Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Distributions

Francesco Amato, Julien Jacques, Isabelle Prim-Allaz

Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Distributions

Francesco Amato [1], Julien Jacques [1], Isabelle Prim-Allaz [2]

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

Abstract. In social sciences, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model-based clustering algorithm for such longitudinal ordinal data. Assuming that an ordinal variable is the discretization of a underlying latent continuous variable, 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 and presented for parameters estimation. An evaluation of the model through synthetic data shows its estimation abilities and its advantages when compared to competitors. A real-world application concerning changes in eating behaviours during the Covid-19 pandemic period in France will be presented.

Keywords. Model-based Clustering. Ordinal 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. The most common kind of questions, and therefore collected data, are ordinal, as for instance in marketing studies where people are asked to evaluate some products or services on an ordinal scale (Dillon et al., 1994). Ordinal data occur when the categories are ordered (Agresti, 2010). Ordinality is a characteristic of the meaning of measurements (Stevens, 1946), and distinct levels of an ordinal variable differ in degree of dissimilarity more than in quantity (Agresti, 2010).

Often, these questionnaires are completed by participants several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population, being especially interested in their time evolution. Nonetheless, modelling temporal evolution is far from trivial. The most basic approach consists in performing analyses independently at each temporal phase, and then trying *a posteriori* to find links between these different analyses, by seeking from one phase to the other to

find similar or different typical behaviours. An example is Selosse, Jacques, Biernacki, and Cousson-Glie, 2019, clustering of ordinal data for an application in psychology. The ideal way to cluster temporal data would be to account for the temporal evolution, modelling all the responses to the questionnaires at the same time. We propose a model-based clustering technique aiming at facilitate such temporal analysis, by grouping together the units behaving similarly in time.

Over the decades, research has produced a vast number of different approaches to clustering. From our prospective, 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: determine the number of clusters, detecting and treating outliers, assessing uncertainty in the clustering (Bouveyron et al., 2019).

Our model proposes to cluster all the ordinal responses at the same time, grouping together the units behaving similarly in time. Moreover, it also aims at being easily understandable and interpretable by practitioners with non-statistical background.

1.1 Related works

Although ordinal data are certainly the type most encountered in questionnaires, they are either transformed according to a Likert scale (Likert, 1932) into quantitative data (Lewis et al., 2005), or transformed into nominal data by ignoring the order (Vermunt et al., 2005). In the first case, even if there is a whole literature on the construction of Likert scales, the introduction of a notion of distance between categories necessarily brings a bias in the analysis (Liddell et al., 2018). In the second case, less often used nevertheless, one loses essential information by not taking into account the notion of order within the categories.

Ordinal data do not have metric information. One classical model to treat ordinal data as in a ordinal-scale model are the traditional *ordered-probit* models (McKelvey et al., 1975, Winship et al., 1984, Becker et al., 1992). This model describes the probability of a ordinal response as the cumulative normal probability between two thresholds on an underlying latent continuous distribution, generally chosen to be Gaussian. This model is generally regarded as one of the standards in both frequentist and Bayesian frameworks (Lynch, 2007, Kruschke, 2015).

More recently, other approaches to deal with such kind of data has been developed. In the clustering context we are interested in, the examples spans from D'Elia et al., 2005, that introduces the CUB model, later developed through the R package `CUB` (Iannario et al., 2016), to Giordan et al., 2011 and more recently Ranalli et al., 2016; Fernandez et al., 2016. In a co-clustering context, the R package `ordinalClust` (Selosse, Jacques, and Biernacki, 2021) makes use of the BOS (Binary Ordinal Search) distribution introduced by Biernacki et al., 2016 and extended for co-clustering by Jacques et al., 2018. A mixture of item

response models was developed to for ordinal response data in the Bayesian framework by McParland et al., 2013, to be later expanded in the frequentist paradigm and to handle mixed data in McParland et al., 2016. More recently, Corneli et al., 2020 proposed a new model that relies on latent continuous random variables to perform co-clustering.

Similarly, several approaches to clustering longitudinal data were developed. In McNicholas et al., 2010 the authors developed a model-based clustering framework for longitudinal continuous data by using Gaussian mixture models and applying the modified Cholesky decomposition to the group covariance matrices. Doing this, the new derived elements can be interpreted as generalized auto-regressive parameters and innovation variances. Moreover, a series of possible constraints are presented in order to give rise to more parsimonious models. In the context of generalized linear latent variable models (GLLVMs), Cagnone et al., 2018 introduced a methodological framework that includes two levels of latent variables: one continuous hidden variable for dimension reduction and clustering and a discrete random variable accounting for the dynamics modelled through a latent Markov model. In the R package `mixAK` (Komrek et al., 2014) the basis for clustering is a mixture of multivariate generalized linear mixed models. In Vvra et al., 2023 a mixture distribution is additionally assumed for random effects.

An other 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 et al., 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, where it was used to cluster Italian provinces based on a longitudinal crime-related score. From a frequentist point of view, 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 et al., 2020). Anderlucci et al., 2015 extends on the work of McNicholas et al., 2010 and incorporates the idea of the modified Cholesky decomposition in the matrix-variate regression model developed by Viroli, 2012, elaborating a family of more parsimonious models. More recently, in Doğru et al., 2016, Gallaugher et al., 2018 and Melnykov et al., 2018, 2019 extensions for non-normal skewed matrix-variate mixture model have been proposed and applied. An attempt to generalize the class of parsimonious models derived by the decomposition of the covariance matrices in a mixture of matrix-normal model has been carried out (Sarkar et al., 2020). A new comprehensive R package to apply this family to clustering continuous three-way data (Zhu et al., 2022) has been proposed, endeavoring the creation of a `mclust` (Scrucca et al., 2016) for three-way continuous data.

1.2 Our idea

As we aims at develop a model easily understandable and interpretable by practitioners with non-statistical background, we found matrix-variate distributions particularly fit, as shown in Alaimo et al., 2023. Moreover, as noticed in Anderlucci et al., 2015, the use of matrix-variate distributions allow to drop the conditional independence assumption, frequently implied in longitudinal latent variable models.

Despite the efficacy of matrix-variate distributions, up to now these methods have only been applied to continuous data. We introduce a Mixture for Ordinal Matrices (MOM) model, aiming at expanding the use to matrix-variate mixtures to ordinal data in an unsupervised learning context.

In the following Sections 2 and 3 we will detail our model and the EM algorithm to perform inference. In Section 4 the results on synthetic data are presented to assess the performance of the model. Finally, in Section 5 an application on real data concerning grocery shopping preferences by a French sample during the Covid-19 pandemic period is outlined.

2 Model

2.1 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 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 matrix-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 various properties, like, for instance, closure under marginalization, conditioning and linear transformations (Gupta et al., 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 $\Phi \otimes \Sigma$. Indeed, for that case the number of independent elements to compute would be $JT(JT+1)/2$, against $J(J+1)/2+T(T+1)/2$ for the matrix-variate one. For example, setting $J = T = 5$, one would have to estimate 325 elements in the multivariate case against 30 elements in the matrix-variate one.

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 $\Theta_k = \{M_k, \Phi_k, \Sigma_k\}$.

2.2 Our model

Let denote by $y_{i,j,t}$ the observation of the j-th variable for the i-th unit at time t ($i = 1, \ldots, N$; $j = 1, .., J$ and $t = 1, \ldots, T$), that is: imagine to observe N units and measuring J different ordinal variables T times throughout the course of the study. 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 $Y_i = (y_{i,j,t}) \in \mathbb{N}^{J\times T}$. The ordered classes are coded by non-negative integers such that each ordinal variable J the ordinal levels are $\{1, 2, \ldots, C_j\}$.
Then, we can assume that each variable $y_{i,j,t}$ is the manifestation of an underlying latent continuous variable $z_{i,j,t}$ which follows a Gaussian distribution, as done in the clustMD model (McParland et al., 2016). At this point, we can assume that each observed ordinal matrix Y_i is indeed the manifestation of a latent continuous random matrix Z_i, which follows a matrix-normal distribution.

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

To map from Y_i to Z_i, let γ_j denote a $C_j + 1$ -dimensional vector of thresholds that partition the real line for the j-th ordinal variable that has C_j levels and let the threshold parameters be constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \ldots \leq \gamma_{j,C_j} = \infty$. If the latent $z_{i,j,t}$ is such that $\gamma_{j,c-1} < z_{i,j,t} < \gamma_{j,c}$ then the observed ordinal response, $y_{i,j,t} = c$.

So, by assuming that each Z_i follows a matrix-normal distribution, we can then cluster our data by means of finite Mixture of Matrix-Normals. In addition to Z_i, we introduce a latent binary K-dimensional vector that indicate whether the unit i belongs to the k-th cluster, $\ell_i = (\ell_{i1}, \dots, \ell_{iK})$, such that $\ell_{ik} = 1$ if the i-th unit belongs to the k-th cluster.

Moreover, let define $\mathcal{O}^{J\times T}$ the set of ordinal matrices of size $J \times T$ whose row j takes values in $\{1, \dots, C_j\}$. Each element of $\mathcal{O}^{J\times T}$ is called a response pattern. Let R be the cardinality of $\mathcal{O}^{J\times T}$. Each response pattern $Y_r \in \mathcal{O}^{J\times T}$ is generated by a portion Ω_r of the latent space $\mathbb{R}^{J\times T}$ according to thresholds $\boldsymbol{\gamma} := \{\gamma_j\}_{j=1}^{J}$. Let the binary vector $\tilde{Y}_i = (\tilde{Y}_{i1}, \dots, \tilde{Y}_{iR})$ 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. We can derive the joint density of $Z_i, \tilde{Y}_i, \ell_i$ as:

$$f(\tilde{Y}_i, Z_i, \ell_i) = f(\tilde{Y}_i|Z_i, \ell_i) f(Z_i|\ell_i) f(\ell_i).$$

Assuming that:

$$\begin{aligned}
\ell_i &\sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} := (\pi_1, \dots, \pi_K) \\
Z_i|\ell_{ik} = 1 &\sim \mathcal{MN}_{(J\times T)}(Z_i|\Theta_k),\ \Theta_k := \{M_k, \Phi_k, \Sigma_k\}, \\
\tilde{Y}_i|Z_i, \ell_{ik} = 1 &\sim \mathcal{M}(1, \xi_i),\ \xi_i := (\mathbf{1}_{\Omega_1}(Z_i), \dots, \mathbf{1}_{\Omega_R}(Z_i))
\end{aligned}$$

we get:

$$f(\ell_i) = \prod_{k=1}^{K} \pi_k^{\ell_{ik}}\ ;\ f(Z_i|\ell_i) = \prod_{k=1}^{K} \left[\phi^{(J\times T)}(Z_i|\Theta_k)\right]^{\ell_{ik}}\ ;\ f(Y_i^R|Z_i, \ell_i) = \prod_{r=1}^{R} \mathbf{1}_{\Omega_r}(Z_i)^{\tilde{Y}_{ir}},$$

where $\mathcal{M}$ indicate the multinomial distribution and $\mathbf{1}_{\Omega_r}(Z_i)$ is the indicator function that equals 1 when the elements in Z_i have values that determine the r-th pattern. Hence, when $\tilde{Y}_{ir} = 1$, the vector $\boldsymbol{\xi}_i$ is a vector whose r-th element equals 1 and all the others equal 0. In the following, $\boldsymbol{Z} := \{Z_i\}_{i=1}^{N}, \boldsymbol{\ell} := \{\ell_i\}_{i=1}^{N}$ and $\boldsymbol{\Theta} := \{\Theta_k, \pi_k\}_{k=1}^{K}$ will indicate the ensembles of Z_i, ℓ_i and of the parameters, respectively. Finally, let $\mathbf{Y} := \{\tilde{Y}_i\}_{i=1}^{N}$ be the collection of the observed response pattern vectors Y_i.

3 Inference

3.1 Thresholds

A key point is of course the choice of the thresholds $\boldsymbol{\gamma}$. It is clear that either the thresholds or the parameters $\boldsymbol{\Theta}$ need to be fixed for the model to be identifiable. Imagine to observe a sample of ordinal categories $c = 1, \dots, C_j$ for variable j and to work in the same framework as Section 2. Let consider each variable separately in an univariate case for sake of simplicity. Then, assume that each observation derive from the discetization of an underlying continuous variable following a normal distribution with parameters (μ_j, σ_j^2), and

consider the $C_j - 1$ dimensional thresholds vector γ_j as parameters to estimate together with the ones of the ones of the underlying normal. Then, the parameters set would be $\theta = (\mu_j, \sigma_j^2, \gamma_j)$, the parameter space $\Theta = (\mathbb{R}, \mathbb{R}^+, \mathbb{R})$, and our model $P = \{p_\theta; \theta \in \Theta\}$, with $p_\theta(y = c) = p(\gamma_{j,c-1} \leq z \leq \gamma_{,jc}), z \sim N(\mu_j, \sigma_j^2)$. It is clear that such a model would not be indentifiable as there is no bijecton $\theta \mapsto p_\theta$. For instance, for a number of ordinal categories $C_j = 2$, $\theta_1 = (1.5, 1, 1.5)$ and $\theta_2 = (0, 1, 0)$ would yield the same distribution $(p_{\theta_1} = p_{\theta_2})$.

Clearly, being the parameters of the mixture the quantities of interest, we need to decide which strategy to follow for fixing the thresholds.
As written in Section 1.1, assuming underlying continuous variables categorized according to some thresholds is not new and there are several ways of specifying such thresholds.
In McParland et al., 2016 the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^J$ are fixed at the very beginning relying on data, by setting them as $\gamma_{j,c} = \varphi^{-1}(\delta_{j,c})$, where $\delta_{j,c}$ is the proportion of variable j which are less than or equal to level c and φ is the standard normal cumulative distribution function. With this assumption, the ordinal distribution of clusters will have the sale global shape, not necessarily uni-modal, which makes clusters interpretation harder.
On the other hand, in Corneli et al., 2020 thresholds are fixed in advance arbitrarily (keeping equidistant the classes) as $\gamma_j = (1.5, 2.5, \ldots, C_j - 0.5)$ and C_j is assumed to be equal for all variables, proposing a scale conversion pre-processing algorithm (Gilula et al., 2019) for cases when this does not hold true. The advantages of such an approach is that a underlying space is related with the range of the ordinal entries, leading to easily interpretable results. Another result of equidistant thresholds is that it produces monotonicity around the mode, creating more separated and interpretable clusters. In the following work this approach will be followed.
It is important to remark that this choice of thresholds does not impose any constraint on the distribution of the ordinal levels, but the monotonic behaviour around the mode.

3.2 EM-algorithm

The EM algorithm (Dempster et al., 1977) 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 $\mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)}) := \mathbb{E}(\log \mathcal{L}_C(\boldsymbol{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \boldsymbol{\ell}) | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}})$, 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 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.

3.3 Complete Likelihood

The complete log-likelihood can be written as

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^{N} \left\{ \sum_{r=1}^{R} \tilde{Y}_{ir} \mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K} \ell_{ik} \left[\log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \right] \right\}.$$

3.4 E-step computation

Conditioning on the parameters computed in the step (s), at the step $(s+1)$ the value of $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(s)})$ is:

$$\begin{aligned} &\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(s)}) := \mathbb{E}(\log \mathcal{L}_C(\boldsymbol{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \boldsymbol{\ell}) | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) = \\ &\mathbb{E}\left(\sum_{i=1}^{N} \left\{ \sum_{r=1}^{R} Y_{ir}^R \mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K} \ell_{ik} \left[\log(\hat{\pi}_k^{(s)}) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\hat{\Phi}_k^{(s)}|) - \right. \right. \right. \\ &\left. \left. \left. \frac{T}{2} \log(|\hat{\Sigma}_k^{(s)}|) - \frac{1}{2} tr[\hat{\Sigma}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})^\intercal] \right] \right\} \middle| \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}} \right) = \end{aligned} \tag{3}$$

$$\sum_{i=1}^{N} \sum_{r=1}^{R} Y_{ir}^R \, \mathbb{E}(\mathbf{1}_{\Omega_r}(Z_i) | \hat{\boldsymbol{\pi}}^{(s)}, \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}^R) + \tag{4}$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}(\ell_{ik} | \hat{\boldsymbol{\pi}}^{(s)}, \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}^R) \left[\log(\hat{\pi}_k^{(s)}) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\hat{\Phi}_k^{(s)}|) - \frac{T}{2} \log(|\hat{\Sigma}_k^{(s)}|) \right] - \tag{5}$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \, \mathbb{E}(\ell_{ik} \, tr[\hat{\Sigma}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})^\intercal] | \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}^R) \tag{6}$$

We can treat each of the three expectations separately, and we get for (4)

$$\mathbb{E}(\mathbf{1}_{\Omega_r}(Z_i) | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) = \mathbb{P}(Z_i \in \Omega_r | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_i).$$

Since we are conditioning on $\tilde{Y}_i$, the observed response pattern is known and therefore the probability of Z_i belonging to Ω_r is equal to 1 when $\tilde{Y}_{ir} = 1$ and 0 otherwise.

By similar reasoning, for (5), since the only unknown quantity is ℓ_{ik}:

$$\mathbb{E}(\ell_{ik}|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)})=\mathbb{P}(\ell_{ik}=1|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)})=$$

$$=\frac{\mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)})\mathbb{P}(Y_{ir}^{R}=1|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)})}{\mathbb{P}(Y_{ir}^{R}=1|\hat{\boldsymbol{\pi}}^{(s)},\hat{\boldsymbol{\Theta}}^{(s)})}=$$

$$=\frac{\pi_k^{(s)}\int_{\Omega_r}f(Z|\Theta_k^{(s)})dZ}{\sum_{k=1}^{K}\pi_k^{(s)}\int_{\Omega_r}f(Z|\Theta_k^{(s)})dZ}=:\boldsymbol{\tau}_{ik}^{(s+1)},$$

where the integral can be approximated through a Monte-Carlo approach applied on the vectorized reparametrization of the matrix-variate distribution.

On the other hand, (6) is less straightforward, and we will need some tricks to deal with it. As done in McParland et al., 2016, we can break down as

$$\begin{aligned}&\mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)},\tilde{\mathbf{Y}})\times\\&\qquad\mathbb{E}(tr[\hat{\Sigma}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})^{\intercal}]|\ell_{ik}=1,\tilde{\mathbf{Y}},\hat{\boldsymbol{\Theta}}^{(s)}).\end{aligned}\tag{7}$$

By opening the matrix product in the second term we get:

$$\hat{\Sigma}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})^{\intercal}=\tag{8}$$

$$=\hat{\Sigma}_k^{-1(s)}Z_i\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}-\hat{\Sigma}_k^{-1(s)}Z_i\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal}-\hat{\Sigma}_k^{-1(s)}\hat{M}_k^{(s)}\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}+\hat{\Sigma}_k^{-1(s)}\hat{M}_k\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal}.\tag{9}$$

It is easy to realize that its solution requires the computation of $\mathbb{E}(Z_i|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1)$ and of the expectation of a matrix quadratic forms, specifically for $\mathbb{E}(Z_i\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1)$. As we will in Section 3.5, we will also need to compute $\mathbb{E}(Z_i^{\intercal}\hat{\Sigma}_k^{-1(s+1)}Z_i|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1)$ for the M-step. The computation of the expectation of Z_i and of such quadratic form necessitates in turn to compute the moments of a truncated matrix-variate Gaussian. However, that is a complex task, so we will need to work the issue around.

We can bypass the problem concerning the expectation of Z_i by defining with $z_i\in\mathbb{R}^{JT\times 1}$ the vectorized version of Z_i and computing

$$\mathbb{E}(z_i|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)})=:m_{ik}^{(s+1)}\tag{10}$$

through the use of a Monte Carlo approach and specifically the use of a Gibbs sampler to sample from a truncated multivariate normal distribution. Moreover, the samples

generated to calculate the first moment $m_{ik}^{(s+1)}$ can be reused to compute the matrix $S_{ik}^{(s+1)} := \mathbb{E}(z_i z_i^\intercal | \ell_{ik} = 1, \tilde{Y}_{ir} = 1, \hat{\boldsymbol{\Theta}}^{(s)}) \in \mathbb{R}^{JT \times JT}$, that can be approximated by calculating the inner product of the vectors used to compute $m_{ik}^{(s+1)}$ then calculating the sample mean of these inner products.

Subsequently, we can find $\mathbb{E}(Z_i \hat{\Phi}_k^{-1(s+1)} Z_i^\intercal | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1)$ by computing it element-by-element. In order to do that, we can define $D_{ik}^{(s)} := \mathbb{E}(Z_i \hat{\Phi}_k^{-1(s+1)} Z_i^\intercal | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1))$, $\hat{\varphi}_{k,gd}^{(s)}$ as the $(g,d)^{th}$ element of $\hat{\Phi}_k^{-1(s)}$. Then, the $(h,t)^{th}$ element of $Z_i^\intercal \hat{\Phi}_k^{-1(s)} Z_i$ would be $\sum_{d=1}^{T} \sum_{g=1}^{T} z_{i,hg} \hat{\varphi}_{k,gd}^{(s)} z_{i,td}$ and we would get

$$
\begin{aligned}
D_{ik}^{(s)} &:= \mathbb{E}(Z_i \hat{\Phi}_k^{-1(s+1)} Z_i^\intercal | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1)) = && (11)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T} \sum_{g=1}^{T} z_{i,hg} \hat{\varphi}_{k,gd}^{(s)} z_{i,td} \Big)_{h,t} | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right) && (12)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T} \sum_{g=1}^{T} z_{i,hg} z_{i,td} \hat{\varphi}_{k,gd}^{(s)} \Big)_{h,t} | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T} \sum_{g=1}^{T} S_{ik,[(g-1)J+h,(d-1)J+t]} \hat{\varphi}_{k,gd}^{(s)} \Big)_{h,t} | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right), && (13)
\end{aligned}
$$

where in Equation 13 we make use of the the elements of S_{ik}.

As written above, we would also need to compute $\mathbb{E}(Z_i^\intercal \hat{\Sigma}_k^{-1(s+1)} Z_i | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1)$, which we can do by following the same reasoning. By defining $C_{ik}^{(s+1)} := \mathbb{E}(Z_i^\intercal \hat{\Sigma}_k^{-1(s+1)} Z_i | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1)$ and by denoting by $\hat{\sigma}_{k,gd}^{(s+1)}$ the $(g,d)^{th}$ element of $\hat{\Sigma}_k^{-1(s+1)}$. Then, the $(h,t)^{th}$ element of $Z_i^\intercal \hat{\Sigma}_k^{-1(s)} Z_i$ is $\sum_{d=1}^{J} \sum_{g=1}^{J} z_{i,gh} \hat{\sigma}_{k,gd}^{(s+1)} z_{i,dt}$, and we get

$$
\begin{aligned}
C_{ik}^{(s+1)} &:= \mathbb{E}(Z_i^\intercal \hat{\Sigma}_k^{-1(s+1)} Z_i | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1) = && (14)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{J} \sum_{g=1}^{J} z_{i,gh} \hat{\sigma}_{k,gd}^{(s+1)} z_{i,dt} \Big)_{h,t} | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right) && (15)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{J} \sum_{g=1}^{J} z_{i,gh} z_{i,dt} \hat{\sigma}_{k,gd}^{(s+1)} \Big)_{h,t} | \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right) && (16)\\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T} \sum_{g=1}^{T} S_{ik,[(h-1)J+g,(t-1)J+d]} \hat{\sigma}_{k,gd}^{(s+1)} \Big)_{h,t} | \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{Y}_{ir} = 1 \right). && (17)
\end{aligned}
$$

Finally, this means that computing $\mathcal{Q}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}^{(s)})$ requires to compute:

- $\mathbb{E}(\ell_{ik}|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) = \frac{\pi_k^{(s)}\int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ}{\sum_{k=1}^K \pi_k^{(s)}\int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ} =: \boldsymbol{\tau}_{ik}^{(s+1)}$,
- $\mathbb{E}(z_i|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: m_{ik}^{(s+1)}$,
- $\mathbb{E}(z_i z_i^\intercal|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: S_{ik}^{(s+1)}$, whose elements are required for the computation of $D_{ik}^{(s+1)}$ and $C_{ik}^{(s+1)}$.

3.5 M-step

By taking the first derivatives of Equation 3, the maximum likelihood estimators of the parameters are given by

$$\hat{\pi}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\boldsymbol{\tau}}_{ik}^{(s+1)}}{N} \quad , \quad \hat{M}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\boldsymbol{\tau}}_{ik}^{(s+1)}\hat{M}_{ik}^{(s+1)}}{\sum_{i=1}^N \hat{\boldsymbol{\tau}}_{ik}^{(s+1)}}, \tag{18}$$

where $\hat{M}_{ik}^{(s+1)} := \mathbb{E}(Z_i|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) = \mathbb{E}(\text{vec}_{J\times T}^{-1}(z_i)|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) = \text{vec}_{J\times T}^{-1}(m_{ik})$, and $\text{vec}_{J\times T}^{-1}$ is the inverse of the vectorization function, i.e. the function mapping from a JT-dimensional vector to a $J\times T$ matrix. The two covariance matrices are interdependent and require the computation of $C_{ik}^{(s+1)}$ and $D_{ik}^{(s+1)}$ to be done using the parameters updated

$$\hat{\Sigma}_k^{(s+1)} = \frac{\sum_{i=1}^N \boldsymbol{\tau}_{ik}^{(s+1)}[D_{ik}^{(s+1)} - \hat{M}_k^{(s+1)}\hat{\Phi}_k^{-1(s)}M_{ik}^{\intercal(s+1)} - M_{ik}^{(s+1)}\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal(s+1)} + \hat{M}_k^{(s+1)}\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal(s+1)}]}{T\sum_{i=1}^N \boldsymbol{\tau}_{ik}^{(s+1)}}, \tag{19}$$

$$\hat{\Phi}_k^{(s+1)} = \frac{\sum_{i=1}^N \boldsymbol{\tau}_{ik}^{(s+1)}[C_{ik}^{(s+1)} - \hat{M}_k^{\intercal(s+1)}\hat{\Sigma}_k^{-1(s+1)}M_{ik}^{(s+1)} - M_{ik}^{\intercal(s+1)}\Sigma_k^{-1(s+1)}\hat{M}_k^{(s+1)} + \hat{M}_k^{\intercal(s+1)}\hat{\Sigma}_k^{-1(s+1)}\hat{M}_k^{(s+1)}]}{J\sum_{i=1}^N \boldsymbol{\tau}_{ik}^{(s+1)}}. \tag{20}$$

We remark the invariance of the order, but it is important to use the updated parameters coherently, so that if $\hat{\Phi}_k^{(s+1)}$ is computed before $\hat{\Sigma}_k^{(s+1)}$, then $D_{ik}^{(s+1)}$ will need to be computed using the new $\hat{\Phi}_k^{(s+1)}$.

3.6 Initialization

To find the initial values of $\hat{\boldsymbol{\Theta}}^{(0)}$ mentioned in Section 3.2, our proposal is the following. Identity matrices are chosen for the initialization of the covariance matrices Φ_k and Σ_k, while $\pi_k = 1/K$. For the initialization of M_k, two solutions are proposed and tested in Section 4.2. The first is a Kmeans++ (Arthur et al., 2007) initialization, that is performed on the vectorized data. The second is a multiple random initialization: the mean matrices M_k are chosen by uniform sampling K matrices among the N observed

data matrices. Since the EM algorithm is not guaranteed to converge toward a global optimum, the algorithm is applied multiple times and the results with the highest log-likelihood is selected. For simulations in Section 4.2, 5 random initializations proved to be enough, but for more complex setting a higher number might be needed.

3.7 Selection of the number of cluster K

The number of cluster K is selected by minimizing the BIC (Schwarz, 1978) criterion. The BIC for a number of cluster k is defined as

$$\text{BIC}_k := -2 \log \mathcal{L}_O(\boldsymbol{\Theta}; \tilde{\boldsymbol{Y}}) + \nu_k \log(N),$$

where ν_k is the total number of model parameters:

$$\nu_k := k[1 + JT + J(J+1)/2 + T(T+1)/2] - 1,$$

and $\mathcal{L}_O(\boldsymbol{\Theta}; \tilde{\boldsymbol{Y}})$ is the observed likelihood of the model, that is

$$\mathcal{L}_O(\boldsymbol{\Theta}; \tilde{\boldsymbol{Y}}) := \prod_{i=1}^{N} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \pi_k \int_{\Omega_r} f(Z|\Theta_k) dZ \right)^{\tilde{Y}_{ir}},$$

since $\tilde{Y}_i$ would still have multinomial distribution but, not conditioning on Z_i and ℓ_{ik}, ξ_i is not deterministic.
To select the model with the optimal K, the algorithm needs to be executed for every $k = 1, ..., K$ and the model with with the lowest BIC_k is chosen.

4 Evaluation

This section presents numerical experiments on simulated data in order to illustrate the behavior of the proposed model regarding the influence of the initialization procedure and sample size, the robustness to different noise ratio in the data, the model selection and in comparison with its continuous counterpart when used on ordinal data treated like quantitative data.

4.1 Simulation Setup

20 different samples have been simulated for increasing number of units $N \in \{300, 1500, 3000\}$, with $K = 3$, $J = 5$, $T = 5$, $\pi = (0.3, 0.4, 0.3)$ and $C_j = 5$ levels $\forall j = 1, \ldots, J$. Each sample has been drawn from a matrix-variate Gaussian and then discretized according to the thresholds chosen in Section 3.1. The distributions parameters were chosen as following: identity covariance matrices Φ and Σ and mean matrices M chosen such that

the estimated the optimal Adjusted Rand Index (ARI) (Rand, 1971), computed by performing one step of the clustering algorithm using the true parameters, would be around 0.85. This setting led to the choice three mean matrices $M_1 = 1.75 \cdot \mathbf{1}_5\mathbf{1}_5^\top$, $M_2 = 2.5 \cdot \mathbf{1}_5\mathbf{1}_5^\top$ and $M_3 = 3.25 \cdot \mathbf{1}_5\mathbf{1}_5^\top$, where $\mathbf{1}_5$ is a 5-dimensional vector whose elements are all 1.
Moreover, three scenarios are derived from this setting by adding some noise fraction within the clusters by simulating a percentage τ of units using a uniform distribution on levels C_j, allocated to the three clusters proportionally to the clusters' size: 0% (scenario 1), 10% (scenario 2), 20% (scenario 3).
The two different kinds of initialization described in Section 3.6 have been tested.

Finally, in our case we use a double stopping rule: either the difference between observed log-likelihood at step $(s+1)$ and (s) is lower than 0.001, or a maximum number of steps is reached. We fixed the number of maximum steps to 200.

Regarding the algorithm setup, we set to 100 iterations as the burn-in period of Gibbs sampler in the E-step, and a thinning equal to 2 to prevent too correlated samples. The number of simulated samples is set to 100. Computation time for one iteration on 2.40 GHz 11th Gen Intel Core i5-1135G7 with 16 Go RAM for one step of the algorithm with Kmeans++ initialization is about 8 seconds for $N = 300$ and about 80 second for $N = 3000$.

4.2 Influence of initialization & sample size

This first experiment aims at studying the ability of MOM to recover the simulated model depending on the type of initialization of the EM algorithm. Figure 1 shows the quality of estimated partitions assessed by means of ARI. We recall that an ARI of 1 indicates that the partition provided by the algorithm is perfectly aligned with the simulated one. Conversely, an ARI of 0 indicates that the two partitions could as well be some random matches. On the graph, the optimal ARI (≈ 0.85) according to the simulation scheme is represented by a horizontal line. The boxplots do not seem to show any significant difference in the median values of the ARI measurements between the two initialization methods, but for sample size equal to 300 there seems to be a greater variability in the results for Kmeans++, probably indicating that good results for this initialization could be better than good results of the other, but that bad results tend to be worse as well.

Overall, from a partitioning point of view, the two initialization techniques do not seem to produce significantly different results. We decided to measure their performance also by computing the Mean Absolute Percentage Error (MAPE) on their estimation of the distribution parameters. The MAPE calculates the average percentage difference between the actual and predicted values of a variable, therefore providing a relative measure of error. For a sample of N units, for a generic parameter θ it is expressed through the formula:

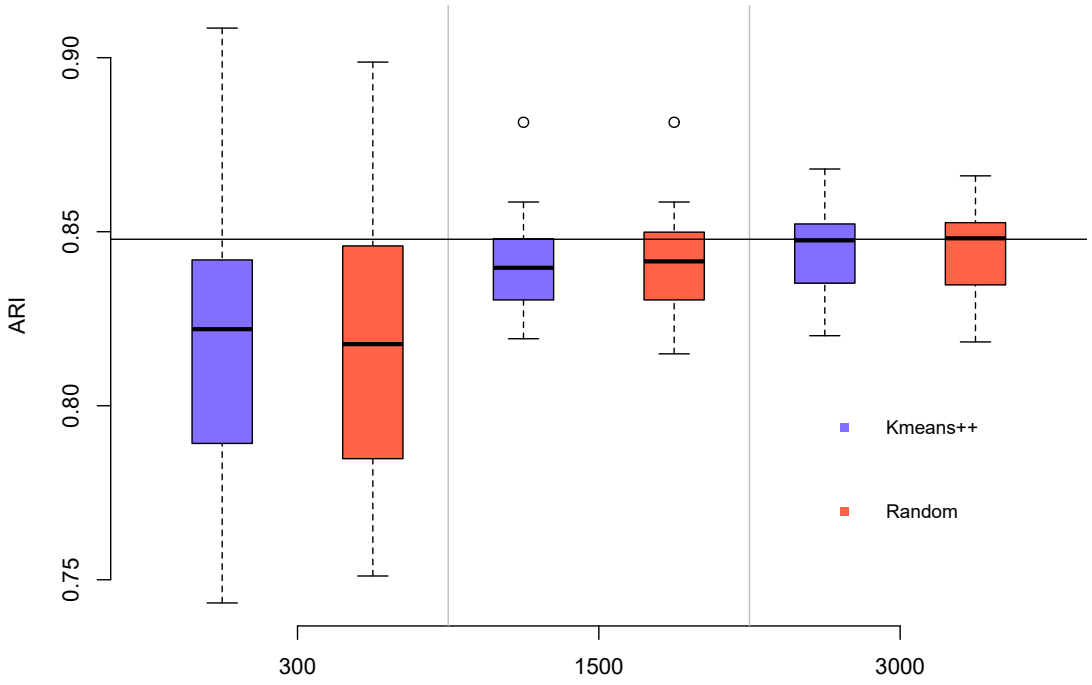


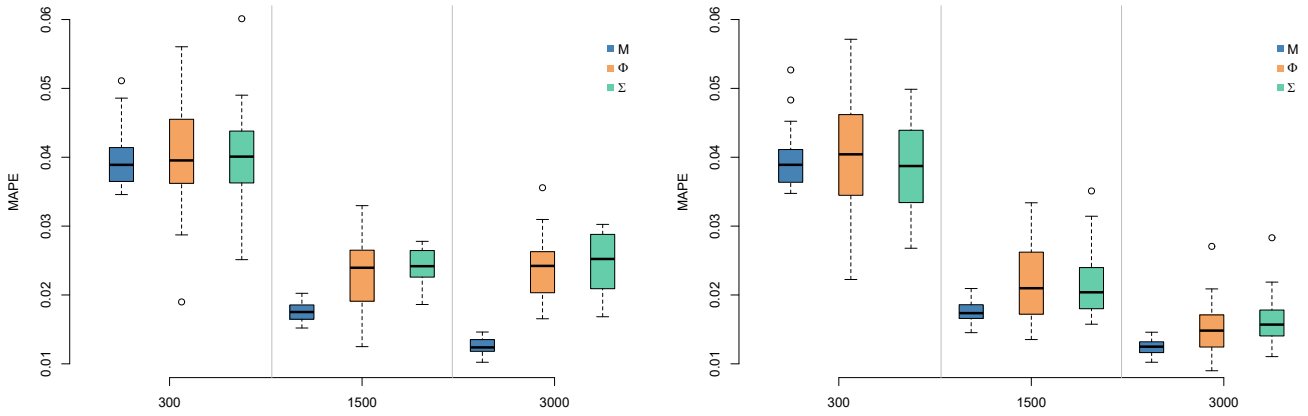
Figure 1: Influence of initialization. The horizontal line represents the estimated Bayesian error.

$$\text{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{\theta_i - \hat{\theta}_i}{\theta_i} \right|,$$

where $\hat{\theta}_i$ is the estimated parameter and θ_i is the true parameter. MAPE has some limitations, such as the fact that it cannot be used when actual values are zero or close to zero. This is why for the covariance matrices only the diagonal elements are considered.

Results are shown in Figure 2. While for $N = 300$ there not seems to be a clear difference, the random initialization appears to have lower median values for the two covariance matrices for $N = 1500$. The difference appears even more evident for increasing sample sizes.

Concerning overall the influence of the sample size, the model behaves as expected: as the sample size increases, the partitioning capabilities improve and tend towards the optimal error. The same happens when we observe the errors concerning the parameter estimations for both the initialization procedures, even if the median MAPE values appear to reach a stable value already for $N = 1500$ for the Kmeans++ initialization, while the values improve further for $N = 3000$ for a random initialization.



(a) Kmeans++ init

(b) Random init.

Figure 2: MAPE for increasing N

Globally, there not seems to be a significant difference in terms of performance results for the two initialization procedures regarding the partitioning capabilities. However, as the sample size increases, parameters estimation quality improves more with a random initialization than with a Kmeans++ one. Nonetheless, it is worth noticing that the random initialization is to some extent a greedy procedure which requires to compute the algorithm several times with the purpose of selecting the best result, and therefore, depending on the number of random initializations chosen, it can easily become time-consuming and computationally costly.
In the following, given the similarities in performance and the computationally advantages, we will carry out most of the analysis using only the Kmeans++ initialization.

4.3 Robustness to noise

As written in Section 4.1, we also simulated some noisy data to study the behaviour of MOM when the underlying normality assumption is not fully respected. ARI for different noise proportions were measured and the results are visible in Figure 3. We decided to measure two quantities: the overall ARI for all the units and the ARI just for the non-noisy ones.

As we would expect, the overall quality of partitioning estimates decreases as the level of noise increases, indicating that MOM is actually disturbed by the noise.

Interestingly, for N large enough, the model proofs itself robust and it classes perfectly non-noisy data, reaching the optimal ARI, represented by the horizonalt black line in the graph. For $N = 300$, the noise disturbs the model estimate, and we do not get a good ARI, even for non-noisy data. The clustering therefore seems well disturbed by outliers

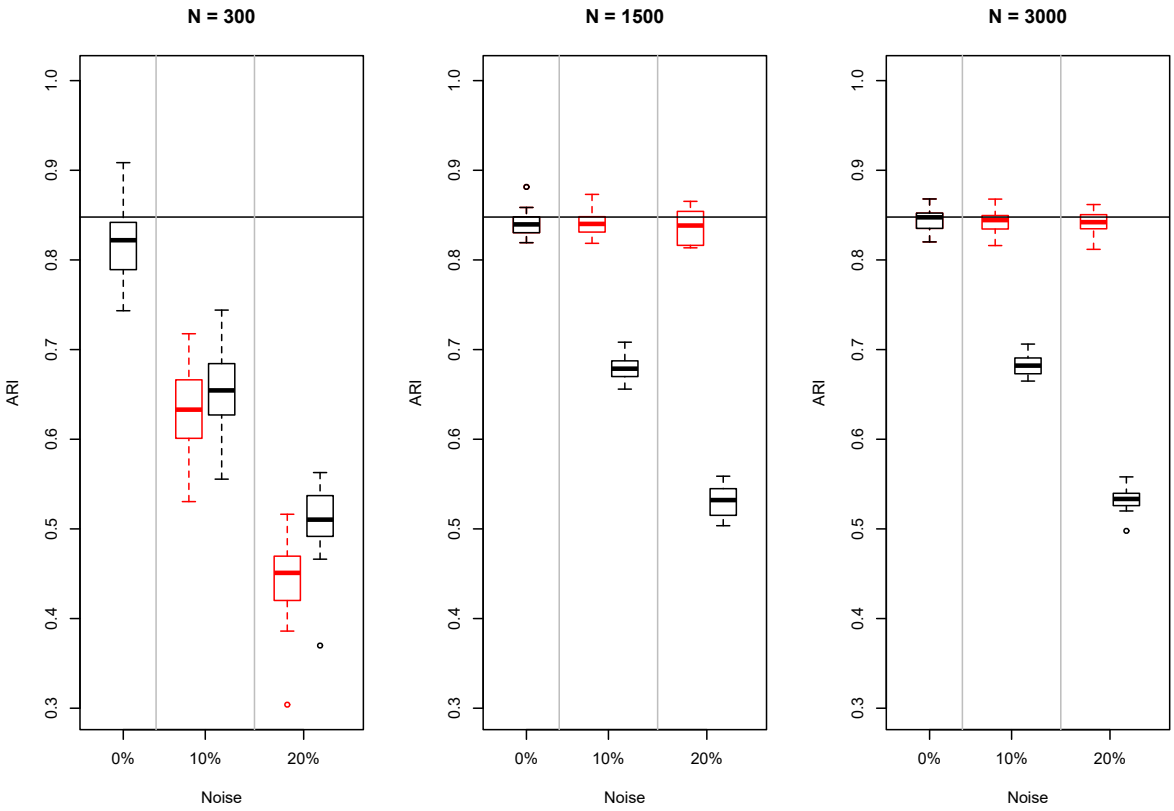


Figure 3: ARI for increasing noise proportions and increasing N.
In red the ARI for non-noisy units, in black for all of them.

when N is small, but it corrects when N increases.

4.4 Model Selection

Following the setup described in Section 4.1, by varying $N \in \{300, 1500, 3000\}$ and adding increasing noise ratios $\tau \in \{0, 0.1, 0.2\}$, 9 different scenarios have derived for testing the model selection capabilities. We recall that for each scenario and each N, 20 data sets have been drawn. Model selection has been performed through BIC, as described in Section 3.7. The results are shown in Table 1.

For $N = 300$ and $\tau = 0$, 70% of the simulated data sets have a lower BIC for K equal to 2 than 3. However, for larger sample sizes, the model with $K = 3$ is selected for each synthetic data sets in each scenario. The model seems therefore sensitive to sample sizes as small as 300, and seems prone to select a value for K smaller than the actual one for small samples. Probably, this is due to the higher number of parameters that the true value of K requires to estimate compared to a lower one, making $N = 300$ not sufficient

to effectively compute them.

Scenario $\tau = 0$							Scenario $\tau = 0.1$							Scenario $\tau = 0.2$						
N/K	1	2	3	4	5	6	N/K	1	2	3	4	5	6	N/K	1	2	3	4	5	6
300	0	14	6	0	0	0	300	0	20	0	0	0	0	300	0	20	0	0	0	0
1500	0	0	20	0	0	0	1500	0	0	20	0	0	0	1500	0	0	20	0	0	0
3000	0	0	20	0	0	0	3000	0	0	20	0	0	0	3000	0	0	20	0	0	0

Table 1: Frequency of selection of each model K by MOM through BIC among the 20 simulated data sets, for increasing N. The actual value for K is 3. Kmeans++ initialization.

Looking at the performances in selecting the right K in presence of noise, we can say that overall the model seems able to handle well some noise in the data, provided a sufficient number of remaining non-noisy units to draw its inference from is given. It keeps optimal classification results for units which follow the distributional assumption and selects the correct model even for $\tau = 0.2$.
At the same time, the presence of noise makes more extreme the problem of selection of K for small sample size described in the previous paragraph, as the model has even fewer non-noisy units to compute the parameters from.

Finally, the reader should acknowledge that the same experiment for $N = 300$ with random multistart initialization was performed, and it yielded the same result as for the Kmeans++ initialization.

4.5 Comparison with Continuous Model

Finally, we compared the results obtained for our model MOM to the ones we given by its continuous version MMN, by treating our ordinal data as continuous ones, as frequently done by practitioners.

In Figure 4 the results for the partitioning task are shown. The difference in the ARI measurement is bigger for smaller number of units, and decreases as the N increases. However, MOM outperforms MMN consistently.

The difference is clearer when comparing the MAPE values for the parameters estimation. As shown in Figure 5, the distance in error increases as N increases for M and Σ, especially for the latter. On the other hand, the same does not happens for the diagonal of Φ, for which not only the difference in error is less pronounced, but for $N = 3000$ the two methods seem equivalent in that regard.

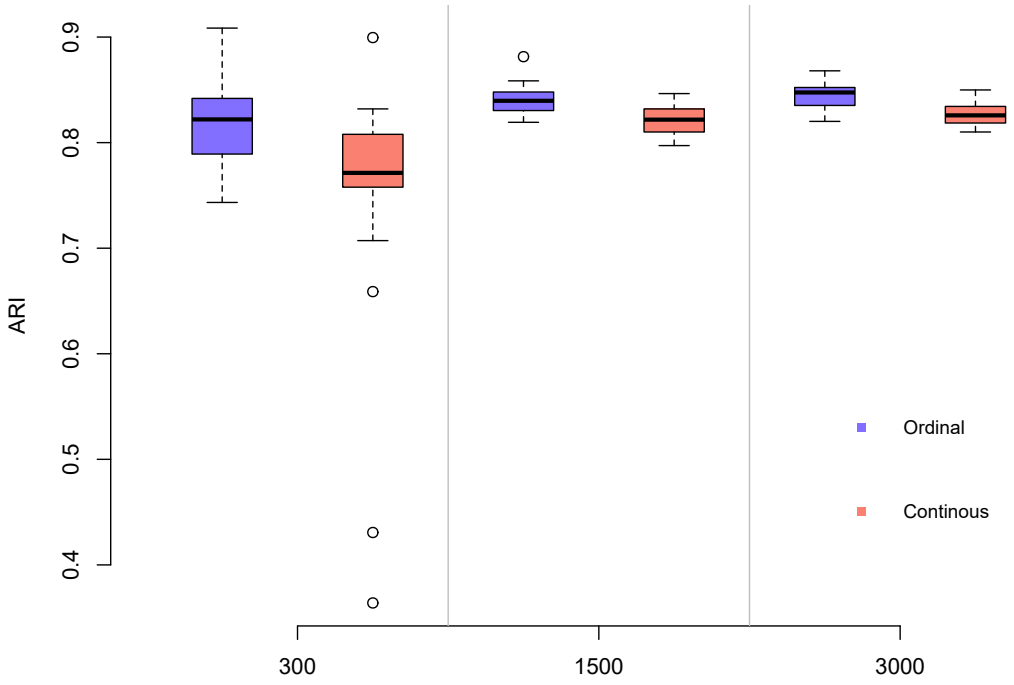
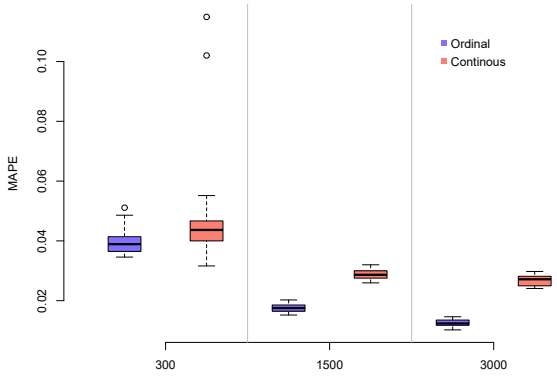


Figure 4: ARI for MOM and MMN. Kmeans++ initialization.

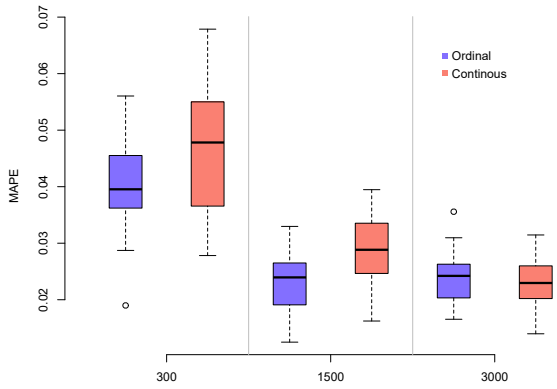
5 Real Data

5.1 Data

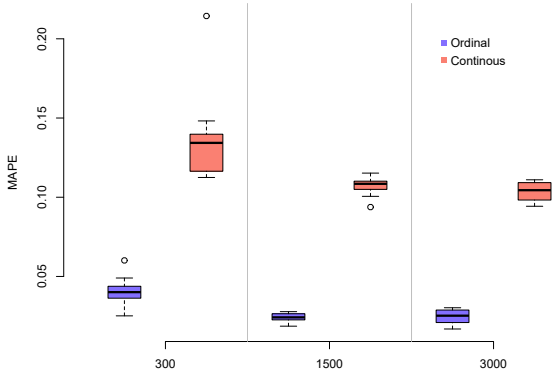
After the evaluation of the model through simulations, a real data application concerning preferences for grocery shopping during the Covid-19 pandemic in France (Franois-Lecompte et al., 2020) has been performed. The surveys consists of 78 questions for the first survey (T1), 73 questions for the second (T2) and 55 questions for the remaining three surveys (T3, T4, T5). The answers are mainly on an ordinal scale, and has been conduced at 5 period during the two years of pandemic's intermittent lockdowns to a French sample. The five period at which the surveys has been conduced are: March 26 - April 5, 2020 (beginning of the 1st lockdown); April 30 - May 11, 2020 (end of the 1st lockdown); June 9 - June 16, 2020 (post-lockdown); October 28 - November 9, 2020 (beginning of the 2nd lockdown); March 5 - March 25, 2021 (just before the 3rd lockdown). As part of a preliminary analysis on the data, we have selected 11 questions coming from 3 macro-area of questioning (quoted as Q5, Q8 and Q12). The total number of participants answering for these 11 questions at each of the 5 surveys is 337. Translated to English, the questions are the following:



(a) MAPE for M



(b) MAPE for Φ



(c) MAPE for Σ

Figure 5: MAPE results for parameter matrices. MOM vs MMN. Kmeans++ init. Note the difference in the scales.

- Q5: In the last month, you would say that you have preferred in your purchases…
  - (1) Seasonal products
  - (2) Products "Bio"
  - (3) Local products
  - (4) Fair trade products
  - (5) Bulk products (excluding fruit and vegetables)
- Q8: Choose the appropriate answer for each item
  - (1) About the foods, you have the impression of wasting

  - (2) You have paid attention to the expiration dates
  - (3) You have prepared anti-waste cooking recipes
- Q12: Would you say
  - (1) This period is ideal to rethink our way of consuming
  - (2) This period is ideal to test more environmentally responsible ways of living
  - (3) This period is ideal to learn how to consume less

For each question, the participant have to answer on an ordinal scale 7 levels: for the macro-group Q5 and Q8 the range is from 1 for "much less than before confinement" to 7 for "much more than before confinement", while for the macro-group Q12 from 1 for "high disagreement" to 7 for "high agreement". In all of the cases the 4th level express some form of "neutrality".
It is worth noticing that the item Q8(1) is an inverse item. As we will see, this will not impact our clustering, as our model is able to handle such items without the need to reverse them, but it is necessary to keep in mind their nature at the moment of interpretation, as it would impact the direction of the correlation with the other items.

So, to sum up, we have N = 337 units for J = 11 variables (questions) and T = 5 times.

5.2 Results

After performing our clustering algorithm with a number of clusters K ranging from 1 to 6 using Kmeans++ initialization, the model with the lowest BIC is with K = 3 (Figure 10). The number of units in first cluster is 124, in the second one they are 149 and in the third 64. The estimated parameters are reported in Table 2 for the mean M, Table 5 for the time covariances Φ and in Table 6 for the variable covariances Σ. To gain interpretability, covariances matrices have been transformed in correlation matrices in Table 3 for Φ and in Table 4 for Σ. In the tables the questions are named using their codes. Moreover, the correlation matrices Φ and Σ are represented by correlation plots in Figures 8 and 9, respectively.

Figure 6 represents the 337 units (individuals) using a non-metric MDS (Venables et al., 2002), specifically through the function isoMDS of the R package MASS. In non-metric MDS only the order of dissimilarities is important rather than the amount of dissimilarities, that makes it suitable to be used for ordinal data, as in our case. For this representation, the temporal structure has been discarded and we have transformed our units from 11×5-dimensional matrices to 55-dimensional vectors. Each individual is represented by a circle whose color depends on its cluster.

Figure 7 plots, using the same non-metric MDS, the cluster means at each of the 5 times. Such plot allows to visualize the time evolution of each cluster.

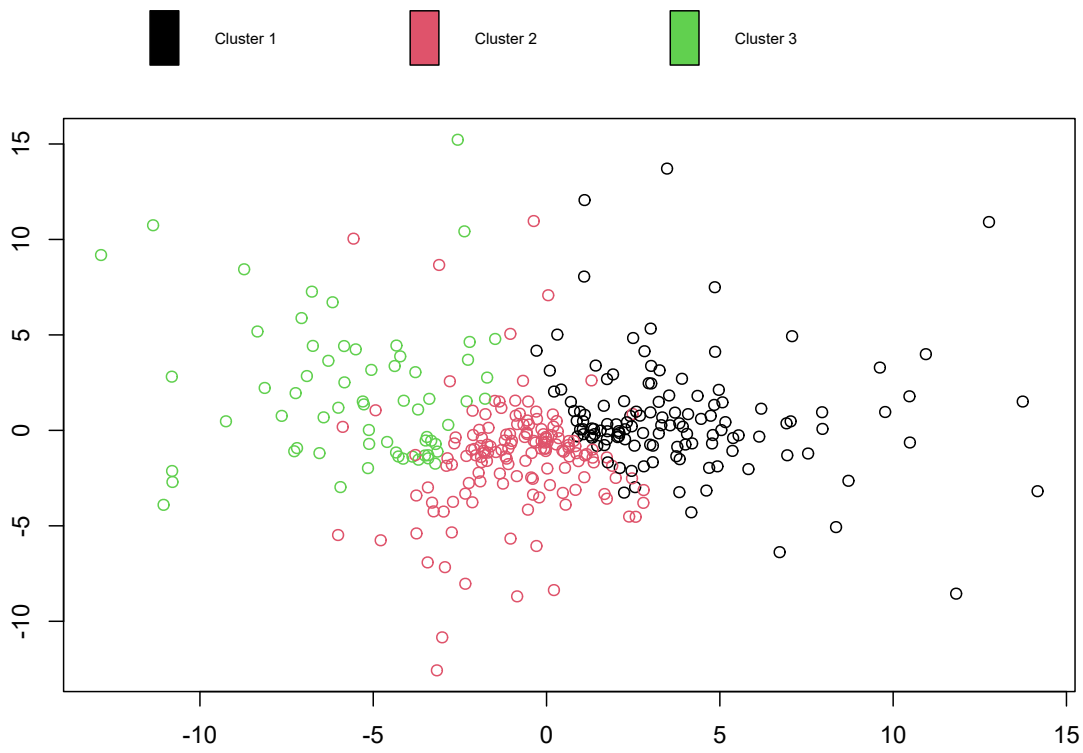


Figure 6: Units represented through isoMDS and colored by cluster allocation.

5.3 Interpretation

Even if data are represented by means of a dimensionality reduction technique, discarding the temporal structure of data, we can see on Figure 6 that the clusters are well separated. In particular, Cluster 2 is between Cluster 1 and Cluster 3. This fact can be confirmed by looking more finely at Table 2.
Moreover, from Figure 7 it is possible to visualize the comprehensive evolution in time for the clusters means. Indeed, one can see that Cluster 2 and Cluster 3 starts relatively close to one another, but Cluster 2 then evolves and approaches Cluster 1 in T3, to then stabilizing on a more intermediate space. Cluster 1 appears to be the most stable one, moving itself on a confined area of the graph. Cluster 3, despite starting on values close to Cluster 2, evolves differently from the others.

In the following, we give a summary description for each cluster and we will try to draw some interpretations. We will start by interpreting Clusters 1 and Cluster 3, which are the most characteristic, to finish with the Cluster 2, which could be seen as an intermediary cluster between the other two.

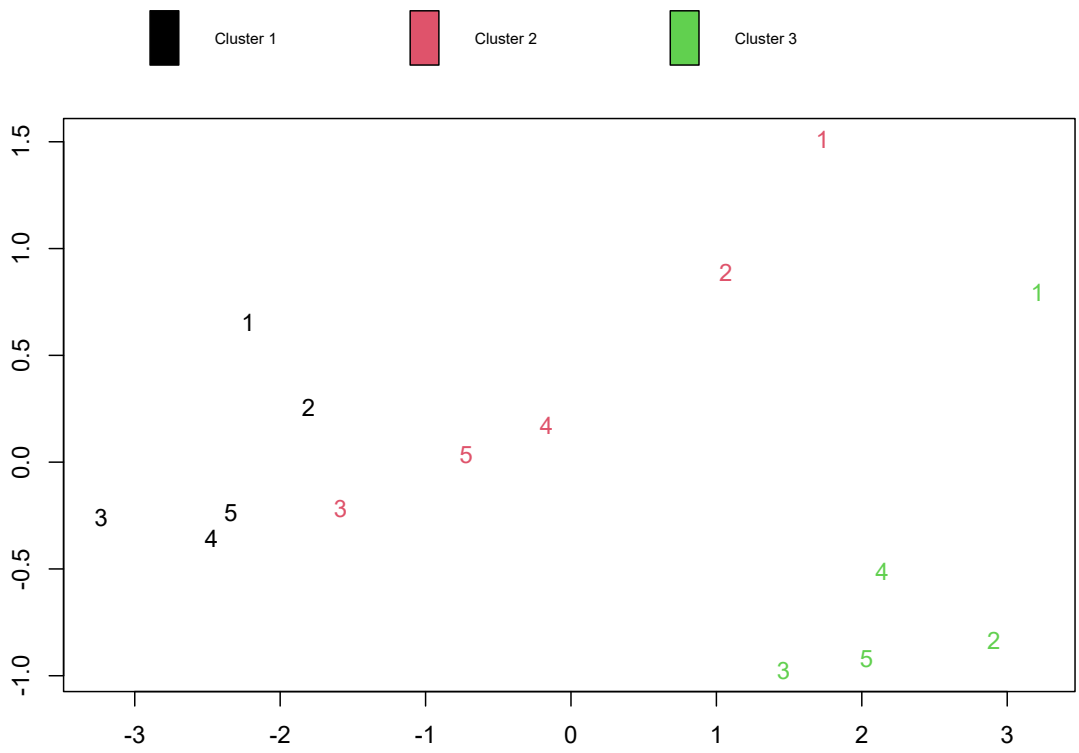


Figure 7: Evolution in time of cluster means. Representation through isoMDS. Numbers represent the time intervals and the colors the clusters.

- **Cluster 1**: 124 units.
  - **Correlation in time**: the cluster is characterized by a fading correlation of T1 with other times and by generally higher correlations than other clusters, with the exception of just a small rift between T2 and T4.
  - **Means**: this is the cluster with overall lowest and most stable mean values, around neutrality level. The only values lower than neutrality are for Q8(1), an inverse item.
  - **Correlation among questions**: generally positive correlations or feeble ones, the cluster is mainly characterized by some positive correlations between macro-area Q5 and Q12, and some negative correlations between those areas and Q8(1).

  We can characterize Cluster 1 as the cluster with overall neutrality-level and stable means. Indeed, considering that levels range from 1 to 7 as detailed in Section 5.1,

the values tend to be around the "neutrality" level, the level coded as 4. Therefore, the cluster is actually a cluster composed by people who were generally neutral with respect to the questions, and did not evolve on this neutrality much during the study period.
Looking at Table 2, it is evident that the questions that discriminate the most among the clusters in terms of average level of response are the ones in Q12, the ones regarding rethinking our lifestyle, as they show different average levels for each cluster. For cluster 1, the average response shows neutrality even in that regard.
This cluster is also the ones that has the highest correlations between Q8(3), anti-wasting recipes, and questions in Q5 group. Overall, observing the behaviour of the correlations, seems clear that Cluster 1, despite being the most neutral cluster in terms of average responses, could be defined as the most consistent cluster, since responses that regarding preferences for sustainable grocery shopping are positively correlated with preparations anti-waste recipes and rethinking our way of life.
The generally positive correlations among some of the other questions may indicate a certain coherence around the neutrality, given that preference for a more sustainable grocery shopping is positively correlated to the anti-wasting behaviours and the belief that the pandemic period should inspire a change in the life habits. This signals that the subjects' responses to those topics move likewise within the cluster. In other words, Cluster 1 did not really change its habits (as level 4 means "as before") and appears not to have felt very impacted by the health crisis, as the neutral level on rethinking its way of life may indicate.

- **Cluster 3**: 64 units.
  - **Correlation in time**: Cluster 3 seems defined by two correlations blocks; one composed by T1 and T2 and the second by T3,T4 and T5.
  - **Means**: with respect to the other clusters, this cluster is characterized by the highest levels for the macro-area Q12 and the lowest values for Q8(1), coherent with the inverse item.
  - **Correlation among questions**: the cluster is the most varied one compared to the other clusters. Intra-macro-area correlations are weaker as well. Some noteworthy negative correlation between Q8(2) and Q5(2) and between Q12(3) and Q5(3).

  Cluster 3 also has generally neutrality-level values for most of questions belonging to Q5 and Q8 macro-groups throughout the study period, as Cluster 1, despite having some lower values for Q8(1) and some higher ones for Q5(3). The main difference is however in the Q12 macro group, the one we can define as composed by the "rethinking-way-of-life" questions. Cluster 3 has remarkably high values here, meaning that this group of people really found that the pandemic period was

stimulating a reflection on our lifestyle. As it turns out, this opinion fades as we advance towards T3 to then re-approach higher levels. It is interesting to observe that T3 corresponds to the beginning of June 2020, that is after the end of the first lockdown, while T4 is at the end of October and beginning of November 2020, after the summer and at the beginning of the second lockdown, and that T5 is in March 2021, when the country was approaching a third lockdown. So, apparently, the second lockdown brings back a reflection on how to live. It seems that people need crises to reflect on their lifestyle.
In this cluster we also observe some negative correlations between question Q12(3), concerning less consumption, and Q5(3), which measures the preference for local products, and also between Q8(2), paying attention to expiring dates, and Q5(2) and Q5(4), the preference for "bio" products and fair trade ones. This may signal that the people composing this cluster who pay more attention to buy "local" (such as going to the local markets), "bio" and sustainable fair product may also be the ones who tend to be less concerned regarding consuming less, probably because they already satisfy their concerns by orienting their grocery shopping to more sustainable products. They satisfy their concerns for consuming less by consuming better.

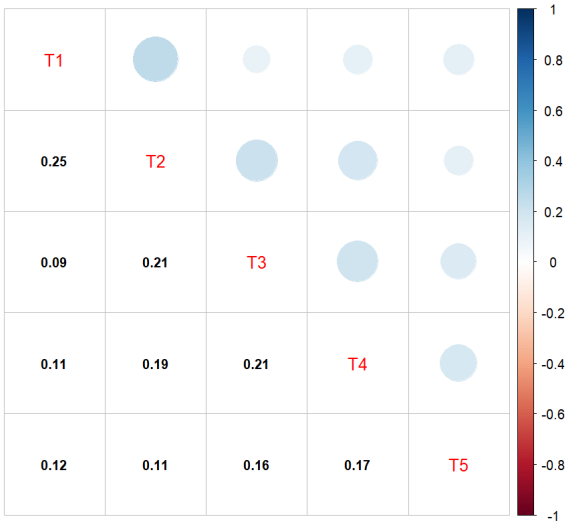
- **Cluster 2**: 149 units.
  - **Correlation in time**: cluster 2 presents notably overall fading correlations in time.
  - **Means**: responses for macro-areas Q5 and Q8 show levels around neutrality, while for macro-area Q12 the levels are middle-high, intermediary between the other two clusters.
  - **Correlation among questions**: cluster characterized by generally low correlations among questions of different macro-areas. Some weak negative correlations among Q12(3) and Q5(5) and Q8(2).

  Cluster 2, as already said, seems composed by subjects whose answers to the questionnaires can be seen as intermediate between the Cluster 1 and Cluster 3. Levels for questions in the Q5 group tend to be lower at the beginning of the inquiry to then have a slight increase over the study period. Questions of the macro-group Q12, that we saw characterized cluster 3 for their high levels in the answers, have an high level for this cluster at the beginning as well, even if not as high as cluster 3. Yet, their value tend towards the "neutrality" approaching T3, to then have a slight increase. We can think of these subjects as people that highly agreed with changing their way of life at the beginning of the inquiry, to then become more and more disaffected as the strict lockdown period gives way to reestablish a more 'ordinary' way of life.

One characteristic of Cluster 2 is that there are not clearly strong correlations outside macro-area blocks, ans even for block Q8 they are not as strong as other clusters. This may indicate heterogeneity in the answers' patterns to the questioners outside the blocks, giving rise not so strong correlations.
Some weak negative correlations between Q12(3) and Q5(2) and between Q8(2) and Q5(2) may signal a similar behaviours as in Cluster 3 regarding satisfying their concerns for consuming less by consuming better, even if less pronounced.

Finally, overall, there are some comments to be made about Q8(1) and intra-group correlations. Question Q8(1) has indeed negative correlations with other questions. The question asks whether the respondent has the impression to waste. Its negative correlation with questions in Q5 and Q12 group, even if only slightly sometimes, means that people that in general have the impression of wasting food are the ones that report lower values regarding preferences for "sustainable" grocery shopping and rethinking our way of consuming, while, vice-versa, subjects whose responses have higher values regarding buying local and seasonal product, like people who go to local markets, tend to have a lower impression of wasting, probably because they actively try not to. This indeed connects to the general negative correlation that question Q8(1) and Q8(3) have: as Q8(3) asks whether the respondent has prepared anti-wasting recipes, the negative correlation seems natural.
On a final note, it is worth pointing out that the cluster that has the lowest correlations for Q8(1) is Cluster 2, as maybe it contains people that try to buy locally and seasonal but do not arrive at making the effort to prepare anti-waste recipes.

(a) Cluster 1

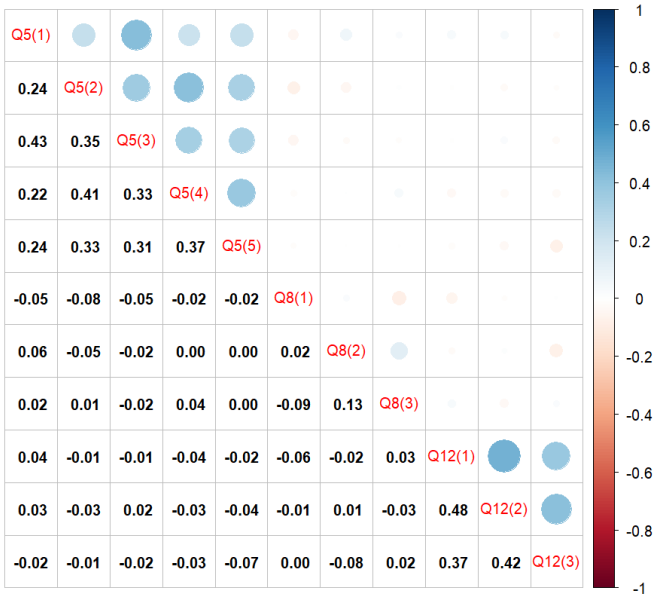


(b) Cluster 2

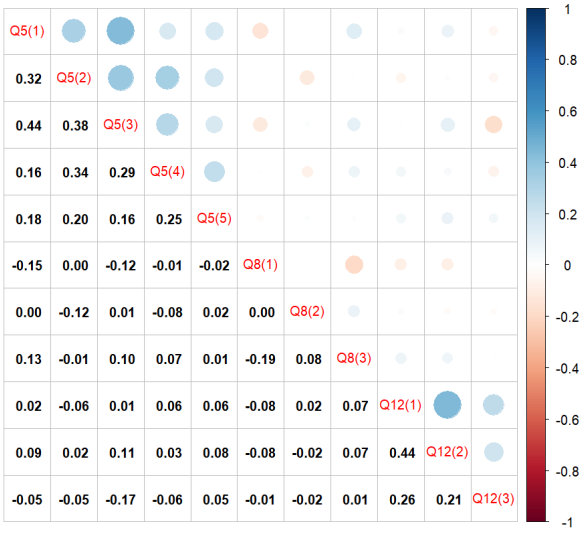
(c) Cluster 3

Figure 8: Clusters' corr-plots among time.

(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Figure 9: Clusters' corr-plots among variables.

6 Conclusions

In this work we have presented a novel approach for modeling longitudinal ordinal data with unobserved heterogeneity. The model presented does not require the conditional independence assumption and respects the the true nature of ordinal data. The matrix-variate structure allows for a more parsimonious modelling. Also, it can explicitly model the temporal structure and the association among the responses, that can vary among clusters. An EM algorithm to perform inference has been proposed and described. The efficacy of the algorithm has been tested on synthetic data under different sample sizes and different noise ratios. We proved the goodness of this framework to cluster longitudinal ordinal data and to get cluster that are easy to interpret and to work with even by non-statisticians.

However, the proposed model has some limitations. In this paper we focused only on the simplest structure of matrix-normal distribution. While considerably more parsimonious than a mixture of multivariate normal distributions, the model seems sensitive to small sample sizes, as seen in Section 4.4, since, as the number of clusters increases, the number of parameters to estimate can still became troublesome. To improve this aspect,the covariance matrices can be further decomposed to obtain more flexible and parsimonious models, as done for example in Anderlucci et al., 2015 and in Sarkar et al., 2020. Moreover, EM algorithm can be leveraged to extend the model to deal with incomplete data under the missing at random (MAR).

Furthermore, typically the data collected in questionnaires are not just ordinal, but rather mixed. Consequently, our final aim is to extend the proposed model to handle longitudinal mixed data, following the frame proposed by McParland et al., 2016.
Moreover, one could as well think of implying, with proper adjustments, different underlying continuous distributions, such as heavy-tailed (Tomarchio et al., 2020), skewed (Gallaugher et al., 2018, Melnykov et al., 2018) or t-student (Doğru et al., 2016) distributions to endow the clustering model with different desired properties.

Acknowledgment

This work has been realised thanks to the financial support provided by Project IADoc@UdL of the University of Lyon and Université Lumière - Lyon 2 as part of the call for "doctoral contracts in artificial intelligence 2020" (ANR-20-THIA-0007-01).
We want to thank Agnès François-Lecompte, Morgane Innocent and Dominique Kréziak, co-authors for their work in Franois-Lecompte et al., 2020 and for sharing their data.
We would also like to thank Brendan Murphy invaluable inputs and support throughout the research process. His insights and expertise were instrumental in shaping the direction of this project.

References

[1] Rensis Likert. "A technique for the measurement of attitudes." In: *Archives of psychology* 140 (1932), pp. 5–55.

[2] S. S. Stevens. "On the Theory of Scales of Measurement". In: *Science* 103.2684 (June 1946), pp. 677–680. ISSN: 0036-8075. DOI: 10.1126/science.103.2684.677.

[3] William M. Rand. "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336 (Dec. 1971), pp. 846–850. ISSN: 0162-1459. DOI: 10.1080/01621459.1971.10482356.

[4] Richard D. McKelvey and William Zavoina. "A statistical model for the analysis of ordinal level dependent variables". In: *Journal of Mathematical Sociology* 4.1 (Jan. 1975), pp. 103–120. ISSN: 0022-250X. DOI: 10.1080/0022250X.1975.9989847.

[5] 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.

[6] Gideon Schwarz. "Estimating the Dimension of a Model". In: *Annals of Statistics* 6.2 (Mar. 1978), pp. 461–464. ISSN: 0090-5364. DOI: 10.1214/aos/1176344136.

[7] Christopher Winship and Robert D Mare. "Regression models with ordinal variables". In: *American sociological review* (1984), pp. 512–525. DOI: 10.2307/2095465.

[8] 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.

[9] William E. Becker and Peter E. Kennedy. "A Graphical Exposition of the Ordered Probit". In: *Econometric Theory* 8.1 (Mar. 1992), pp. 127–131. ISSN: 1469-4360. DOI: 10.1017/S0266466600010781.

[10] Williams R Dillon, Thomas J Madden, and NH Firtle. "Marketing research in a marketing environment: Irwin". In: *Homewood, IL* (1994).

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

[12] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. New York, NY, USA: Springer, 2002. ISBN: 978-0-387-21706-2.

[13] Angela D'Elia and Domenico Piccolo. "A mixture model for preferences data analysis". In: *Computational Statistics and Data Analysis* 49.3 (June 2005), pp. 917–934. ISSN: 0167-9473. DOI: 10.1016/j.csda.2004.06.012.

[14] S. J. G. Lewis et al. “Heterogeneity of Parkinson’s disease in the early clinical stages using a data driven approach”. In: *Journal of Neurology, Neurosurgery, and Psychiatry* 76.3 (Mar. 2005), pp. 343–348. ISSN: 0022-3050. DOI: 10.1136/jnnp.2003.033530.

[15] Jeroen K. Vermunt and Jay Magidson. *Latnt GOLD 4.0 User’s Guide*. Belmont, Massachusetts, USA: Statistical Innovations Inc., 2005.

[16] David Arthur and Sergei Vassilvitskii. “k-means++: the advantages of careful seeding”. In: *SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. USA: Society for Industrial and Applied Mathematics, Jan. 2007, pp. 1027–1035. ISBN: 978-0-89871624-5. DOI: 10.5555/1283383.1283494.

[17] Scott M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY, USA: Springer, 2007. ISBN: 978-0-387-71265-9.

[18] Alan Agresti. *Analysis of Ordinal Categorical Data, 2nd Edition*. Wiley, Apr. 2010. ISBN: 978-0-470-08289-8.

[19] Paul D. McNicholas and T. Brendan Murphy. “Model-based clustering of longitudinal data”. In: *Canadian Journal of Statistics / La Revue Canadienne de Statistique* 38.1 (Mar. 2010), pp. 153–168. ISSN: 0319-5724. DOI: 10.1002/cjs.10047.

[20] Marco Giordan and Giancarlo Diana. “A Clustering Method for Categorical Ordinal Data”. In: *Communications in Statistics - Theory and Methods* 40.7 (Mar. 2011), pp. 1315–1334. ISSN: 0361-0926. DOI: 10.1080/03610920903581010.

[21] 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.

[22] 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.

[23] Cinzia Viroli. “On matrix-variate regression analysis”. In: *Journal of Multivariate Analysis* 111 (Oct. 2012), pp. 296–309. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2012.04.005.

[24] Damien McParland and Isobel Claire Gormley. “Clustering Ordinal Data via Latent Variable Models”. In: *Algorithms from and for Nature and Life*. Springer, July 2013, pp. 127–135. DOI: 10.1007/978-3-319-00035-0_12.

[25] Arnot Komrek and Lenka Komrkov. “Capabilities of R Package mixAK for Clustering Based on Multivariate Continuous and Discrete Longitudinal Data”. In: *Journal of Statistical Software* 59.12 (Sept. 2014), pp. 1–38. ISSN: 1548-7660. DOI: 10.18637/jss.v059.i12.

[26] Laura Anderlucci and Cinzia Viroli. “Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data”. In: *Annals of Applied Statistics* 9.2 (June 2015), pp. 777–800. ISSN: 1932-6157. DOI: 10.1214/15-AOAS816.

[27] John K. Kruschke. *Doing Bayesian Data Analysis*. Elsevier, Academic Press, 2015. ISBN: 978-0-12-405888-0.

[28] Christophe Biernacki and Julien Jacques. "Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm". In: *Statistics and Computing* 26.5 (Sept. 2016), pp. 929–943. ISSN: 0960-3174. DOI: 10.1007/s11222-015-9585-2.

[29] Fatma Zehra Doğru, Yakup Murat Bulut, and Olcay Arslan. "Finite mixtures of matrix variate t distributions". In: *Gazi University Journal of Science* 29.2 (2016), pp. 335–341.

[30] D. Fernandez, R. Arnold, and S. Pledger. "Mixture-based clustering for the ordered stereotype model". In: *Computational Statistics & Data Analysis* 93 (Jan. 2016), pp. 46–75. ISSN: 0167-9473. DOI: 10.1016/j.csda.2014.11.004.

[31] Maria Iannario and Domenico Piccolo. "A generalized framework for modelling ordinal data". In: *Statistical Methods & Applications* 25.2 (June 2016), pp. 163–189. ISSN: 1613-981X. DOI: 10.1007/s10260-015-0316-9.

[32] 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.

[33] Monia Ranalli and Roberto Rocci. "Mixture models for ordinal data: a pairwise likelihood approach". In: *Statistics and Computing* 26.1-2 (Jan. 2016), pp. 529–547. ISSN: 0960-3174. DOI: 10.1007/s11222-014-9543-4.

[34] Luca Scrucca et al. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". In: *The R Journal* 8.1 (2016), pp. 289–317. DOI: 10.32614/RJ-2016-021. URL: https://doi.org/10.32614/RJ-2016-021.

[35] Silvia Cagnone and Cinzia Viroli. "Multivariate Latent Variable Transition Models of Longitudinal Mixed Data: An Analysis on Alcohol Use Disorder". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 67.5 (Nov. 2018), pp. 1399–1418. ISSN: 0035-9254. DOI: 10.1111/rssc.12285.

[36] 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.

[37] Julien Jacques and Christophe Biernacki. "Model-based co-clustering for ordinal data". In: *Computational Statistics & Data Analysis* 123 (July 2018), pp. 101–115. ISSN: 0167-9473. DOI: 10.1016/j.csda.2018.01.014.

[38] Torrin M. Liddell and John K. Kruschke. "Analyzing ordinal data with metric models: What could possibly go wrong?" In: *J. Exp. Soc. Psychol.* 79 (Nov. 2018), pp. 328–348. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2018.08.009.

[39] 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.

[40] 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.

[41] Zvi Gilula et al. "A study into mechanisms of attitudinal scale conversion: A randomized stochastic ordering approach". In: *Quantitative Marketing and Economics* 17.3 (Sept. 2019), pp. 325–357. ISSN: 1573-711X. DOI: 10.1007/s11129-019-09209-3.

[42] 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.

[43] Margot Selosse, Julien Jacques, Christophe Biernacki, and Florence Cousson-Glie. "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.

[44] 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.

[45] Agns Franois-Lecompte et al. "Confinement et comportements alimentaires - Quelles volutions en matire d'alimentation durable ?" In: *Revue Franaise de Gestion* 46.293 (Nov. 2020), pp. 55–80. ISSN: 0338-4551. DOI: 10.3166/rfg.2020.00493.

[46] 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.

[47] Salvatore D. Tomarchio, Antonio Punzo, and Luca Bagnato. "Two new matrix-variate distributions with application in model-based clustering". In: *Computational Statistics & Data Analysis* 152 (Dec. 2020), p. 107050. ISSN: 0167-9473. DOI: 10.1016/j.csda.2020.107050.

[48] 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.

[49] Margot Selosse, Julien Jacques, and Christophe Biernacki. "ordinalClust: An R Package to Analyze Ordinal Data". In: *The R Journal* 12.2 (2021), pp. 173–188. DOI: 10.32614/RJ-2021-011. URL: https://doi.org/10.32614/RJ-2021-011.

[50] 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.

[51] Leonardo Salvatore Alaimo et al. "A Comparison of Migrant Integration Policies via Mixture of Matrix-Normals". In: *Social Indicators Research* 165.2 (Jan. 2023), pp. 473–494. ISSN: 1573-0921. DOI: 10.1007/s11205-022-03024-2.

[52] Jan Vvra and Arnot Komrek. "Classification based on multivariate mixed type longitudinal data with an application to the EU-SILC database". In: *Advances in Data Analysis and Classification* 17.2 (June 2023), pp. 369–406. ISSN: 1862-5355. DOI: 10.1007/s11634-022-00504-8.

Appendices

A Tables

Table 2: Clusters' means over time. The estimated parameter $\hat{\pi}$ = (0.37,0.44,0.19)

Cluster 1	T1	T2	T3	T4	T5
Q5(1)	3.99	4.16	4.20	4.22	4.17
Q5(2)	3.60	3.77	4.02	4.02	4.15
Q5(3)	3.89	4.22	4.19	4.42	4.35
Q5(4)	3.51	3.78	3.95	3.98	4.03
Q5(5)	3.32	3.64	3.86	4.14	4.03
Q8(1)	3.36	3.42	3.61	3.66	3.64
Q8(2)	4.06	4.17	4.08	4.10	4.03
Q8(3)	4.12	4.16	4.15	4.19	4.09
Q12(1)	4.30	4.53	3.73	4.10	4.23
Q12(2)	4.13	4.50	3.53	3.94	4.15
Q12(3)	4.38	4.41	3.67	4.10	4.02
Cluster 2	T1	T2	T3	**T4**	**T5**
Q5(1)	3.80	4.08	4.22	4.21	4.18
Q5(2)	3.72	3.79	4.10	4.07	4.13
Q5(3)	3.73	4.03	4.35	4.30	4.25
Q5(4)	3.49	3.78	3.99	3.97	3.98
Q5(5)	3.37	3.61	3.96	4.00	4.11
Q8(1)	3.30	3.49	3.70	3.70	3.57
Q8(2)	4.04	4.23	3.99	4.05	3.97
Q8(3)	4.05	4.27	4.07	4.14	4.08
Q12(1)	6.69	6.15	4.66	5.14	4.92
Q12(2)	6.69	6.38	4.49	5.56	5.18
Q12(3)	6.49	6.07	4.70	5.69	5.29
Cluster 3	**T1**	**T2**	**T3**	**T4**	**T5**
Q5(1)	4.27	5.04	4.92	4.59	4.85
Q5(2)	3.83	4.36	4.48	4.35	4.47
Q5(3)	4.49	5.43	5.16	5.23	5.28
Q5(4)	3.53	4.08	4.26	4.34	4.44
Q5(5)	3.69	3.78	4.21	4.30	4.39
Q8(1)	2.15	2.26	2.55	3.03	2.74
Q8(2)	4.12	4.00	4.06	4.15	4.11
Q8(3)	4.35	4.76	4.53	4.49	4.64
Q12(1)	7.20	7.10	6.36	6.76	6.59

Q12(2)	7.22	6.93	6.08	6.61	6.65
Q12(3)	7.32	6.72	6.04	6.48	6.24

Table 3: Clusters' time correlation

Cluster 1	T1	T2	T3	T4	T5
T1	1.00	0.28	0.18	0.16	0.09
T2	0.28	1.00	0.23	0.17	0.25
T3	0.18	0.23	1.00	0.25	0.27
T4	0.16	0.17	0.25	1.00	0.25
T5	0.09	0.25	0.27	0.25	1.00
Cluster 2	T1	T2	T3	T4	T5
T1	1.00	0.25	0.09	0.11	0.12
T2	0.25	1.00	0.21	0.19	0.11
T3	0.09	0.21	1.00	0.21	0.16
T4	0.11	0.19	0.21	1.00	0.17
T5	0.12	0.11	0.16	0.17	1.00
Cluster 3	T1	T2	T3	T4	T5
T1	1.00	0.20	0.19	0.09	0.09
T2	0.20	1.00	0.25	0.12	0.17
T3	0.19	0.25	1.00	0.19	0.23
T4	0.09	0.12	0.19	1.00	0.33
T5	0.09	0.17	0.23	0.33	1.00

Table 4: Clusters' variables correlation

Cluster 1	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)
Q5(1)	1.00	0.23	0.45	0.18	0.15	-0.10	0.01	0.16	0.15	0.08	0.07
Q5(2)	0.23	1.00	0.31	0.47	0.34	0.01	0.06	0.05	0.11	0.10	0.01
Q5(3)	0.45	0.31	1.00	0.29	0.22	-0.11	0.02	0.16	0.16	0.09	0.09
Q5(4)	0.18	0.47	0.29	1.00	0.32	0.03	0.07	-0.00	0.08	0.07	0.02
Q5(5)	0.15	0.34	0.22	0.32	1.00	-0.01	0.04	-0.00	0.01	0.06	-0.02
Q8(1)	-0.10	0.01	-0.11	0.03	-0.01	1.00	-0.05	-0.17	-0.09	-0.07	-0.05
Q8(2)	0.01	0.06	0.02	0.07	0.04	-0.05	1.00	0.19	0.07	0.09	0.04
Q8(3)	0.16	0.05	0.16	-0.00	-0.00	-0.17	0.19	1.00	0.09	0.05	0.07
Q12(1)	0.15	0.11	0.16	0.08	0.01	-0.09	0.07	0.09	1.00	0.58	0.50
Q12(2)	0.08	0.10	0.09	0.07	0.06	-0.07	0.09	0.05	0.58	1.00	0.48
Q12(3)	0.07	0.01	0.09	0.02	-0.02	-0.05	0.04	0.07	0.50	0.48	1.00
Cluster 2	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)

Q5(1)	1.00	0.24	0.43	0.22	0.24	-0.05	0.06	0.02	0.04	0.03	-0.02
Q5(2)	0.24	1.00	0.35	0.41	0.33	-0.08	-0.05	0.01	-0.01	-0.03	-0.01
Q5(3)	0.43	0.35	1.00	0.33	0.31	-0.05	-0.02	-0.02	-0.01	0.02	-0.02
Q5(4)	0.22	0.41	0.33	1.00	0.37	-0.02	0.00	0.04	-0.04	-0.03	-0.03
Q5(5)	0.24	0.33	0.31	0.37	1.00	-0.02	-0.00	-0.00	-0.02	-0.04	-0.07
Q8(1)	-0.05	-0.08	-0.05	-0.02	-0.02	1.00	0.02	-0.09	-0.06	-0.01	0.00
Q8(2)	0.06	-0.05	-0.02	0.00	-0.00	0.02	1.00	0.13	-0.02	0.01	-0.08
Q8(3)	0.02	0.01	-0.02	0.04	-0.00	-0.09	0.13	1.00	0.03	-0.03	0.02
Q12(1)	0.04	-0.01	-0.01	-0.04	-0.02	-0.06	-0.02	0.03	1.00	0.48	0.37
Q12(2)	0.03	-0.03	0.02	-0.03	-0.04	-0.01	0.01	-0.03	0.48	1.00	0.42
Q12(3)	-0.02	-0.01	-0.02	-0.03	-0.07	0.00	-0.08	0.02	0.37	0.42	1.00
Cluster 3	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)
Q5(1)	1.00	0.32	0.44	0.16	0.18	-0.15	-0.00	0.13	0.02	0.09	-0.05
Q5(2)	0.32	1.00	0.38	0.34	0.20	-0.00	-0.12	-0.01	-0.06	0.02	-0.05
Q5(3)	0.44	0.38	1.00	0.29	0.16	-0.12	0.01	0.10	0.01	0.11	-0.17
Q5(4)	0.16	0.34	0.29	1.00	0.25	-0.01	-0.08	0.07	0.06	0.03	-0.06
Q5(5)	0.18	0.20	0.16	0.25	1.00	-0.02	0.02	0.01	0.06	0.08	0.05
Q8(1)	-0.15	-0.00	-0.12	-0.01	-0.02	1.00	0.00	-0.19	-0.08	-0.08	-0.01
Q8(2)	-0.00	-0.12	0.01	-0.08	0.02	0.00	1.00	0.08	0.02	-0.02	-0.02
Q8(3)	0.13	-0.01	0.10	0.07	0.01	-0.19	0.08	1.00	0.07	0.07	0.01
Q12(1)	0.02	-0.06	0.01	0.06	0.06	-0.08	0.02	0.07	1.00	0.44	0.26
Q12(2)	0.09	0.02	0.11	0.03	0.08	-0.08	-0.02	0.07	0.44	1.00	0.21
Q12(3)	-0.05	-0.05	-0.17	-0.06	0.05	-0.01	-0.02	0.01	0.26	0.21	1.00

Table 5: Clusters' time covariances

Cluster 1	T1	T2	T3	T4	T5
T1	1.34	0.36	0.20	0.17	0.09
T2	0.36	1.25	0.25	0.18	0.26
T3	0.20	0.25	0.88	0.21	0.23
T4	0.17	0.18	0.21	0.87	0.22
T5	0.09	0.26	0.23	0.22	0.85
Cluster 2	T1	T2	T3	T4	T5
T1	1.17	0.30	0.09	0.10	0.12
T2	0.30	1.22	0.21	0.19	0.11
T3	0.09	0.21	0.84	0.17	0.13
T4	0.10	0.19	0.17	0.81	0.14
T5	0.12	0.11	0.13	0.14	0.86
Cluster 3	T1	T2	T3	T4	T5

T1	1.50	0.32	0.27	0.13	0.14
T2	0.32	1.78	0.38	0.19	0.28
T3	0.27	0.38	1.33	0.26	0.33
T4	0.13	0.19	0.26	1.42	0.49
T5	0.14	0.28	0.33	0.49	1.48

Table 6: Clusters' variables covariances

Cluster 1	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)
Q5(1)	0.58	0.14	0.31	0.10	0.10	-0.06	0.00	0.08	0.15	0.08	0.07
Q5(2)	0.14	0.62	0.22	0.28	0.25	0.01	0.03	0.03	0.11	0.10	0.01
Q5(3)	0.31	0.22	0.84	0.20	0.18	-0.08	0.01	0.09	0.19	0.11	0.12
Q5(4)	0.10	0.28	0.20	0.56	0.22	0.02	0.03	-0.00	0.08	0.07	0.02
Q5(5)	0.10	0.25	0.18	0.22	0.83	-0.01	0.02	-0.00	0.01	0.08	-0.02
Q8(1)	-0.06	0.01	-0.08	0.02	-0.01	0.62	-0.03	-0.09	-0.09	-0.08	-0.06
Q8(2)	0.00	0.03	0.01	0.03	0.02	-0.03	0.45	0.09	0.06	0.08	0.03
Q8(3)	0.08	0.03	0.09	-0.00	-0.00	-0.09	0.09	0.43	0.08	0.05	0.07
Q12(1)	0.15	0.11	0.19	0.08	0.01	-0.09	0.06	0.08	1.66	1.01	0.91
Q12(2)	0.08	0.10	0.11	0.07	0.08	-0.08	0.08	0.05	1.01	1.79	0.91
Q12(3)	0.07	0.01	0.12	0.02	-0.02	-0.06	0.03	0.07	0.91	0.91	2.00
Cluster 2	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)
Q5(1)	0.55	0.13	0.30	0.12	0.16	-0.03	0.03	0.01	0.04	0.03	-0.02
Q5(2)	0.13	0.58	0.25	0.23	0.23	-0.05	-0.02	0.01	-0.01	-0.02	-0.01
Q5(3)	0.30	0.25	0.89	0.23	0.27	-0.04	-0.01	-0.01	-0.01	0.03	-0.03
Q5(4)	0.12	0.23	0.23	0.56	0.25	-0.01	0.00	0.02	-0.03	-0.03	-0.03
Q5(5)	0.16	0.23	0.27	0.25	0.83	-0.01	-0.00	-0.00	-0.02	-0.04	-0.09
Q8(1)	-0.03	-0.05	-0.04	-0.01	-0.01	0.79	0.01	-0.05	-0.07	-0.01	0.00
Q8(2)	0.03	-0.02	-0.01	0.00	-0.00	0.01	0.42	0.06	-0.02	0.01	-0.07
Q8(3)	0.01	0.01	-0.01	0.02	-0.00	-0.05	0.06	0.44	0.02	-0.03	0.02
Q12(1)	0.04	-0.01	-0.01	-0.03	-0.02	-0.07	-0.02	0.02	1.52	0.73	0.63
Q12(2)	0.03	-0.02	0.03	-0.03	-0.04	-0.01	0.01	-0.03	0.73	1.53	0.71
Q12(3)	-0.02	-0.01	-0.03	-0.03	-0.09	0.00	-0.07	0.02	0.63	0.71	1.87
Cluster 3	Q5(1)	Q5(2)	Q5(3)	Q5(4)	Q5(5)	Q8(1)	Q8(2)	Q8(3)	Q12(1)	Q12(2)	Q12(3)
Q5(1)	0.90	0.26	0.42	0.13	0.16	-0.14	-0.00	0.12	0.02	0.08	-0.05
Q5(2)	0.26	0.74	0.33	0.24	0.16	-0.00	-0.09	-0.00	-0.05	0.02	-0.05
Q5(3)	0.42	0.33	1.01	0.24	0.15	-0.12	0.01	0.10	0.01	0.11	-0.19
Q5(4)	0.13	0.24	0.24	0.68	0.19	-0.00	-0.06	0.06	0.05	0.03	-0.06
Q5(5)	0.16	0.16	0.15	0.19	0.84	-0.02	0.01	0.01	0.06	0.08	0.05
Q8(1)	-0.14	-0.00	-0.12	-0.00	-0.02	1.00	0.00	-0.18	-0.08	-0.08	-0.01

Q8(2)	-0.00	-0.09	0.01	-0.06	0.01	0.00	0.87	0.08	0.02	-0.02	-0.02
Q8(3)	0.12	-0.00	0.10	0.06	0.01	-0.18	0.08	0.91	0.07	0.06	0.01
Q12(1)	0.02	-0.05	0.01	0.05	0.06	-0.08	0.02	0.07	1.05	0.46	0.30
Q12(2)	0.08	0.02	0.11	0.03	0.08	-0.08	-0.02	0.06	0.46	1.00	0.24
Q12(3)	-0.05	-0.05	-0.19	-0.06	0.05	-0.01	-0.02	0.01	0.30	0.24	1.26

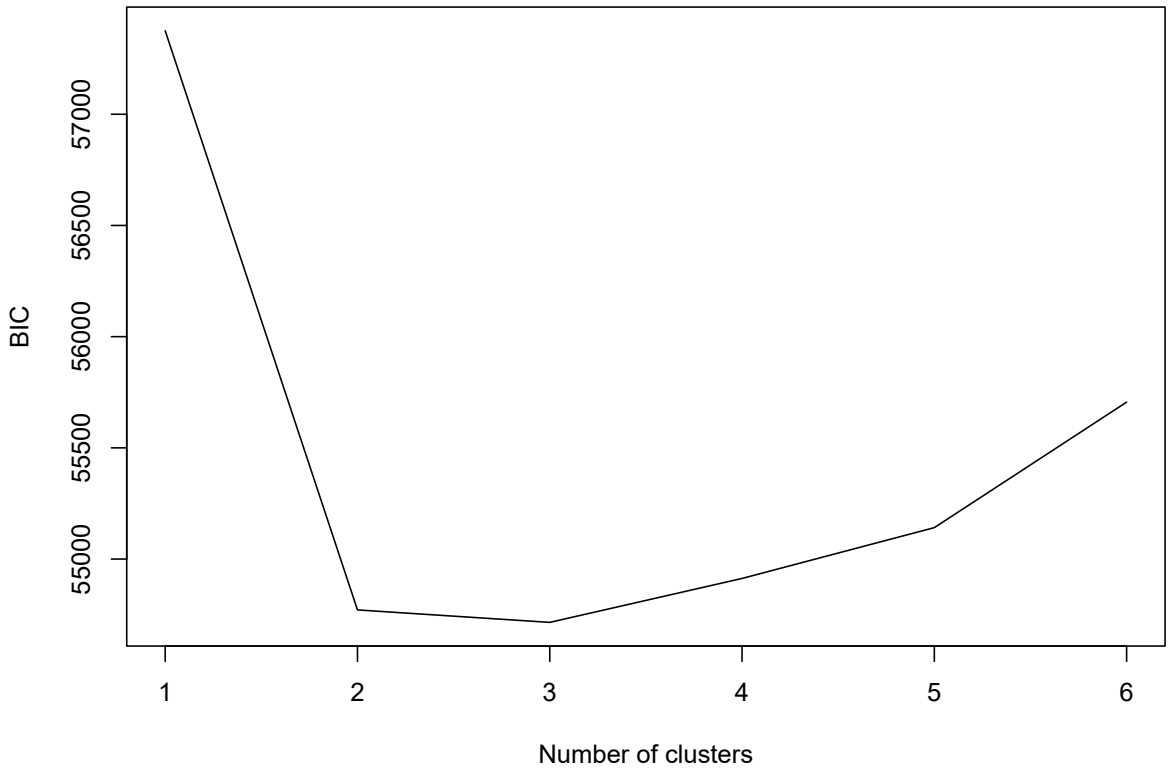


Figure 10: Visualization of BIC for K as results of application on real data. Kmeans++ initialization.