

# Classifying three-way data through Matrix-Normal Mixtures

Cinzia Viroli

**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 and show that they can be a powerful tool for classifying three-way data in unsupervised problems.

**Key words:** Model based clustering, random matrix, three-way data, EM-algorithm

## 1 Introduction

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 when data are collected from experiments in which individuals provide multiple ratings for multiple objects (Vermunt, 2007). 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 & 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 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 in 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 one, 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).

Some different solutions for clustering three-way data are based on a least-square approach (Gordon & 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 deal with three-way data (see Basford & McLachlan, 1985). In this approach it is assumed that each unit  $j$  belongs to one of  $k$  possible groups in proportions  $\pi_1, \dots, \pi_k$  respectively, so that in a given occasion  $l$ , with  $l = 1, \dots, r$ ,

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

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

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

This approach has been further extended by Hunt & 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-square approach), 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 represents a very general framework which includes, as special cases, both mixtures of multivariate normals and the variant proposed by Basford & McLachlan (1985) for the analysis of three-way data. 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. To deal with this problem some model extensions will be presented in order to obtain more parsimonious models.

## 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, A. P., 1981; de Wall, 1988).

Suppose we observe  $n$  independent and identically distributed random matrices  $Y_1, \dots, Y_n$  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;  $\Phi$  a  $r \times r$  covariance matrix containing the variances and covariances between the  $r$  occasions or times; and  $\Omega$  is a  $p \times p$  covariance matrix containing the variance and covariances of the  $p$  variables or locations. The matrices  $\Phi$  and  $\Omega$  are commonly referred to as the *between* and the *within* covariance matrices, respectively. The  $r \times p$  matrix normal distribution is defined as

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

or in compact notation

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

## 3 Finite Mixtures of Matrix Normals

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  $M_i$ ,  $\Phi_i$  and  $\Omega_i$ , 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

$$f(Y_j | \pi_1, \dots, \pi_k, \Theta_1, \dots, \Theta_k) = \sum_{i=1}^k \pi_i \phi_i^{(r \times p)}(Y_j; M_i, \Phi_i, \Omega_i), \quad (4)$$

where  $\Theta_i = \{M_i, \Phi_i, \Omega_i\}$  collectively denotes the set of matrix normal parameters. The weights  $\pi_i$  with  $i = 1, \dots, k$  represent the prior probabilities of belonging to each sub-population corresponding to a mixture component. The posterior probability  $\tau_{ij}(Y_j | \pi_1, \dots, \pi_k, \Theta_1, \dots, \Theta_k)$  that the observed matrix  $Y_j$  belongs to the  $i$ th component of the mixture can be expressed by Bayes's theorem as

$$\tau_{ij}(Y_j | \pi_1, \dots, \pi_k, \Theta_1, \dots, \Theta_k) = \frac{\pi_i \phi_i^{(r \times p)}(Y_j; M_i, \Phi_i, \Omega_i)}{\sum_{h=1}^k \pi_h \phi_h^{(r \times p)}(Y_j; M_h, \Phi_h, \Omega_h)} = \frac{\pi_i \phi_{ij}}{\sum_{h=1}^k \pi_h \phi_{hj}}. \quad (5)$$

Parameters in (4) can be efficiently estimated through the EM algorithm which alternates between the expectation and the maximization steps until convergence. It is possible to show that estimates for the model parameters can be obtained in closed form:

$$\hat{M}_i = \frac{\sum_{j=1}^n \tau_{ij} Y_j}{\sum_{j=1}^n \tau_{ij}}, \quad (6)$$

$$\hat{\Phi}_i = \frac{\sum_{j=1}^n \tau_{ij} (Y_j - \hat{M}_i) \Omega_i^{-1} (Y_j - \hat{M}_i)^\top}{p \sum_{j=1}^n \tau_{ij}}, \quad (7)$$

$$\hat{\Omega}_i = \frac{\sum_{j=1}^n \tau_{ij} (Y_j - \hat{M}_i)^\top \hat{\Phi}_i^{-1} (Y_j - \hat{M}_i)}{r \sum_{j=1}^n \tau_{ij}}, \quad (8)$$

$$\hat{\pi}_i = \frac{\sum_{j=1}^n \tau_{ij}}{n}. \quad (9)$$

## 4 A simulation study

The performance of the proposed mixture model is first evaluated in two Monte Carlo experiments, with the aim of measuring the capability of some information criteria to detect the correct specification of  $k$  and 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.

### 4.1 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., 1999) 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 in 100 replicates.

|         | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ |
|---------|---------|---------|---------|---------|---------|
| 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 the times. Although finite mixture models do not satisfy the regularity conditions upon which the BIC is defined, several results in the literature suggest its appropriateness and good performance in the model-based clustering context (see, for a complete discussion, Fraley & Raftery, 2002).

### 4.2 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 2 reports the means of the error 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.

**Table 2** 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) |

## 5 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 (Celeux & Govaert, 1995). This parametrization consists in expressing the component covariance matrix of a mixture model, say  $\Sigma_i$ , in terms of its eigenvalue decomposition,  $\Sigma_i = \lambda_i D_i A_i D_i^\top$ , where  $D_i^\top$  is the matrix of eigenvectors,  $A_i$  is a diagonal matrix whose elements are proportional to the eigenvalues of  $\Sigma_i$  and  $\lambda_i$  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  $\Phi_i$  and the within covariances  $\Omega_i$  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 3 illustrates the number of parameters for the fully unconstrained model and for the different parameterizations. By taking into account the notation given in Fraley & Raftery (2002), the label VVV refers to heteroscedastic components, EEE denotes homoscedastic components, VVI denotes diagonal but varying

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

**Table 3** Number of parameters of the 36 mixture models, with  $\alpha = k - 1 + krp$ .

variability components, EEI refers to diagonal and homoscedastic components and finally VII and EII denote spherical components with and without varying volume.

**Remark 1.** By taking  $\Phi_i = \mathbf{I}_r$  for each component  $i$ , with  $i = 1, \dots, k$ , mixtures of matrix normals coincide with the mixtures of multivariate normals proposed by Basford & 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  $\Sigma_i$  on the diagonal. Then we have that each component density is

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

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  $i = 1, \dots, k$ , we obtain a variant of the Basford & McLachlan 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 components  $M_i$ . 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).

## References

1. Basford, K.E. & Mclachlan, G.J., 1985, The Mixture Method of Clustering applied to three-way data. *Journal of Classification*, 2, 109-125.
2. Biernacki, C., Celeux, G. & Govaert, G., 1999, An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model. *Pattern Recognition Letters*, 20, 267-272.
3. Carroll, J.D. & Arabie, P., 1980, Multidimensional Scaling. *Annual Review of Psychology*, 31, 607-649.
4. Celeux, G. & Govaert, G., 1995, Gaussian parsimonious clustering models. *Pattern Recognition*, 28, 781-793.
5. Chang, W.C., 1983, On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-275.
6. Dawid, A.P., 1981, Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68, 265-274.
7. De Wall, D.J., 1988, Matrix-variate distributions. In: *Encyclopedia of Statistical Sciences*, 5, Knotz, S. and Johnson, N.L. (Eds., Wiley, New York, 326-333.
8. Dutilleul, P., 1999, The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64, 105-123.
9. Fraley, C. & Raftery, A.E., 2002, Model-based clustering, discriminant analysis & density estimation. *Journal of the American Statistical Association*, 97, 611-631.
10. Gordon, A.D. & Vichi, M., 1998, Partitions of Partitions. *Journal of Classification*, 15, 265-285.
11. Gupta, A.K. & Nagar, D.K., 2000, *Matrix variate distributions*. Chapman and Hall, CRC, U.S.
12. Hunt, L.A. & Basford, K.E., 1999, Fitting a Mixture Model to three-mode three-way data with categorical and continuous variables. *Journal of Classification*, 16, 283-296.
13. Joe, H., 2006, Generating Random Correlation Matrices Based on Partial Correlations. *Journal of Multivariate Analysis*, 97, 2172-2189.
14. Mclachlan, G.J. and Basford, K.E., 1988, *Mixture models: inference and application to clustering*. New York, Marcel Dekker.
15. Mclachlan, G.J. and Peel, D., 2000, *Finite Mixture Models*. New York: John Wiley & Sons INC.
16. Nel, H.M., 1977, On distributions and moments associated with matrix normal distributions. *Mathematical Statistics Department Technical Report*, 24, University of the Orange Free State, Bloemfontein, South Africa.
17. Vermunt, J.K., 2003, Multilevel latent class models. *Sociological Methodology*, 33, 213-239.
18. Vermunt, J.K., 2007, A hierarchical mixture model for clustering three-way data sets. *Computational Statistics & Data Analysis*, 51, 5368-5376.
19. Vichi, M., 1999, One mode classification of a three-way data set. *Journal of Classification*, 16, 27-44.
20. Vichi, M., Rocci, R. & Kiers, A.L., 2007, Simultaneous Component and Clustering Models for three-way data: within and between approaches. *Journal of Classification*, 24, 71-98.