Clustering Longitudinal Mixed Data

Francesco Amato [1], Julien Jacques [1]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*francesco.amato, julien.jacques*}*@univ-lyon2.fr*

Résumé Nous présentons un algorithme de clustering pour données longitudinales mixtes. En supposant que les variables non continues sont la discrétisation de variables continues latentes, le modèle s'appuie sur un mélange de lois normales matricielles, capable de prendre en compte simultanément des structures de dépendance entre variables et temporelles. Le modèle est ainsi capable de modéliser simultanément l'hétérogénéité des données, l'association entre les réponses et la structure de dépendance temporelle. Un algorithme EM est développé pour l'estimation des paramètres.
Mots-clés Clustering probabiliste. Données longitudinales mixtes. Données à trois voies. Modèles de mélange. Lois Gaussiennes matricielles.

Abstract. We present a model-based clustering algorithm to cluster longitudinal mixed data. Assuming that the non-continuous variables are the discretization of underlying latent continuous variables, the model relies on a mixture of matrix-variate normal distributions, accounting simultaneously for within- and between-time dependence structures. The model is thus able to concurrently model the heterogeneity, the association among the responses and the temporal dependence structure. An EM algorithm is developed for parameters estimation.
Keywords. Model-based Clustering. Mixed longitudinal data. Three-way data. Mixture models. Matrix-variate Gaussians.

1 Context

In many areas of humanities and social sciences, the studies are based on questionnaires completed by participants. Often, these questionnaires are completed several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population.
However, the statistical analysis of these questionnaires is far from simple, for several reasons. First, the answers to the questions are often of different types. The analysis of such mixed data is a current research problem in the fields of statistics and machine learning. The second scientific obstacle is the modelling of the temporal evolution of the answers to the questions. Currently, too frequently the analyses are done independently at each temporal phase, then researchers try *a posteriori* to find links between these different analyses, by seeking from one phase to the other to find similar typical behaviour. We can

for example cite Selosse et al., 2019 in the case of clustering of longitudinal ordinal data for an application in psychology. The ideal way to model these data would be through modelling all the responses to the questionnaires at the same time.
In this work we aim at providing a tool to perform model-based clustering on questionnaires repeated over time. Probabilistic (or model-based) clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering (Bouveyron et al., 2019).

2 Related work

While several approaches exist for the clustering longitudinal and mixed data separately, literature is poor when they are to be dealt with simultaneously.
An approach to clustering longitudinal data consists in arranging the data in a three-way format and modelling them through a matrix-variate mixture model. This approach offers the advantage of accounting for the overall time-behavior, grouping together the units that have a similar pattern across and within time. While not being new (Basford and McLachlan, 1985), matrix-variate distributions have recently gained attention, and mixtures of matrix-normals (MMN) have been developed and applied both in a frequentist framework in Viroli, 2011a and within a Bayesian one by Viroli, 2011b. These models represent a natural extension of the multivariate normal mixtures to account for temporal (or even spatial) dependencies, and have the advantage of being also relatively easy to estimate by means of EM algorithm (a nice short description of the EM application to MNN is provided in §2.1 of Wang and Melnykov, 2020). More recently, in Gallaugher and McNicholas, 2018 and Melnykov and Zhu, 2018, 2019 extensions for non-normal skewed cases have been proposed and applied. However, matrix-variate models suffer from over-parametrization that leads to estimation issues. To overcome this issue a more parsimonious model (Sarkar et al., 2020) and a new R package (Zhu, Sarkar, and Melnykov, 2022) has been proposed. Despite their efficacy, up to now these methods have only been applied to continuous data.
Our model expands the use of matrix-variate mixtures to mixed data, by building on the framework proposed by McParland and Gormley, 2016 and further developed by Choi, Ahn, and Kim, 2023.

3 Preliminaries

Let $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, that is a matrix-variate normal distribution where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix

containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is given by

$$f(Z|M,\Phi,\Sigma) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^{\intercal}]\right\}. \quad (1)$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Z \sim \mathcal{MN}_{(J\times T)}(M,\Phi,\Sigma)$, then $\text{vec}(Z) \sim \mathcal{MVN}_{JT}(\text{vec}(M),\Phi\otimes\Sigma)$, where vec(.) is the vectorization operator, that is the function mapping from a a $J\times T$ matrix to a JT-dimensional vector, and $\otimes$ denotes the Kronecker product. The property of rewriting the general covariance matrix $\Psi\in\mathbb{R}^{JT\times TJ}$ as $\Psi=\Phi\otimes\Sigma$ is called separability condition. Then, the mean and the variance of the multivariate normal normal distribution are:

$$\mathbb{E}(\text{vec}(Z)|M,\Phi,\Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M,\Phi,\Sigma) = \Sigma\otimes\Phi. \quad (2)$$

Being a special case of the multivariate normal distribution, the matrix-normal distribution shares the same properties, like, for instance, closure under marginalization, conditioning and linear transformations (Gupta and Nagar, 2000). The separability condition of the covariance matrix has two advantages. First, it allows the modeling of the temporal pattern of interest directly on the covariance matrix Φ. Second, it represents a more parsimonious solution than that of the unrestricted $\Phi\otimes\Sigma$.

Introduced by Viroli, 2011a, the pdf of the finite Mixture of Matrix-Normals (MMN) model is given by

$$f(Z|\boldsymbol{\pi},\boldsymbol{\Theta}) = \sum_{k=1}^{K}\pi_k\phi^{(J\times T)}(Z|M_k,\Phi_k,\Sigma_k),$$

where $\phi^{(J\times T)}$ represents the density function of a $J\times T$-dimensional matrix-variate normal, K is the number of mixture components, $\boldsymbol{\pi}=\{\pi_k\}_{k=1}^K$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^K\pi_k=1$ and $\boldsymbol{\Theta}=\{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k=\{M_k,\Phi_k,\Sigma_k\}$.

4 Model

Denote by y_{ijt} the observation of the j-th ($j=1,..,J$) variable for the i-th ($i=1,\ldots,N$) unit at time t ($t=1,\ldots,T$), that is: imagine to observe N units and measuring J different mixed variables T times throughout the course of the study. We can divide the J mixed variables into C continuous variables and O the non-continuous ones, such that $C+O=J$. Let us reorganize this data in a random-matrix form such that $\mathbf{Y}=\{Y_i\}_{i=1}^N$ is a sample of $J\times T$-variate matrix observations $Y_i=(y_{ijt})\in[\mathbb{R}^{C\times T},\mathbb{N}^{O\times T}]^{\intercal}$, $J=C+O$. The ordered classes are coded by positive integers such that each ordinal variable o the

ordinal levels are $\{1, 2, \ldots, C_o\}$, while the binary classes are coded as 0 and 1.
Assuming our population is heterogeneous and partitioned into K clusters, we define $\ell_i = (\ell_{i1}, \ldots, \ell_{iK})$ as a one-hot encoding representation of group membership, such that $\ell_{ik} = 1$ if the i-th unit belongs to the k-th cluster.
Then, we can assume that each variable y_{ijt} is the manifestation of an underlying latent continuous variable z_{ijt}.

4.1 Modelling continuous variables

We assume that the observed continuous variables y_{ijt} match exactly the latent variable:

$$y_{ijt} = z_{ijt}$$

4.2 Modelling ordinal variables

To map ordinal data, we follow Amato, Jacques, and Prim-Allaz, 2024. Let the generic ordinal o-th variable have C_o levels. Let γ_o denote a $C_o + 1$ -dimensional vector of thresholds that partition the real line for the corresponding o-th underlying continuous variable, and let the threshold parameters be constrained such that $-\infty = \gamma_{o,0} \leq \gamma_{o,1} \leq \ldots \leq \gamma_{o,C_o} = \infty$. If the latent $z_{i,o,t}$ is such that $\gamma_{o,c-1} < z_{i,o,t} < \gamma_{o,c}$ then the observed ordinal response, $y_{i,o,t} = c$.
Moreover, let define $\mathcal{O}^{O \times T}$ the set of ordinal matrices of size $J \times T$ whose row o takes values in $\{1, \ldots, C_o\}$. Each element of $\mathcal{O}^{O \times T}$ is called a response pattern. Let R be the cardinality of $\mathcal{O}^{O \times T}$. Each response pattern $Y_r \in \mathcal{O}^{O \times T}$ is generated by a portion Ω_r of the latent space $\mathbb{R}^{O \times T}$ according to thresholds $\boldsymbol{\gamma} := \{\gamma_o\}_{o=1}^{O}$. Let the binary vector $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iR})$ be one-hot encoding of Y_i such that if the r-th pattern is observed then $\tilde{Y}_{ir} = 1$ and any other entry in the vector equals zero.

A key point is of course the choice of the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{o=1}^{O}$. To avoid identifiability and computational complexity issues, thresholds are fixed and not considered as parameters. There are different ways to do it. We decide to follow Corneli, Bouveyron, and Latouche, 2020, where the thresholds are chosen as $\gamma_o = (-\infty, 1.5, 2.5, \ldots, C_o - 0.5, \infty)$.

4.3 Modelling categorical variables

For non-ordered categorical data with P levels we can consider a one-hot encoding for $P - 1$ levels and treat them as binary variables. Binary variables can be considered as a special case of ordinal variables where the number of classes $C_o = 2$. The threshold cutting the underlying continuous variable is set to 0.

4.4 Joint model

At this point, we can assume that each observed matrix Y_i is indeed the manifestation of a latent random matrix Z_i, and that this underlying random matrix is linked through different relations to the observed matrix Y_i, depending on the type of variable each element $y_{i,j,t}$, as described in Section 4.

So, we can think of Y_i as a block matrix, and conveniently split it between the first C rows, representing the observed continuous variables, and the remaining $J - C = O$ rows, representing the observed ordinal and categorical variables. Notice that the slicing happens just over rows but not over columns. Then, for notation's sake we can write $Y_i = [Y_i^\alpha, Y_i^\beta]^\intercal$, where $Y_i^\alpha \in \mathbb{R}^{C\times T}$ is the block containing the continuous variables and $Y_i^\beta \in \mathbb{N}^{O\times T}$ gathers the ordinal and categorical ones (that we coded via integers).

$$\begin{pmatrix}\mathbb{R}^{C\times T}\\ \mathbb{N}^{O\times T}\end{pmatrix} \ni Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T}\\ \vdots & \ddots & \vdots & \cdots & \vdots\\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T}\\ \vdots & \cdots & \vdots & \ddots & \vdots\\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T}\end{pmatrix} \longleftarrow Z_i = \begin{pmatrix} z_{i,1,1} & \cdots & z_{i,1,t} & \cdots & z_{i,1,T}\\ \vdots & \ddots & \vdots & \cdots & \vdots\\ z_{i,j,1} & \cdots & z_{i,j,t} & \cdots & z_{i,j,T}\\ \vdots & \cdots & \vdots & \ddots & \vdots\\ z_{i,J,1} & \cdots & z_{i,J,t} & \cdots & z_{i,J,T}\end{pmatrix} \in \mathbb{R}^{J\times T}$$

Again, we can write $Z = [Z_i^\alpha, Z_i^\beta]^\intercal$, applying the same logic as for Y_i. Then, we assume a mixture of matrix-normal distributions on the latent space Z_i.

Assuming that:

$$\begin{aligned}\ell_i \sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} := (\pi_1, \dots, \pi_K)\\ Z_i|\ell_{ik} = 1 \sim \mathcal{MN}_{(J\times T)}(Z_i|\Theta_k),\ \Theta_k := \{M_k, \Phi_k, \Sigma_k\},\end{aligned}$$

we get:

$$f(\ell_i) = \prod_{k=1}^{K} \pi_k^{\ell_{ik}}\ ;\ f(Z_i|\ell_i) = \prod_{k=1}^{K} \left[\phi^{(J\times T)}(Z_i|\Theta_k)\right]^{\ell_{ik}}\ ;$$

where $\mathcal{M}$ indicates the multinomial distribution.

In the following, $\boldsymbol{Z} := \{Z_i\}_{i=1}^N, \boldsymbol{\ell} := \{\ell_i\}_{i=1}^N$ will indicate the ensembles of Z_i, ℓ_i. Finally, let $\mathbf{Y} := \{Y_i\}_{i=1}^N$ be the collection of the observed matrices Y_i.

5 Estimation

To estimate the model, since we do not observe neither Z nor ℓ, we resort to the EM algorithm (Dempster, Laird, and Rubin, 1977).

The EM algorithm is an iterative algorithm alternates two steps: the expectation step (E-step) and the maximization step (M-step). It start from an initialization $\hat{\boldsymbol{\Theta}}^{(0)}$ of the parameters. Then, let denote with the superscript $(s+1)$ the parameters estimated in the current step and with (s) the ones computed in the previous step.
The E-step consists of evaluating $\mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)}) := \mathbb{E}(\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, that is the expectation of the complete log-likelihood conditioned on the parameters computed in the previous step and on the observed data.
In the M-step the parameters are updated by maximizing the expected complete log-likelihood found on the E step, that is $\hat{\boldsymbol{\Theta}}^{(s+1)} := \arg\max_{\Theta} \mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)})$. The iteration process is repeated until convergence on the log-likelihood is met.

5.1 Complete log-likelihood

The complete log-likelihood can be written, up to some constant c, as:

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \ell_{ik} \Bigg[\log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \Bigg] + c. \quad (3)$$

The unknown parameters to be estimated are $\boldsymbol{\Theta} := \{\pi_k, M_k, \Phi_k, \Sigma_k\}_{k=1}^K$

5.2 E-step

Looking at 3, keeping in time the block-structure of Z_i and the links we defined in 4, it is easy to see that the expected values to be computed are $\mathbb{E}(\ell_{ik} | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, $\mathbb{E}(\ell_{ik} Z_i^\beta | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ and of $\mathbb{E}(\ell_{ik} Z_i^\beta \Phi_k^{-1(s)} Z_i^{\beta\intercal} | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ or $\mathbb{E}(\ell_{ik} Z_i^{\beta\intercal} \Sigma_k^{-1(s)} Z_i^\beta | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ by the cyclic property of the trace. We will compute both as they are both needed in the M-step.

The first involves computing a cumulative probability of a matrix-variate normal distribution according to the thresholds described in Section 4. This in turn means solving a complex high-dimensional integral, which is hardly tractable analytically. However, it can be approximated through a Monte-Carlo approach applied on the vectorized reparametrization of the matrix-variate distribution according to Section 3.

The remaining three require the computation of the first and second moments of a truncated matrix-variate distribution. However, again that is a complex task with no close

solution, so we will need to work the issue around. We can bypass the problem by again working on the vectorized version of the distribution through the use of a Monte Carlo approach and specifically the use of a Gibbs sampler to sample from a truncated multivariate normal distribution. The samples generated to calculate the first moment can be reused to compute the second moment by calculating the inner product of the vectors used to compute the first then calculating the sample mean of these inner products.

5.3 M-step

To maximize the expected complete log-likelihood we can take the derivatives of Equation 3 with respect to the parameters. All updating equations have closed form and can be computed thanks to the expectations found in the E-step.

6 Conclusions

Mixture of matrix-variate normal distributions can be an efficient way to cluster longitudinal continuous data. Assuming that non-continuous variables are a discretization of latent continuous variables allows us to extend the use of these MMN to cluster longitudinal mixed data sets. Numerical study on synthetic data sets as well as real data application concerning diet choice during the pandemic (François-Lecompte et al., 2020) will be presented.

References

[1] Francesco Amato, Julien Jacques, and Isabelle Prim-Allaz. "Clustering longitudinal ordinal data via finite mixture of matrix-variate distributions". In: *Statistics and Computing* 34.2 (Apr. 2024). ISSN: 1573-1375. DOI: 10.1007/s11222-024-10390-z.

[2] Kaye E. Basford and Geoffrey J. McLachlan. "The mixture method of clustering applied to three-way data". In: *Journal of Classification* 2.1 (Dec. 1985), pp. 109–125. ISSN: 1432-1343. DOI: 10.1007/BF01908066.

[3] Charles Bouveyron et al. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge, England, UK: Cambridge University Press, June 2019. ISBN: 978-1-10864418-1. DOI: 10.1017/9781108644181.

[4] Young-Geun Choi, Soohyun Ahn, and Jayoun Kim. "Model-Based Clustering of Mixed Data With Sparse Dependence". In: *IEEE Access* 11 (July 2023), pp. 75945–75954. DOI: 10.1109/ACCESS.2023.3296790.

[5] Marco Corneli, Charles Bouveyron, and Pierre Latouche. "Co-Clustering of Ordinal Data via Latent Continuous Random Variables and Not Missing at Random Entries". In: *Journal of Computational and Graphical Statistics* 29.4 (Oct. 2020), pp. 771–785. ISSN: 1061-8600. DOI: 10.1080/10618600.2020.1739533.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (Sept. 1977), pp. 1–22. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[7] Agnès François-Lecompte et al. "Confinement et comportements alimentaires - Quelles évolutions en matière d'alimentation durable ?" In: *Revue Française de Gestion* 46.293 (Nov. 2020), pp. 55–80. ISSN: 0338-4551. DOI: 10.3166/rfg.2020.00493.

[8] Michael P. B. Gallaugher and Paul D. McNicholas. "Finite mixtures of skewed matrix variate distributions". In: *Pattern Recognition* 80 (Aug. 2018), pp. 83–93. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.02.025.

[9] Arjun Kumar Gupta and Daya Krishna Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.

[10] Damien McParland and Isobel Claire Gormley. "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2 (June 2016), pp. 155–169. ISSN: 1862-5355. DOI: 10.1007/s11634-016-0238-x.

[11] Volodymyr Melnykov and Xuwen Zhu. "On model-based clustering of skewed matrix data". In: *Journal of Multivariate Analysis* 167 (Sept. 2018), pp. 181–194. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2018.04.007.

[12] Volodymyr Melnykov and Xuwen Zhu. "Studying crime trends in the USA over the years 2000–2012". In: *Advances in Data Analysis and Classification* 13.1 (Mar. 2019), pp. 325–341. ISSN: 1862-5355. DOI: 10.1007/s11634-018-0326-1.

[13] Shuchismita Sarkar et al. "On parsimonious models for modeling matrix data". In: *Computational Statistics & Data Analysis* 142 (Feb. 2020), p. 106822. ISSN: 0167-9473. DOI: 10.1016/j.csda.2019.106822.

[14] Margot Selosse et al. "Analysing a quality-of-life survey by using a co-clustering model for ordinal data and some dynamic implications". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.5 (Nov. 2019), pp. 1327–1349. ISSN: 0035-9254. DOI: 10.1111/rssc.12365.

[15] Cinzia Viroli. "Finite mixtures of matrix normal distributions for classifying three-way data". In: *Statistics and Computing* 21.4 (Oct. 2011), pp. 511–522. ISSN: 1573-1375. DOI: 10.1007/s11222-010-9188-x.

[16] Cinzia Viroli. "Model based clustering for three-way data structures". In: *Bayesian Analysis* 6.4 (Dec. 2011), pp. 573–602. ISSN: 1936-0975. DOI: 10.1214/11-BA622.

[17] Yang Wang and Volodymyr Melnykov. "On variable selection in matrix mixture modelling". In: *Stat* 9.1 (Jan. 2020), e278. ISSN: 2049-1573. DOI: 10.1002/sta4.278.

[18] Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. "MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling". In: *Journal of Classification* 39.1 (Mar. 2022), pp. 147–170. ISSN: 1432-1343. DOI: 10.1007/s00357-021-09401-9.