbol{\Theta})\right], \qquad (16)$$

which is equivalent to maximizing the following function with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$:

$$L\left(\boldsymbol{\pi},\boldsymbol{\Theta}|Y_1,\dots,Y_n,\tau_{ij}\right)$$
$$=\sum_{i=1}^{k}\sum_{j=1}^{n}\tau_{ij}\log\left[\pi_i\phi_i^{(r\times p)}(Y_j;M_i,\Phi_i,\Omega_i)\right]. \qquad (17)$$

4.1 EM algorithm

In order to maximize expression (17), the conditional distribution of the allocation variable given the observed data $Pr(z_{ij}=1|Y)$, $(i=1,\dots,k;\ j=1,\dots,n)$ must be computed at each iteration of the E-step, as function of the current set of parameters.

In the M-step of the algorithm the two terms in (16) can be separately maximized since they depend on a different set of parameters and all cross-derivatives are zero.

The optimal values for the mixture parameters, M_i, Φ_i and Ω_i are obtained by maximizing $E_{z|Y;\boldsymbol{\pi}',\boldsymbol{\Theta}'}[\log f(Y|z;\boldsymbol{\pi},\boldsymbol{\Theta})]$ where $f(Y|z;\boldsymbol{\pi},\boldsymbol{\Theta})$ has the matrix normal density in (15).

We derive

$$E_{z|Y;\boldsymbol{\pi}',\boldsymbol{\Theta}'}\left[\sum_{j=1}^{n}\log f\left(Y_j|z_j;\boldsymbol{\pi},\boldsymbol{\Theta}\right)\right]$$
$$=\sum_{i=1}^{k}f(z_{ij}|Y_j;\boldsymbol{\pi}',\boldsymbol{\Theta}')\left[-\frac{rpn}{2}\log(2\pi)\right.$$
$$-\frac{pn}{2}\log|\Phi_i|-\frac{rn}{2}\log|\Omega_i|$$
$$\left.-\frac{1}{2}\sum_{j=1}^{n}\operatorname{tr}\Phi_i^{-1}(Y_j-M_i)\Omega_i^{-1}(Y_j-M_i)^\top\right].$$

By applying the first derivatives with respect to the mixture parameters, estimates can be obtained in closed form:

$$\hat{M}_i = \frac{\sum_{j=1}^{n} \tau_{ij} Y_j}{\sum_{j=1}^{n} \tau_{ij}}, \tag{18}$$

$$\hat{\Phi}_i = \frac{\sum_{j=1}^{n} \tau_{ij} (Y_j - \hat{M}_i) \Omega_i^{-1} (Y_j - \hat{M}_i)^\top}{p \sum_{j=1}^{n} \tau_{ij}}, \tag{19}$$

$$\hat{\Omega}_i = \frac{\sum_{j=1}^{n} \tau_{ij} (Y_j - \hat{M}_i)^\top \hat{\Phi}_i^{-1} (Y_j - \hat{M}_i)}{r \sum_{j=1}^{n} \tau_{ij}}. \tag{20}$$

Estimates for the weights of the mixture can be obtained by evaluating the score function of $E_{z|Y;\boldsymbol{\pi}',\boldsymbol{\Theta}'}[\log f(z|\boldsymbol{\pi}, \boldsymbol{\Theta})]$ under the constraints that they must be positive and sum to one. This estimates are similar to the ones in case of Gaussian mixtures (see McLachlan and Peel 2000, for major details):

$$\hat{\pi}_i = \frac{\sum_{j=1}^{n} \tau_{ij}}{n}.$$

The algorithm has been implemented in R code and is available from the author's homepage.

5 A simulation study

The performance of the proposed mixture model is first evaluated in two Monte Carlo experiments. In the first simulation study mixtures of matrix normals with different number of components have been fitted on simulated data with three classes, with the aim of measuring the capability of some information criteria to detect the correct specification of k. The aim of the second experiment is to assess the classification performance of the proposed mixture model with increasing dimensionality, p, different sample sizes, n, and different starting points of the EM algorithm.

5.1 Simulation design 1

A sample of 300 observations has been drawn from three matrix normals of proportions $\pi_1 = 0.3$, $\pi_2 = 0.4$ and $\pi_3 = 0.3$ and dimensionality $r = 3$ and $p = 5$. The three matrix means have been set to $\text{vec}(M_1) = \{0.5, 0.5, 0, \dots, 0\}$, $\text{vec}(M_2) = \{0, 0, \dots, 0\}$ and finally $\text{vec}(M_3) = \{-0.5, 0.5, 0, \dots, 0\}$. The three between and within covariance matrices have been randomly generated through the methodology proposed in Joe (2006). A reasonable level of noise, generated according to a centered Gaussian with variance equal to 0.2, has been finally added to the data. Table 1 shows the best values of the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Integrated Classification Likelihood Criterion (ICL-BIC, Biernacki et al. 2000) obtained in a sequence of 100 multistart estimation procedures.

Table 1 Frequencies with which each model is selected according to the information criteria BIC, AIC and ICL-BIC

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
BIC	0	7	93	0	0
AIC	0	0	39	28	33
ICL-BIC	0	13	87	0	0

Results show that in this simulation study all the criteria suggest the correct number of components, $k = 3$, but AIC being characterized by a smaller penalty term seems to overestimate the number of components in some situations. Among the three criteria, the BIC suggests the correct model most of times. Although finite mixture models do not satisfy the regularity conditions upon which the BIC is defined, several results suggest its appropriateness and good performance in the model-based clustering context (see, for a complete discussion, Fraley and Raftery 2002a).

With reference to the best model selected by the BIC, we have computed the Mahalanobis distance between the estimated clusters. A Mahalanobis measure for three-way data can be easily obtained by applying the properties of the matrix normal distribution. More specifically, let C_1 and C_2 be two clusters with centroids given by M_1 and M_2; let Ψ_0 and Ω_0 be the weighted means of their within and between covariances. By using the matrix normal representation in (4), the Mahlabobis distance between the two vectorized centroids is:

$$D_M(C_1, C_2) = (\text{vec}(M_1) - \text{vec}(M_2))^\top [\Psi_0 \otimes \Omega_0]^{-1} \times (\text{vec}(M_1) - \text{vec}(M_2))$$

which is equivalent to

$$D_M(C_1, C_2) = \text{tr}\left[\Psi_0^{-1} (M_1 - M_2) \Omega_0^{-1} (M_1 - M_2)^\top\right]. \tag{21}$$

The three Mahalanobis distances between the estimated clusters are $D_M(C_1, C_2) = 0.004$, $D_M(C_2, C_3) = 0.002$ and $D_M(C_1, C_3) = 0.011$ respectively, which indicate that clusters 1 and 3 are better separated. This is also confirmed by considering the classification matrix of the estimated classification reported in Table 2.

5.2 Simulation design 2

Given the previously described setting for the model parameters, 100 data sets with $p = 3, 5, 7, 9$, $r = 2, 4, 6, 8$ and with different sample sizes $n = 100, 300, 500$ have been independently drawn. Table 3 reports the means of the error

Table 2 Confusion matrix

Model clustering	True clustering		
	1	2	3
1	6	123	0
2	74	8	0
3	0	2	87

Table 3 Means of the error rates obtained across the 100 replicates of each model setting. In brackets standard errors are reported

	$n = 100$	$n = 200$	$n = 500$
$p = 3, r = 2$	0.270 (0.124)	0.223 (0.114)	0.215 (0.097)
$p = 5, r = 4$	0.100 (0.084)	0.055 (0.039)	0.058 (0.064)
$p = 7, r = 6$	0.029 (0.064)	0.015 (0.039)	0.017 (0.047)
$p = 9, r = 8$	0.013 (0.056)	0.013 (0.058)	0.007 (0.041)

rates obtained across the 100 replicates of each model setting with different randomly chosen starting values for the EM algorithm. Results indicate that classification performance improves as the sample size increases, since in all the model settings when n increases the error rate decreases. Moreover, classification performance seems to be quite robust to the model dimensionality, even in the highest dimensional setting ($p = 9$ and $r = 8$). This is due to the high flexibility of matrix-normal mixtures when both the covariance matrices are assumed to be unconstrained. However, the number of parameters to be estimated could rapidly increase as the number of component increases, as better described in the next section.

6 Limitations and extensions

6.1 More parsimonious mixture models

If no restriction is imposed on the mixture parameters, the proposed mixture model is very flexible since classes can differ with respect to locations and according to the variability of the two modes. However, the number of parameters to be estimated rapidly increases as the number of components increases, especially due to the fully unconstrained within and between covariance matrices. Moreover, in certain situations, the within or between variability across the different components could be homogenous or isotropic. In order to explore various clustering situations, some restrictions on the model parameters can be introduced with the aim of obtaining more parsimonious models which are still appropriate and sufficiently flexible for clustering purposes.

In multivariate normal mixtures, restrictions typically consist of constraining the class-specific covariance matrices by a parametrization of the generic component-covariance matrix based on its spectral decomposition (Banfield and Raftery 1993; Celeux and Govaert 1995).

This parametrization consists of expressing the component covariance matrix of a mixture model, say Σ_i, in terms of its eigenvalue decomposition as follows

$$\Sigma_i = \lambda_i D_i A_i D_i^\top \tag{22}$$

where $D_i^\top$ is the matrix of eigenvectors, A_i is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_i and λ_i is the associated constant of proportionality. By allowing some but not all of these quantities to vary between clusters, a family of different models can be estimated. With reference to the proposed mixture model, the most interesting situations are those in which the between covariances Φ_i and the within covariances Ω_i are: homoscedastic, diagonal but heteroscedastic, diagonal and homoscedastic, spherical allowing for varying volumes and isotropic. By combining these restrictions, a family of $36 = 6 \times 6$ possible sub-models can be defined. Table 4 illustrates the number of parameters for the fully unconstrained model and for the different parameterizations. By taking into account the notation given in Fraley and Raftery (2002a), the label VVV refers to heteroscedastic components, EEE denotes homoscedastic components, VVI denotes diagonal but varying variability components, EEI refers to diagonal and homoscedastic components and finally VII and EII denote spherical components with and without varying volume.

The estimation details of the covariance matrices for the eight situations are illustrated in Appendix B.

Remark 1 By taking $\Phi_i = \mathbf{I}_r$ for each component i, with $i = 1, \ldots, k$, mixtures of matrix normals coincide with the mixtures of multivariate normals proposed by Basford and McLachlan (1985) for modelling three-way data. In fact, in this particular case, $\mathbf{I}_r \oplus \Sigma_i$ is a block diagonal matrix which contains Σ_i on the diagonal. Then by (5) and (6) we have that each component density is

$$\phi_i^{(rp)}(\mathrm{vec}(M_i), \mathbf{I}_r \oplus \Sigma_i) = \prod_{i=1}^{r} \phi_i^{(p)}(M_{il}, \Sigma_i)$$

and expression (1) is easily obtained with $M_{il} = \mu_{il}$.

Remark 2 By taking $\Omega_i = \mathbf{I}_r$ for each component i, with $i = 1, \ldots, k$, we obtain a variant of the Basford and McLachaln model for the situation in which all the observed variables are sphered. On the contrary, correlations between different occasions are not null and can vary across classes. The double constraint $\Phi_i = \mathbf{I}_r$ and $\Omega_i = \mathbf{I}_r$ leads to the most simplified mixture model.

Remark 3 Another parsimonious variant can be obtained by imposing some constraints on the mean matrix compo-

Table 4 Number of parameters of the 36 mixture models, with $\alpha = k - 1 + krp$

$\Phi_i \backslash \Omega_i$	VVV	EEE	VVI	EEI	VII	EII
VVV	$\alpha + kr(r+1)/2 + kp(p+1)/2$	$\alpha + kr(r+1)/2 + p(p+1)/2$	$\alpha + kr(r+1)/2 + kp$	$\alpha + kr(r+1)/2 + p$	$\alpha + kr(r+1)/2 + k$	$\alpha + kr(r+1)/2 + 1$
EEE	$\alpha + r(r+1)/2 + kp(p+1)/2$	$\alpha + r(r+1)/2 + p(p+1)/2$	$\alpha + r(r+1)/2 + kp$	$\alpha + r(r+1)/2 + p$	$\alpha + r(r+1)/2 + k$	$\alpha + r(r+1)/2 + 1$
VVI	$\alpha + kr + kp(p+1)/2$	$\alpha + kr + p(p+1)/2$	$\alpha + kr + kp$	$\alpha + kr + p$	$\alpha + kr + k$	$\alpha + kr + 1$
EEI	$\alpha + r + kp(p+1)/2$	$\alpha + r + p(p+1)/2$	$\alpha + r + kp$	$\alpha + r + p$	$\alpha + r + k$	$\alpha + r + 1$
VII	$\alpha + k + kp(p+1)/2$	$\alpha + k + p(p+1)/2$	$\alpha + k + kp$	$\alpha + k + p$	$\alpha + k + k$	$\alpha + k + 1$
EII	$\alpha + 1 + kp(p+1)/2$	$\alpha + 1 + p(p+1)/2$	$\alpha + 1 + kp$	$\alpha + 1 + p$	$\alpha + 1 + k$	$\alpha + 1 + 1$

nents M_i. For instance, if the occasions are repeated measures of the same variables, one could assume that the class specific matrix means do not vary across the third mode, which means that all the rows of M_i are equal (since M_i is a matrix of dimension $r \times p$). Alternatively the rows could be proportional or modeled by a linear relation with occasions (see, for an example, Vermunt 2007).

Remark 4 Let Y be distributed according to (7), with fully unconstrained or constrained component parameters. Thanks to Theorem 1, mixtures of matrix normals include mixtures of multivariate or univariate normals as special case. In particular, if A is a $1 \times r$ vector the linear transformation $X = AY$ has a distribution given by a mixture of kp-variate normals. If B is a $p \times 1$ vector the linear transformation $X = YB$ leads to a mixture of kr-variate normals.

Remark 5 The constraints here introduced are not the only possible ones for defining more parsimonious models. Recently, Bouveyron et al. (2007) proposed a further extension of this parametrization by allowing the spectral decomposition of the covariance matrix is divided in two blocks, one relevant and one not relevant for clustering purposes. By fixing some parameters of the two subspaces to be common within or between classes they obtain a family of different regularized mixture models. Rather than by a restricted eigenvalue decomposition, the structure of the covariance matrices can also be simplified by lower-dimensional representations using factor analytic structures (Yung 1997; McLachlan et al. 2003; Montanari and Viroli 2010). A factor analysis model for matrix-variate variables has been proposed by Xie et al. (2008).

6.2 Discriminant analysis

In the different perspective of supervised classification, mixture models have been widely applied as a flexible tool for density estimation (McLachlan and Peel 2000; Fraley and Raftery 2002a, 2002b). In a typical supervised classification problem, the number of groups, k, is assumed to be known and the main purpose is to define a rule for predicting the class membership of some units on the basis of the complete knowledge of a training sample. The training sample on which the rule is built consists of the p predictors and of an indicator of the class membership for a set of n units. In probabilistic discriminant analysis, the so-called Bayes decision rule suggests to assign a unit $\mathbf{y}_0$ to the population $\hat{i}$ such that

$$\hat{i} = \underset{i=1,\dots,k}{\arg\max} \left\{ f_i(\mathbf{y}_0)\pi_i \right\}, \tag{23}$$

where f_i, with $i, \dots, k$, denotes the class conditional density and π_i the *a priori* probability of observing an individual from population i. If the class conditional densities are Gaussian, expression (23) simply yields the well known linear or quadratic discriminant functions according to whether the condition of homoscedasticity is fulfilled or not. A way of generalizing linear or quadratic discriminant analysis is to allow the density f_i to be a mixture of normals (Scott 1992; McLachlan 1992; Hastie and Tibshirani 1996). Mixture based discriminant analysis can be extended for the analysis of three way data, by allowing each class conditional density to be modelled by a mixture of matrix normals. In the most flexible and general situation, the number of components and the parameterization of the between and within covariance matrices, above introduced, that best describe each class are data driven choices.

7 Real examples

Finite mixtures of matrix normals, with unconstrained and constrained parameters, have been fitted on three real examples. The aim of the first real application is to compare the results of the proposed mixture model with those obtained by the previous proposal of Basford and McLachlan (1985). In the second real example we want to evaluate the clustering performance on real data in which the true cluster membership is known and compare it with the conventional model based clustering on the matricized data set along the third mode. The last application is a problem of discriminant analysis, previously analyzed with different methodologies. Here we apply mixtures of matrix normals for modelling each class conditional density.

7.1 Soybean data

This data originated by an experiment of Mungomery et al. (1974). Fifty-eight soybean lines were evaluated at four locations in southeastern Queensland in two years, 1970 and 1971. For each of the eight location-year combinations ($r = 8$ environments), several chemical and agronomic attributes were observed. Basford and McLachlan (1985) have previously analyzed this data set considering $p = 2$ attributes: seed yield (kg/ha) and seed protein percentage. They found $k = 7$ groups with a BIC of 2635. On this data we applied mixtures of matrix normals with k ranging from 1 to 8. The best model has been chosen according to the BIC in the family of the 16 sub-models illustrated in Table 4. Since the starting values of the EM-algorithm could affect the estimated final likelihood, a random multistart strategy between 50 initial points has been considered. The selected mixture model has a BIC value equal to 2443 and consists of $k = 3$ components with homoscedastic between covariances, Φ_i (EEE), and diagonal, but varying, within covariances Σ_i, (VVI). In other terms, the relation between the environmental conditions do not change among the three groups. Moreover, the two variables are uncorrelated but have a different variability in the three estimated classes. It is interesting to note that the three groups, obtained by computing the posterior probabilities in (8), are the fusion of the partition into the 7 groups obtained by Basford and Mclachaln (1985), as shown in Table 5.

Table 5 Soybean data: number of cases in each of the k groups detected by mixtures of matrix normal (rows) and by the mixture based approach by Basford and Mclachaln, 1985 (columns)

k	1	2	3	4	5	6	7
1	0	0	0	9	0	0	0
2	0	0	0	0	9	9	16
3	3	3	9	0	0	0	0

In order to give an interpretation to the obtained classes, the estimated component means are reported in Table 6. Group 2 consists of lines with the lowest values of yield but high protein, group 1 and 3 mainly differentiate each other with respect to the first variable. On the whole 8 environments, group 3 seems to be the one with the highest seed yield.

7.2 Landsat satellite data

This dataset is obtained from the UCI machine-learning repository homepage and consists of satellite images purchased from NASA by the Australian Centre for Remote Sensing. The data set consists of $r = 4$ digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each digital image is represented in a 3×3 square neighbourhood for a total of $p = 9$ pixels. Satellite images are of different types. For our purposes, we have selected only the images belonging to the classes grey soil, damp grey soil and soil with vegetation stubble, because of their similarity. The cases contained in the test data set available from the UCI website, are 397, 211 and 237 for the three classes respectively, for a total of $n = 845$ observations. Due to the high dimensionality of the data, classification is usually based on the four central spectral values for a total of $p = 4$ attributes. In so doing, information coming from the other pixels is discarded and the three-way structure of the data is collapsed into a two-way structure of dimension 845×4. Classification on this reduced dataset by Gaussian mixture models with heteroscedastic components implemented with the `R` library `Mclust` leads to a classification error rate of 0.258. If we consider all the pixels (without separating the variability within and between the $r = 4$ spectral bands) data can be structured in a matrix of dimensions 845×36. In this second situation the error classification rate obtained with `Mclust` is 0.283.

On the three-way data of dimensions $845 \times 4 \times 9$ we have fitted a mixture of $k = 3$ matrix normals. The EM algorithm has been implemented with a multistart strategy for the initial values of parameters and the optimal model has been selected according to the BIC criterion. The best model is the fully unconstrained mixture for both the between and within covariances, (VVV, VVV), with a value of BIC equal to 23416. The obtained classification error rate is 0.116.

7.3 Handwritten digit recognition

This example contains samples of handwritten digits (0, 1, ..., 9) collected from the handwritten ZIP codes on envelopes from US postal mail (Hastie et al. 2001). Each image is a matrix of 16×16 pixels ranging in intensity from

Table 6 Soybean data: estimated component means

	Seed yield			Seed protein percentage		
$l\backslash k$	1	2	3	1	2	3
1	2.59	2.08	2.38	37.50	40.20	39.30
2	1.09	1.45	2.08	38.10	41.40	38.50
3	2.26	1.92	1.60	34.80	37.20	37.00
4	1.66	0.99	3.09	38.90	42.30	38.00
5	2.64	2.36	2.77	37.70	41.40	38.30
6	2.40	2.16	3.18	38.80	43.50	39.60
7	2.67	2.23	2.25	39.00	41.80	40.80
8	2.34	1.39	2.37	39.90	45.00	39.30
	2.21	*1.82*	*2.46*	*38.1*	*41.6*	*38.8*

Table 7 Digit classification results. The first 4 rows are taken by Hastie and Tibshirani (1996). The fifth row reports the error rates obtained with Mixture of Matrix Normals (MNN)

Technique	Error rates (%)	
	Training	Test
LDA	1.6	8.7
MDA (filtered −64 df)	2.7	7.1
MDA (filtered −64 df, 4 subclasses)	2.3	6.3
MDA (filetered −49 df, 4 subclasses, shrunk)	2.7	5.8
MMN (6 subclasses)	4.0	7.1

0 to 255. From a training set of 1756 cases, we want to predict the identity of images 3, 5 and 8 of 492 digits in a validation set. This problem has been previously addressed by Hastie and Tibshirani (1996) with classification results which are reported in Table 7. We approximated each class conditional density by a Mixture of Matrix Normals (MMN) specified in order to minimize the classification error rate in the training set. According to this criterion, we have chosen fully unconstrained mixture models with 6 components for each class conditional density. Results are in line, but not better, with the previous error rates obtained with mixture discriminant analysis, as reported in the last row of the table.

8 Concluding remarks

In this work finite mixtures of matrix normal distributions are defined and investigated. As observed, the matrix normal distribution represents a natural way for modeling random matrices, describing experiments characterized by the simultaneous observation of continuous variables in different situations or locations. It is not surprising that interest towards this distribution (and towards the family of matrix-variate distributions) has began many years ago, but a wider investigation of many potential applications of this class of distributions has not been fully explored. A possible motivation is that the choice of modeling the distribution of matrices instead of the distribution of units can be computationally high demanding. However a direct analysis of three-way data structures allows to consider and estimate correlations within and between variables and occasions. For this reason they represent a powerful tool for model based clustering of observed matrices. In so doing, a necessary condition is that the number of observed variables in the different occasions does not vary. Possible extensions in this direction are not here analyzed, but they could be cast in a multilevel approach. In the different perspective of discriminant analysis, mixture of matrix normals can be applied as a flexible tool for density estimation. A family of different 36 more parsimonious sub-models, originated by differently parameterizing the between and within covariance matrices, has been considered. An EM algorithm for the estimation of the proposed family of mixture models has been developed. Possible limitations of the proposed algorithm are that it can converge to degenerate solutions; it is quite sensitive to the choice of the starting points and it can be quite slow when n is large. A Bayesian approach could alleviate such problems by simultaneously solving the issue of the model selection.

Appendix A

In this section proofs of theorems and properties described in Sect. 3 are given.

Proof of Proposition 1 We assume that classical regularity conditions hold; then the expected value of Y is given by

$$E(Y)=\int\sum_{i=1}^{k}\pi_i Y\phi_i^{(n\times p)}(Y;M_i,\Phi_i,\Omega_i)dY$$
$$=\sum_{i=1}^{k}\pi_i\int Y\phi_i^{(n\times p)}(Y;M_i,\Phi_i,\Omega_i)dY=\sum_{i=1}^{k}\pi_i M_i,$$

from which expression (9) is straightforward. In order to derive the variance of $y=\operatorname{vec}(Y)$ we need to compute the second moment as follows:

$$E(yy^\top)=\int\sum_{i=1}^{k}\pi_i yy^\top\phi_i^{(n\times p)}(Y;M_i,\Phi_i,\Omega_i)dydy^\top$$
$$=\sum_{i=1}^{k}\pi_i\int yy^\top\phi_i^{(n\times p)}(Y;M_i,\Phi_i,\Omega_i)dydy^\top$$
$$=\sum_{i=1}^{k}\pi_i(\Phi_i\otimes\Omega_i)+\sum_{i=1}^{k}\pi_i\left(\operatorname{vec}(M_i)\operatorname{vec}(M_i)^\top\right).$$

Expression (10) derives by the relation $\operatorname{var}(y)=E(yy^\top)-E(y)E(y)^\top$. □

Proof of Theorem 1 The property can be demonstrated as follows. Let A^* be the orthogonal matrix consisting of $\{A,A^\perp\}$. It splits the coordinates of the original space as follows

$$X^*=A^*Y=(A,A^\perp)Y=(X,X^\perp).$$

The marginal distribution of X is defined as

$$f(X)=\int f(X,X^\perp)dX^\perp=\int\frac{1}{J}f(A^{*\top}X^*)dX^\perp,$$

where $J=\det(\partial X^*/\partial Y)=\det(A^*)$ is the Jacobian related to coordinate transformation. Since $\det(A^*)=1$,

$$f(X)=\int f(A^{*\top}X^*)dX^\perp$$
$$=\int\sum_i\pi_i\phi_i^{(r\times p)}(A^{*\top}X^*;M_i,\Phi_i,\Omega_i)dX^\perp$$
$$=\sum_i\pi_i\int\phi_i^{(r\times p)}(X^*;A^*M_i,A^*\Phi_iA^{*\top},\Omega_i)dX^\perp$$
$$=\sum_i\pi_i\phi_i^{(m\times p)}(X;AM_i,A\Phi_iA^\top,\Omega_i),$$

where the property of closure under linear transformations of the matrix normal has been used (see, Gupta and Nagar 2000, for major details). □

Proof of Corollary 1 Being an extension of Theorem 1, proof is straightforward. □

Appendix B

Celeux and Govaert (1995) provided maximum likelihood estimation for a big family of 14 constrained Gaussian mixture models. In this section we develop the EM estimators under the different constraints for the within and between covariance matrices expressed as $\Phi_i=\lambda_iD_iA_iD_i^\top$ and $\Omega_i=\zeta_iL_iC_iL_i^\top$.

Let B be the between cluster scattering matrix:

$$B=\sum_{i=1}^{k}\sum_{j=1}^{n}\tau_{ij}(Y_i-M_i)\Omega_i^{-1}(Y_j-M_i)^\top=\sum_{i=1}^{k}B_i,$$

where $B_i=\sum_{j=1}^{n}\tau_{ij}(Y_i-M_i)\Omega_i^{-1}(Y_j-M_i)^\top$. The within cluster scattering matrix is defined as

$$W=\sum_{i=1}^{k}\sum_{j=1}^{n}\tau_{ij}(Y_i-M_i)^\top\Phi_i^{-1}(Y_j-M_i)=\sum_{i=1}^{k}W_i,$$

with $W_i=\sum_{j=1}^{n}\tau_{ij}(Y_i-M_i)^\top\Phi_i^{-1}(Y_j-M_i)$.

Model EEE The maximization of (17) with respect to Φ_i under the constraint $\Phi_i=\Phi=\lambda DAD^\top$ becomes

$$\max_{\Phi}\sum_{i=1}^{k}\sum_{j=1}^{n}\tau_{ij}\log\left[\pi_i\phi_i^{(r\times p)}(Y_j;M_i,\Phi,\Omega_i)\right]$$
$$=\min_{\Phi}np\log|\Phi|$$
$$+\operatorname{tr}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\tau_{ij}\Phi^{-1}(Y_i-M_i)\Omega_i^{-1}(Y_j-M_i)^\top\right]$$
$$=\min_{\Phi}np\log|\Phi|+\operatorname{tr}[\Phi^{-1}B]$$

where B is the between cluster scattering matrix. The homoscedastic between covariance matrix is estimated by

$$\hat{\Phi}=\frac{B}{np}.$$

Analogously the maximization of (17) with respect to Ω_i under the constraint $\Omega_i=\Omega=\zeta LCL^\top$ leads to

$$\hat{\Omega}=\frac{W}{nr},$$

where W is the within cluster scattering matrix.

Model VVI In the situation $\Phi_i = \lambda_i A_i$, maximizing equation (17) leads to the minimization of

$$\begin{aligned}
&\min_{\Phi_i} p \sum_{i=1}^{k} \sum_{j=1}^{n} \tau_{ij} \log|\Phi_i| \\
&\quad + \operatorname{tr}\left[\sum_{i=1}^{k} \sum_{j=1}^{n} \tau_{ij} \Phi_i^{-1} (Y_i - M_i) \Omega_i^{-1} (Y_j - M_i)^\top\right] \\
&= \min_{\Phi_i} p \sum_{i=1}^{k} n_i \log|\Phi_i| + \operatorname{tr}\left[\sum_{i=1}^{k} \Phi_i^{-1} B_i\right] \\
&= \min_{\Phi_i} p \sum_{i=1}^{k} n_i \log \lambda_i^r + \sum_{i=1}^{k} \frac{1}{\lambda_i} \operatorname{tr}[A_i^{-1} B_i]
\end{aligned}$$

with $n_i = \sum_{j=1}^{n} \tau_{ij}$. By computing the first derivatives with respect to each λ_i and A_i the estimates are:

$$\hat{A}_i = \frac{\operatorname{diag}[B_i]}{|\operatorname{diag}[B_i]|^{1/r}}, \qquad \hat{\lambda}_i = \frac{|\operatorname{diag}[B_i]|^{1/r}}{p n_i}.$$

Analogously, for $\Omega_i = \zeta_i C_i$, we have

$$\hat{C}_i = \frac{\operatorname{diag}[W_i]}{|\operatorname{diag}[W_i]|^{1/p}}, \qquad \hat{\zeta}_i = \frac{|\operatorname{diag}[W_i]|^{1/p}}{r n_i}.$$

Model EEI In the situation $\Phi_i = \lambda A$ the maximization of (17) leads to the minimization of

$$\min_{\Phi} np \log|\Phi| + \operatorname{tr}\left[\Phi^{-1} B\right] = npr \log \lambda + \frac{1}{\lambda} \operatorname{tr}\left[A^{-1} B\right]$$

from which

$$\hat{A} = \frac{\operatorname{diag}[B]}{|\operatorname{diag}[B]|^{1/r}}, \qquad \hat{\lambda} = \frac{|\operatorname{diag}[B]|^{1/r}}{np}.$$

Analogously, for $\Omega = \zeta C$, estimates are

$$\hat{C} = \frac{\operatorname{diag}[W]}{|\operatorname{diag}[W]|^{1/p}}, \qquad \hat{\zeta} = \frac{|\operatorname{diag}[W]|^{1/p}}{nr}.$$

Model VII If $\Phi_i = \lambda_i \mathbf{I}_r$, we have the minimization problem

$$\min_{\Phi_i} p \sum_{i=1}^{k} n_i \log \lambda_i^r + \sum_{i=1}^{k} \frac{1}{\lambda_i} \operatorname{tr}[B_i],$$

from which

$$\hat{\lambda}_i = \frac{\operatorname{tr}[B_i]}{p r n_i}.$$

Similarly, for $\Omega_i = \zeta_i \mathbf{I}_p$, we have

$$\hat{\zeta}_i = \frac{\operatorname{tr}[W_i]}{p r n_i}.$$

Model EII In the isotropic situation $\Phi_i = \lambda \mathbf{I}_r$ the maximization of equation (17) leads to the minimization of

$$\min_{\Phi} npr \log \lambda + \frac{1}{\lambda} \operatorname{tr}[B],$$

from which

$$\hat{\lambda} = \frac{\operatorname{tr}[B]}{prn}.$$

Analogously, for $\Omega = \zeta \mathbf{I}_p$, we get

$$\hat{\zeta} = \frac{\operatorname{tr}[W]}{prn}.$$

References

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

Basford, K.E., McLachlan, G.J.: The mixture method of clustering applied to three-way data. J. Classif. **2**, 109–125 (1985)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. PAMI **22**, 719–725 (2000)

Billard, L., Diday, E.: From the statistics of data to the statistics of knoweledge: symbolic data analysis. J. Am. Stat. Assoc. **98**, 470–487 (2003)

Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Comput. Stat. Data Anal. **52**, 502–519 (2007)

Carroll, J.D., Arabie, P.: Multidimensional scaling. Ann. Rev. Psychol. **31**, 607–649 (1980)

Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**, 781–793 (1995)

Chang, W.C.: On using principal components before separating a mixture of two multivariate normal distributions. Appl. Stat. **32**, 267–275 (1983)

Dawid, A.P.: Some matrix-variate distribution theory: notational considerations and a Bayesian application. Biometrika **68**, 265–274 (1981)

De Wall, D.J.: Matrix-variate distributions. In: Knotz, S., Johnson, N.L. (eds.): Encyclopedia of Statistical Sciences, vol. 5, pp. 326–333. Wiley, New York (1988)

Dempster, N.M., Laird, A.P., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)

Dutilleul, P.: The MLE algorithm for the matrix normal distribution. J. Stat. Comput. Simul. **64**, 105–123 (1999)

Fraley, C., Raftery, A.E.: MCLUST: Software for model-based cluster analysis. J. Classif. **16**, 297–206 (1999)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. **97**, 611–631 (2002a)

Fraley, C., Raftery, A.E.: MCLUST: Software for model-based clustering, discriminant analysis and density estimation, Technical Report No. 415, Department of Statistics, University of Washington (2002b)

Fraley, C., Raftery, A.E.: Enhanced Software for model-based clustering, discriminant analysis and density estimation: MCLUST. J. Classif. **20**, 263–286 (2003)

Ganesalingam, S., McLachlan, G.J.: A case study for two clustering methods based on maximum likelihood. Stat. Neerl. **33**, 81–90 (1979)

Gordon, A.D., Vichi, M.: Partitions of partitions. J. Classif. **15**, 265–285 (1998)

Gupta, A.K., Nagar, D.K.: Matrix Variate Distributions. Chapman and Hall/CRC, London/Boca Raton (2000)

Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. J. R. Stat. Soc. B **58**, 155–176 (1996)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)

Hunt, L.A., Basford, K.E.: Fitting a Mixture Model to three-mode three-way data with categorical and continuous variables. J. Classif. **16**, 283–296 (1999)

Joe, H.: Generating random correlation matrices based on partial correlations. J. Multivar. Anal. **97**, 2177–2189 (2006)

Jones, M.C., Sibson, R.: What is projection pursuit? (with discussion). J. R. Stat. Soc. A **150**, 1–38 (1987)

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

McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York (1992)

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

McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

McLachlan, G.J., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**, 379–388 (2003)

Montanari, A., Viroli, C.: Heteroscedastic factor mixture analysis. Stat. Modell. Int. J. (2010, forthcoming)

Mungomery, V.E., Shorter, R., Byth, D.E.: Genotype x environment interactions and environmental adaption. I. Pattern analysis—application to soya bean populations. Austr. J. Agric. Res. **25**, 59–72 (1974)

Nel, H.M.: On distributions and moments associated with matrix normal distributions. Mathematical Statistics Department Technical Report, 24, University of the Orange Free State, Bloemfontein, South Africa (1977)

Rowe, B.R.: Multivariate Bayesian Statistics. Chapman and Hall/CRC, London/Boca Raton (2003)

Scott, D.W.: Multivariate Density Estimation. Wiley, New York (1992)

Vermunt, J.K.: Multilevel latent class models. Sociol. Method. **33**, 213–239 (2003)

Vermunt, J.K.: A hierarchical mixture model for clustering three-way data sets. Comput. Stat. Data Anal. **51**, 5368–5376 (2007)

Vichi, M.: One mode classification of a three-way data set. J. Classif. **16**, 27–44 (1999)

Vichi, M., Rocci, R., Kiers, A.L.: Simultaneous component and clustering models for three-way data: within and between approaches. J. Classif. **24**, 71–98 (2007)

Wolfe, J.H.: Pattern clustering by multivariate mixture analysis. Multivar. Behav. Res. **5**, 329–350 (1970)

Xie, X., Yan, S., Kwok, J.T., Huang, T.S.: Matrix-variate factor analysis and its applications. IEEE Trans. Neural Netw. **19**, 1821–1826 (2008)

Yung, Y.F.: Fitting mixtures in confirmatory factor-analysis models. Psychometrika **62**, 297–330 (1997)