# Mixed data temporal clustering for modelling longitudinal surveys

Francesco Amato

# Mixed data temporal clustering for modelling longitudinal surveys


Auteur:

**Francesco AMATO**

*Établissement d'inscription* : Université Lumière Lyon 2
*Laboratoire d'accueil* : Laboratoire ERIC


## Composition du Jury:

| | | |
|---|---|---|
| Christophe BIERNACKI | Professeur, Université de Lille | Président |
| Pierre LATOUCHE | Professeur, Université Clermont Auvergne | Rapporteur |
| Claire GORMLEY | Full Professor, University College Dublin | Rapporteuse |
| Cécile PROUST-LIMA | Directrice de Recherche, INSERM | Examinatrice |
| Pierre VALETTE-FLORENCE | Professeur Émérite, Université Grenoble Alpes | Examinateur |
| Stéphane CHRÉTIEN | Professeur, Université Lumière Lyon 2 | Examinateur |
| Julien JACQUES | Professeur, Université Lumière Lyon 2 | Directeur de thèse |
| Isabelle PRIM-ALLAZ | Professeure, Université Lumière Lyon 2 | Co-directrice de thèse |


En vue de l'obtention du titre de Docteur de l'Université Lumière Lyon 2
*Discipline* : Mathématiques appliquées

**Date de soutenance: 19/05/2025**

*"On s'engage et puis... on voit!"*

- Napoléon Bonaparte, on his approach to battle.

*Mixed data temporal clustering for modelling longitudinal surveys*

Author:
Francesco AMATO (*Établissement d'inscription* : Université Lumière Lyon 2
*Laboratoire d'accueil* : Laboratoire ERIC)

Supervisors:
Julien JACQUES
Isabelle PRIM-ALLAZ

# Abstract

In the realm of social sciences, humanities, and medical research, longitudinal surveys are pivotal for understanding temporal dynamics and behavioural patterns within populations. These surveys often collect mixed-type data—comprising nominal, ordinal, quantitative, and textual responses—which pose significant challenges for statistical analysis. The analysis of such mixed data is a current research problem in the fields of statistics and machine learning, and for ignorance or deficit of existing solutions adapted to their cases, practitioners often tend to transform the data (particularly categorical ordinal and count ones) in order to treat them as one continuous, since they are easier to handle. Such approach is not satisfying since it leads either to the introduction of a bias or to an important information loss. Moreover, it is often the case for these questionnaires to be completed by participants several times over a study period. These data are scientifically known as "longitudinal" data. Researchers then analyse these questionnaires, being especially interested in time evolution of common behaviours. 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.

The scientific context of this work is rooted in the growing need for statistical methods to analyse complex, temporally evolving datasets. Specifically, it is often the case for researchers to be interested in finding "typical" patterns, that is to find group of units exposing similar behaviour. This is called "clustering". Existing methods either fail to capture the intricate relationships between variables and the temporal evolution of clusters, leading to suboptimal results or produce results that are difficult to interpret by non-statisticians.

The contributions of this thesis are multifaceted. Firstly, it provides a survey over the existing frameworks for clustering longitudinal, mixed-type and longitudinal mixed-type data. Then, we gradually approach the problem by starting with longitudinal ordinal data, to finally present the Mixture of Mixed-Matrices (MMM) model: reorganizing the data in a three-way structure and assuming that the non-continuous variables are observations of underlying latent continuous variables, the model relies on a mixture of matrix-variate normal distributions to perform clustering in the latent dimension. The MMM model is thus able to handle continuous, ordinal, binary, nominal and count data and to concurrently model the heterogeneity, the association among the responses and the temporal dependence structure in a parsimonious way and without assuming conditional independence. We finally propose some future directions to overcome the limits of our proposed model.

# Résumé

Dans le domaine des sciences sociales, des sciences humaines et de la recherche médicale, les enquêtes longitudinales sont essentielles pour comprendre les dynamiques temporelles et les schémas comportementaux au sein des populations. Ces enquêtes collectent souvent des données de types mixtes (comprenant des réponses nominales, ordinales, quantitatives et textuelles) qui posent des défis significatifs pour l'analyse statistique. L'analyse de telles données mixtes est un problème de recherche actuel dans les domaines de la statistique et de l'apprentissage automatique. Par ignorance ou manque de solutions existantes adaptées à leurs cas, les praticiens tendent souvent à transformer les données (notamment les données ordinales catégorielles et les comptages) pour les traiter comme des données continues, car elles sont plus faciles à manipuler. Cette approche n'est pas satisfaisante car elle conduit soit à l'introduction d'un biais, soit à une perte importante d'information. De plus, il est souvent demandé aux participants de remplir ces questionnaires plusieurs fois au cours d'une période d'étude. Ces données sont scientifiquement connues sous le nom de données "longitudinales". Les chercheurs analysent ensuite ces questionnaires, s'intéressant particulièrement à l'évolution temporelle des comportements communs. Cependant, modéliser l'évolution temporelle est loin d'être trivial. L'approche la plus basique consiste à effectuer des analyses indépendamment à chaque phase temporelle, puis à essayer *a posteriori* de trouver des liens entre ces différentes analyses, en cherchant d'une phase à l'autre des comportements typiques similaires ou différents.
Le contexte scientifique de ce travail est ancré dans le besoin croissant de méthodes statistiques pour analyser des ensembles de données complexes et évolutives dans le temps. Plus précisément, les chercheurs s'intéressent souvent à la recherche de "schémas typiques", c'est-à-dire à la découverte de groupes d'unités présentant des comportements similaires. Cela s'appelle le "clustering". Les méthodes existantes échouent soit à capturer les relations complexes entre les variables et l'évolution temporelle des clusters, conduisant à des résultats sous-optimaux, soit produisent des résultats difficiles à interpréter par des non-statisticiens. Les contributions de cette thèse sont multiples. Tout d'abord, elle fournit une revue des cadres existants pour le clustering de données longitudinales, de types mixtes et longitudinales de types mixtes. Ensuite, nous abordons progressivement le problème en commençant par les données ordinales longitudinales, pour finalement présenter le modèle Mixture of Mixed-Matrices (MMM) : en réorganisant les données dans une structure tridimensionnelle et en supposant que les variables non continues sont des observations de variables latentes continues sous-jacentes, le modèle repose sur un mélange de distributions normales matricielles pour effectuer le clustering dans la dimension latente. Le modèle MMM est ainsi capable de gérer les données continues, ordinales, binaires, nominales et de comptage et de modéliser simultanément l'hétérogénéité, l'association entre les réponses et la structure de dépendance temporelle de manière parcimonieuse et sans supposer l'indépendance conditionnelle. Enfin, nous proposons quelques directions futures pour surmonter les limites de notre modèle proposé.

# Acknowledgements

Despite being at the top of the manuscript, acknowledgments are usually the last part to be written, and maybe the most rewording one: not only because it means that the struggle of writing has ended, but also because it allows to remember all the people met during the journey. Personally, I consider myself lucky to have many to thank. It will be long, but bear with me.

This journey started when I was in Lisbon, Portugal for my second Erasmus from the University of Bologna, Italy and finished in Saint-Étienne, France, where I am writing these words, via Lyon and Paris. It started in English and it ends in French. It started when I was single and it ends when I am married. What a ride these three and a half years have been.

First, I need to thank Cinzia, who mentored me during my bachelor's and master's degree, suggested to apply to this PhD program, and also wrote the recommendation letter for me. Her teachings, both in and out of the classroom, have been valuable. A special thanks also goes to Brendan, who happened to be in Lyon during my first year, and generously helped me navigate the very beginning of my thesis like a deputy supervisor, despite being super busy and having no obligation towards me.

Additionally, I had the chance of being supervised by two terrific people, both very busy but always available to guide me when I needed it. The distance, both physical, due to our labs locations, and academical, due to our study subjects, never prevented Isabelle to meaningfully advise me on the more practical aspects of the thesis, allowing me to not get lost in the realm of theory, reminding me of the real-world challenges we statisticians need to have in mind when serving our role, that is to be an ancillary science. An even bigger acknowledgement needs to be given to Julien, who had to put up with me much more. I would enter his office with a mixture of anxiety and worry because I was stuck, or made a mistake, or both, but he would always seat and patiently reason with me on how to tackle the issue. He taught me to be pragmatic, to address problems from different angles, to reason on the deeper level of what I am doing and most importantly that a thesis, like research (and maybe life), "is not a linear journey". *Merci*.

Furthermore, I would like to extend my heartfelt thanks to the members of my defense committee for their time and insightful feedback. Your thoughtful questions and constructive comments during the defense greatly enriched the final version of this thesis. I am sincerely grateful for your engagement and for the honor of having you as part of this important milestone in my academic journey.

Next, I would like to express my gratitude to the members of my *comité de suivi*, Jairo and Denys, for their constructive observations and their caring, which have contributed to the quality of this work and to my personal growth. I am also grateful to the permanent members of the lab, who directly or indirectly supported me during this adventure, particularly regarding my teaching experience. I want to acknowledge Guillaume, who for kindness and proximity in age has felt like an older brother, always willing to help out. He sets out the example of the kind of young researcher and teacher I aspire to be. Gratitude goes to Stéphane, Julien, Sabine and Jérôme as well.

Much appreciation goes also to the doctoral colleagues of the ERIC lab, with whom I shared the difficulties, the issues but also the happy moments of daily life. Notably to Jean Steve and Eliz, who started this adventure at the same time and shared with me the same supervisor, but also to Martial, Enzo, Gaël, Irina, Noé, Simon, Rémi and all the others. *Ça a été un plaisir, les gars!*

I want to show my appreciation also to my friends and comrades, old and new, who one way or another, encouraged me to take this road, sustained me and my choices. Thanks to my historical friends from my home-town and my high school, Antonio, Andrea, Andrea, Shady, Marco, Marco and Lorenzo. Even if we are apart, *vi voglio bene, amici miei!* The friends met along my studies, some of whom share the same difficulties as doctoral students as well: Giulia, Matteo, Silvia, Federica, Noemi and everyone else. *Grazie.* A thank you is due to my favourite (acquired) *stéphanois*, Thibaud. *Merci.* And I cannot forget to mention my Italian-Brazilian mates: Agnese, Isabella, Kaique and the little Ikki. *Obrigado.* Lastly, I am appreciative of the chance I had to meet two remarkable Italian colleagues and friends, who happened to be in Lyon for their exchange period and shared with me the start and the end of my PhD, Emiliano and Matteo. It was nice to feel home closer with them.

Last but absolutely not least, the warmest thanks goes to my family. To my mother, father and sister who have put up with me for literally all my life, encouraged me, sustained me, taught me, helped me and loved me. I owe you everything. *Grazie di tutto, davvero.* Finally, my most heartfelt appreciation goes to my husband, Luiz, who after just few months decided to follow me in France, and has sustained with me all the steps, the difficulties, the joyful moments, the worries and the happiness. He took care of me, made my life happier, loved me and ultimately made me a better person. I can never thank him enough. *Ti amo tanto, Fofo!.*

In conclusion, since I am legally obliged but also because it allowed me to pay the bills during this time, 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).

# Contents

# List of Symbols

| | |
|---|---|
| $y_{ijt}$ | Observation of the $j$-th variable for the $i$-th unit at time $t$. |
| $J$ | Number of variables. |
| $N$ | Number of observed units. |
| $T$ | Number of time points. |
| $J_d$ | Number of features of the $d$-th data type. |
| $\ell_i$ | Cluster allocation variable for the $i$-th unit. |
| $\pi_k$ | Probability of belonging to cluster $k$, where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. |
| $x, \xi$ | Covariates affecting the outcome or cluster allocation probability. |
| $\theta$ | Parameters associated with models. |
| $\Sigma_k$ | Covariance matrix parameterization. |
| $\Gamma_k, \Delta_k$ | Matrices associated with eigenvalues and orientation. |
| $\Phi_k$ | Time covariance matrix. |
| $M_k$ | Mean matrix. |
| $\mathcal{N}_J(\mu, \Sigma)$ | Multivariate normal distribution of dimension $J$ with mean $\mu$ and covariance $\Sigma$. |
| $\mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$ | Matrix-normal distribution notation. |
| $f(y\|\theta)$ | Generic density function. |

**Statistical Functions and Operators:**

| | |
|---|---|
| $\mathbb{E}(X)$ | Expectation of random variable $X$. |
| $\mathbb{V}(X)$ | Variance of $X$. |
| $\text{Cov}(X, Y)$ | Covariance between variables $X$ and $Y$. |
| $\mathbb{P}(X)$ | Generic probability notation: density function $f(x)$ for continuous variables, $P(X = x)$ for discrete ones. |
| $\mathcal{L}(\theta; y)$ | Likelihood function. |
| $\log \mathcal{L}(\theta; y), l(\theta; y)$ | Log-likelihood function. |
| $\hat{\theta}$ | Estimated parameter values. |
| $\mathbf{1}_A(x)$ | Indicator function for event $A$. |
| $\int f(x)dx$ | Integral of function $f(x)$. |
| $\sum_{i=1}^{n} X_i$ | Summation notation. |
| $\max(\cdot)$ | Maximum function. |
| $\arg\max_x f(x)$ | Value of $x$ that maximizes $f(x)$. |

# Chapter 1

# Introduction

## 1.1 About me

I am lucky enough to remember the day I got interested in statistics. In Italy, passing a final exams is mandatory to graduate high-school. The exam is composed by different tests, and while they can vary based on the kind of school students attend, one test is equal nation- and schools-wide: a six hours-long session during which candidates must write an essay on one of the several proposed topics. Among the ones indicated for the year of my graduation, one stood out and was chosen by a younger me. The title was "Growth, development and social progress. Is GDP the measure of everything?"[1]. I have been interested in economics, history and politics since I can remember and even more so after the 2011 financial crisis that hurt my country and family. However, that was the first time my focus was on the measurement of phenomena, and how to analyse and contextualise those measures. In my essay, I argued that using GDP as a measure of general development was not a good approach, since GDP was created to measure something else, that is the economic output, and while it could be used as a proxy for some other phenomena we wish to assess, other types of measurements were needed if one really wanted to get a grasp on the general level of development, social progress and well-being of a country. This happened on the 22nd of June 2016, and three weeks later I was applying to join the international statistical sciences program of the University of Bologna. On the 4th of August of the same year, a law was enacted amending the approval process of the Italian State budget, introducing for the first time the measurement of some indicators of "equitable and sustainable well-being"[2] as complementary measures to GDP, and against which compare and evaluate the effects of public policies. Someone must have read my essay.
While my academic path with social sciences somehow parted, I never stopped being interested in them, and this thesis gave me the opportunity to work on a piece of statistics applicable to them.

[1] The curious reader can find here the complete text of the exam and the list of topics to choose from (in Italian). The so called "socio-economic" topic is under Type B, Topic n.2.

[2] Here is the text of the law n.163/2016 (in Italian). The introduction of the described indicators is in Art.14.

## 1.2 Scientific context and motivating question

In many areas of humanities, social sciences and medical sciences, studies are often based on questionnaires. The researchers then analyse these questionnaires to determine typical motivations or intention and to measure typical behaviours within the studied population. But the statistical analysis of these questionnaires is far from simple, for several reasons. First, the answers to the questions are frequently of different types: nominal categorical (for example "what is your socio-professional category?"), ordinal categorical (for example "what is your level of satisfaction: bad, average, good?"), quantitative ("what is your age?"), count ("how many times did you go shopping this month"), textual (for open questions with free answer). These data are scientifically known as "mixed-type" or "mixed" data. The analysis of such mixed data is a current research problem in the fields of statistics and machine learning, and for ignorance or deficit of existing solutions adapted to their cases, practitioners often tend to transform the data (particularly categorical ordinal and count ones) in order to treat them as one continuous, since they are easier to handle. Such approach is not satisfying since it leads either to the introduction of a bias or to an important information loss.

Moreover, it is often the case for these questionnaires to be completed by participants several times over a study period. These data are scientifically known as "longitudinal" data. Researchers then analyse these questionnaires, being especially interested in time evolution of common behaviours. 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-Gélie, 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. Thus, the analysis will exhibit typical temporal evolution behaviours, which are the objects which researchers in human and social sciences wish to study.

When these two scientific challenges are put together, the scientific literature becomes even thinner, as researchers would need a tool that is able to handle longitudinal mixed-type data, as the ones collected in François-Lecompte, Innocent, Kréziak, and Prim-Allaz, 2020.
The scarcity of adapted methods is particularly pronounced when it comes to clustering, that is the task of grouping a set of data units in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups. For example, reducing a large set of individuals into a limited number of groups of individuals exhibiting the same behaviour, therefore extracting typical behaviours, which are generally the objects which researchers in human and social sciences wish to study.
The aim of this thesis is thus to provide a new algorithm for clustering longitudinal mixed-type data. The core of the thesis will be the development of a statistical model, associated inference algorithms and tools to perform this task, with in mind researchers and practitioners with no main statistical background as the final users.

## 1.3 What is clustering

As specified in the previous section, cluster analysis is an ensemble of methods dedicated to finding groups in a set of objects characterized by certain measurements, and it aims at finding these groups

in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. This task has a very wide range of applications such, for instance: biology, textual analysis, economy, or sociology. Indeed, an early and very well known example of clustering is the biological classification system introduced by Linnaeus in 1735 (Bouveyron, Celeux, Murphy, and Raftery, 2019a). Linnaeus categorized plants and animals based on subjective criteria, such as the number of stamens in flowers. However, modern cluster analysis uses automatic methods to group quantitative data. More formally, clustering can be defined as the process of partitioning a dataset into $K$ clusters, where $K$ is a predefined number of groups. The objective is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. In other words, the aim is to ensure that data points within the same cluster are as similar as possible, while data points in different clusters are as dissimilar as possible. Since the 1930s, various algorithms have been developed to partition data into groups with similar characteristics (Bouveyron, Celeux, Murphy, and Raftery, 2019a).
In most applications, researchers do not actually know the true nature, number or composition of these groups, they do not have a "ground truth" to be used either in the inference process or in the post-inference validation. For this reason, clustering is said to belong to the group of "unsupervised learning" methods.

For readers who may not be familiar with this terminology, in the context of machine learning and statistics, it is customary to divide algorithms into three broad classes, depending on the paradigm they employ to "learn"[3]:

- **Supervised learning** paradigm: the model is trained on a labelled dataset. This means that each training example is paired with an output label. The goal of supervised learning is to learn a mapping from input data to output labels so that the model can accurately predict the labels of new, unseen data.
- **Unsupervised learning** paradigm: it involves training a model on data without labelled responses. The goal is to find hidden patterns or intrinsic structures in the input data. Unlike supervised learning, unsupervised learning does not have a predefined output; instead, it aims to discover the underlying structure of the data.
- **Semi-supervised learning** paradigm: it combines a certain proportion of labelled data with a complementary proportion of unlabelled data during training. This approach aims at leveraging the strengths of both supervised and unsupervised learning. The main goal is to improve learning accuracy by using the unlabelled data to enhance the understanding of the labelled data, which can be particularly useful when labelled data is scarce or expensive to obtain.

The unsupervised nature of clustering makes it as much an art as a science (Von Luxburg, Williamson, and Guyon, 2012), as different clustering techniques can prioritize different data characteristics and also endow clusters with different interpretations. For instance, in Hennig, 2015 the author explores the concept of "true clusters" and emphases that the definition of a "true" or "real" cluster depends on the context and the specific goals of clustering. The author argues that there is no universally accepted definition of clusters, as different applications require different clustering

[3]Meaning, the approach they employ to find the parameters they need to compute. This is called "learning" in the context of machine learning, and the process of learning is called "training", while in statistics it is more common to call it "inference".

characteristics, as the usefulness and meaningfulness of the grouping is based on the specificity of each context.

## 1.4 Different types of clustering

The absence of a "true" label to be predicted led to the development of various types of clustering algorithms (Hennig, Meila, Murtagh, and Rocci, 2015), each focusing on a particular philosophical nature of clusters. Among others, we can cite:

- **Centroid-based clustering**: clusters are formed based on the distance between data points, and each cluster is represented by a central point or vector, which is not necessarily a member of the dataset. Clusters are viewed as groups of points close in the measurement space.
- **Hierarchical clustering**: clusters are formed based on the similarity between data points, but each cluster is assumed to be part of a bigger cluster so to build a hierarchy of clusters. These algorithms are also known as "connectivity-based clustering". Clusters are seen as part of extensive hierarchy of clusters that merge with each other at certain distances.
- **Density-based clustering**: clusters are defined as dense regions of data points separated by regions of lower density. Data points in sparse areas are usually considered to be noise and border points. Clusters are seen as areas of dense data points presence.
- **Model-based clustering**: data are seen as generated from a mixture of probability distributions, and the components of this mixture are typically interpreted as clusters. Therefore, clusters are seen as probability distributions and endowed with all the statistical properties that this entails.

Each of the methods above not only defines clusters differently, but also implies different ways to deal with the main questions and challenges that arise when performing clustering, such as the choice of the number of clusters, the assessment of uncertainty of the grouping and its robustness to outliers and noise. From our perspective, model-based (or probabilistic) clustering methods have the ability to leverage the statistical properties of probability distributions and the inferential framework of statistics to answer to these practical questions, and the advantage of clearly stating the assumptions behind the clustering algorithm. This is why in the following chapters of this thesis we will present and work within the framework of probabilistic clustering, which is detailed more formally in Chapter 2.

## 1.5 Content of the thesis

To introduce the less specialist reader to the argument of the thesis, we provide a description of model-based clustering in Chapter 2, where we present its general framework, its wider expression in the form of finite mixture models, the main algorithm used to perform inference on this mode, the EM algorithm, and finally its more common declination, that is the Gaussian mixture model. We will also approach the possible parametrizations of the latter and how to select the appropriate number of clusters. In Chapter 3 we provide a survey on the existing literature building up starting from the literature on model-based clustering for mixed-type cross-sectional data, to then tackle the methods existing in the context of longitudinal data, and finally address the literature for

longitudinal mixed-type data. The chapter is also meant to be read for practical purposes thanks to the resume of existing libraries available in R (R Core Team, 2023) to fit the described models. Then, the main works of the thesis will follow in the next two chapters. In Chapter 4 the first step of our journey towards clustering longitudinal mixed-type data is presented, that is a model to cluster longitudinal ordinal data. We started with this kind of data because categorical ordinal data are usually the most common type in questionnaires as we described in Section 1.2. The chapter is the reproduction of the work published under Amato, Jacques, and Prim-Allaz, 2024. In there, we lay the basis of our model, that is the use of matrix-variate distributions and the assumption that an ordinal variable is the discretization of a underlying latent continuous variable. This way, the model we developed is able to account simultaneously for within- and between-time dependence structures and to concurrently model the heterogeneity, the association among the responses and the temporal correlation structure. In Chapter 5 we detail our model to address longitudinal mixed-type data. The work in this chapter has been submitted for publication and is currently under review. We extend the model proposed in the previous chapter to account simultaneously for continuos, ordinal, binary, categorical and count data, in a way that does not require the conditional independence assumption and can fully account for correlation structure among different types of data. Finally, in Chapter 6, we will draw some conclusions about the developed model in the context of longitudinal mixed-type data analysis, we will outline some limitations and possible fixes and future perspective. Moreover, in Appendix B we present a work done during the first year of the thesis in collaboration with an Italian research team, as it is not strictly related to the topic of the thesis. The work in question was important for us to prove the goodness of the Gaussian matrix-variate mixture model in an application on longitudinal continuous data in a social science context. This work has been published under Alaimo et al., 2023. Specifically, it presents a clustering application on the Migrant Integration Policy Index. We show that the model is suitable for longitudinal data, allowing for the identification of clusters of countries with similar patterns of migrant integration policies over time. The analysis identify 5 clusters of countries, with the aim to facilitate the evaluation and the comparison of the countries within each cluster and between different clusters over time.

## 1.6 List of Publications and Communications

**Publications:**

- Leonardo 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
- 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

**Under review:**

- Francesco Amato and Julien Jacques. "MMM: Clustering Multivariate Longitudinal Mixed-type Data". working paper or preprint. Nov. 2024. URL: https://hal.science/hal-04807626 (submitted to *Journal of Computational and Graphical Statistics*)

# Chapter 2

# Model-based Clustering

## 2.1 Introduction

As noted in Chapter 1, model-based clustering is a statistical approach that leverages probabilistic models to perform cluster analysis. This is because, as outlined by Bouveyron, Celeux, Murphy, and Raftery, 2019b "basing cluster analysis on a probability model has several advantages. In essence, this brings cluster analysis within the range of standard statistical methodology and makes it possible to carry out inference in a principled way".
In the following sections, we will describe the general form of the main algorithms belonging to the probabilistic clustering family, that is the finite mixture models, and the Expectation-Maximisation algorithm, which is widely used to infer their parameters (that is, to "learn"). Then, we will detail the most used of these models, the Gaussian mixture model and its particular parametrization. Finally, we will present the main criteria to select the number of clusters $K$.

## 2.2 Finite Mixture Models

Finite Mixture Models (FMMs) represent the probability density function of random variables as a weighted sum of component densities. This concept was first introduced by Pearson, 1894, where the author modelled the distribution of ratios between forehead width and body length for 1000 Neapolitan crabs using a mixture of two univariate Gaussian distributions. Given a dataset $y = (y_1, \ldots, y_N)$ with $N$ multivariate observations of dimension $J$, a FMM expresses the probability distribution of an observation $y_i$ as:

$$f(y_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k) \tag{2.2.1}$$

Here, $\pi_k \geq 0$ for all $k$, and $\sum_{k=1}^{K} \pi_k = 1$. The parameter $\pi_k$ is the mixing proportion, and $f_k(y_i; \theta_k)$ is the component density with parameters $\theta_k$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. The model can be characterized hierarchically by introducing a latent vector $\ell_i \in \{0, 1\}^K$, where $\ell_{ik} = 1$ if $y_i$ belongs to the $k$-th component, and $v_{ik} = 0$ otherwise. The joint density of $(y_i, \ell_i)$ is:

$$f(y_i, \ell_i; \boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k f_k(y_i; \theta_k)]^{\ell_{ik}} \tag{2.2.2}$$

The marginal density of $y_i$ is then:

$$f(y_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k) \tag{2.2.3}$$

The observed data log-likelihood is:

$$l_o(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k). \tag{2.2.4}$$

When all components $f_k$ belong to the same family of probability distributions and their parametric form is assumed to be known, the we talk about homogeneous parametric FMMs, such as Gaussian Mixture Models (GMMs) (McLachlan and Peel, 2000), t-distribution Mixture Models (Peel and McLachlan, 2000), and Skew-Normal Mixture Models (Lin, Lee, and Yen, 2007). In these models, the component densities arise from the same parametric family. When the component densities come from different parametric families, then we call them heterogeneous parametric FMMs. For non-parametric FMMs, no assumptions are made about the form of the component densities.

## 2.3 Expectation-Maximisation Algorithm

The Expectation-Maximisation (EM) algorithm is the main method for estimating the parameters of FMMs (Titterington, Smith, and Makov, 1985). It involves iterating between the Expectation (E) step and the Maximisation (M) step until convergence. The algorithm was developed by Dempster, Laird, and Rubin, 1977 in the context of Maximum Likelihood Estimation (MLE) problems involving latent variables or incomplete data. This is why in this framework we distinguish between a complete likelihood and an observed likelihood: the complete likelihood is the likelihood function assuming both observed and latent variables are known, while the observed likelihood is obtained by integrating out (marginalizing) the latent variables.
Let $\ell_i \in \{0,1\}^K$ be a $K$ dimensional latent vector that takes values $\ell_{ik} = 1$ if the unit $i$ belongs to the $k$-th cluster and 0 otherwise. Given a sample $\boldsymbol{y} = (y_1, \ldots, y_N)$, $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_N)$ the complete-data log-likelihood function is:

$$l_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\ell}) = \sum_{i=1}^{N} \sum_{g=1}^{K} \ell_{ik} \left\{ \log \pi_k + \log f_k(y_i; \theta_k) \right\}. \tag{2.3.1}$$

The EM algorithm starts with an initial parameter vector $\boldsymbol{\theta}^{(0)}$. 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. It iterates as follows:

**E-Step:** computes the expectation of the complete-data log-likelihood conditional on $y$ using the current parameter estimates $\hat{\boldsymbol{\theta}}^{(s)}$. This expectation is defined as $\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)}) := \mathbb{E}(l_c(\boldsymbol{\theta}; y, \boldsymbol{\ell}) | \hat{\boldsymbol{\theta}}^{(s)}, y)$, and it is:

$$\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{\tau}_{ik}^{(s+1)} \left\{ \log \hat{\pi}_k^{(s)} + \log f_k(y_i; \hat{\theta}_k^{(s)}) \right\} \tag{2.3.2}$$

where $\hat{\tau}_{ik}^{(s+1)}$ is the posterior probability that $y_i$ belongs to the $k$-th component at iteration $s+1$:

$$\hat{\tau}_{ik}^{(s+1)} = \frac{\hat{\pi}_k^{(s)} f_k(y_i; \hat{\theta}_k^{(s)})}{\sum_{k'=1}^{K} \hat{\pi}_{k'}^{(s)} f_{k'}(y_i; \hat{\theta}_{k'}^{(s)})}. \tag{2.3.3}$$

**M-Step:** updates the parameter estimates maximizing $\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)})$:

$$\hat{\boldsymbol{\theta}}^{(s+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)}). \tag{2.3.4}$$

The mixing proportions $\hat{\pi}_k^{(s+1)}$ are updated as:

$$\hat{\pi}_k^{(s+1)} = \frac{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)}}{N}, \forall k \in \{1, \ldots, K\}. \tag{2.3.5}$$

The parameters $\hat{\boldsymbol{\theta}}_k^{(s+1)}$ are updated by solving:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \hat{\tau}_{ik}^{(s+1)} \frac{\partial \log f_k(y_i; \theta_k)}{\partial \boldsymbol{\theta}} = 0. \tag{2.3.6}$$

The EM algorithm alternates between the E and M steps until convergence. While the algorithm is not guaranteed to converge to the global maximum of the observed likelihood function, it is assured to converge to a local maximum or a saddle point under certain conditions on the observed likelihood function and the properties of the parameter space, such as bounded likelihood function or its continuity and differentiability (Wu, 1983). Convergence of the EM algorithm to a local maximum of the observed log-likelihood function can be assessed in several ways. In essence, these consist of seeing whether the algorithm has been moving slowly in the latest iterations. One possible criterion is that the observed log-likelihood has changed very little between the last two iterations; a typical threshold is a change of less than $1 \cdot 10^{-5}$.
The EM algorithm is preferred over other iterative procedures like the Newton-Raphson method (Ypma, 2012) because it is less sensitive to the starting values and more efficient for high-dimensional parameter spaces. However, depending on the chosen distributions $f_k$, Equation 2.3.6 can be lead to explicit estimators or not. In the latter case, more advanced forms of the EM algorithm may be needed. We refer the reader to McLachlan and Krishnan, 2007 for a complete survey.

## 2.4 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are a common type of FMM where the component densities are multivariate Gaussian distributions. The parameters $\theta_k$ include the mean vector $\mu_k$ and covariance matrix $\Sigma_k$. GMMs are popular because for many natural processes is easy to assume Gaussian distributions, making GMMs a suitable model for various applications. The GMM can be written as:

$$f(y_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_J(y_i \mid \mu_k, \Sigma_k), \tag{2.4.1}$$

where

$$\mathcal{N}_J(y_i \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^J |\Sigma_k|}} \exp\left\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\right\} \tag{2.4.2}$$

is the density of a multivariate Gaussian distribution. As an example, Equation 2.3.1 for GMMs becomes:

$$l_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\ell}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \ell_{ik} \left[\log \pi_k + \log \mathcal{N}_J(y_i \mid \mu_k, \Sigma_g)\right]. \tag{2.4.3}$$

And in turn, in the E-step, Equation 2.3.3 becomes:

$$\hat{\tau}_{ik}^{(s+1)} = \frac{\hat{\pi}_k^{(s)} \mathcal{N}_J(y_i \mid \hat{\boldsymbol{\mu}}_k^{(s)}, \hat{\Sigma}_k^{(s)})}{\sum_{k'=1}^{K} \hat{\pi}_h^{(s)} \mathcal{N}_J(y_i \mid \hat{\boldsymbol{\mu}}_{k'}^{(s)}, \hat{\Sigma}_{k'}^{(s)})}. \tag{2.4.4}$$

Note that, in the E-step, the conditioning is on the current parameter estimates, hence the use of hats on the parameters. It follows that the expected value of the complete-data log-likelihood in Equation 2.3.2 is:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(s)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\tau}_{ik}^{(s+1)} \left[\log(\hat{\pi}_k^{(s)}) - \frac{J}{2}\log 2\pi - \frac{1}{2}\log|\hat{\Sigma}_k^{(s)}| - \frac{1}{2}\operatorname{tr}\left\{(y_i - \hat{\mu}_k^{(s)})(y_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1,(s)}\right\}\right] \\ &= \sum_{k=1}^{K} \hat{n}_k^{(s+1)} \log(\hat{\pi}_k) - \frac{NJ}{2}\log 2\pi - \sum_{k=1}^{K} \frac{\hat{n}_k^{(s+1)}}{2} \log|\hat{\Sigma}_k^{(s)}| \\ &\quad - \sum_{k=1}^{K} \frac{\hat{n}_k^{(s+1)}}{2} \operatorname{tr}\left\{\frac{1}{\hat{n}_k^{(s+1)}} \sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)} (y_i - \hat{\mu}_k^{(s)})(y_i - \hat{\mu}_k^{(s)})^T \hat{\Sigma}_k^{-1,(s)}\right\}, \end{aligned} \tag{2.4.5}$$

where $\hat{n}_k = \sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)}$.

In the M-step, the model parameters are updated by maximizing Equation 2.4.5 with respect to $\tau_k, \mu_k$, and $\Sigma_k$. This yields the updates:

$$\hat{\pi}_k^{(s+1)} = \frac{\hat{n}_k^{(s)}}{N}, \quad \hat{\mu}_g^{(s+1)} = \frac{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)} y_i}{\hat{n}_k^{(s+1)}}, \tag{2.4.6}$$

while

$$\hat{\Sigma}_k^{(s+1)} = \frac{1}{\hat{n}_k^{(s+1)}} \sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)} (y_i - \hat{\mu}_k^{(s+1)})(y_i - \hat{\mu}_k^{(s+1)})^T. \tag{2.4.7}$$

Then, as said in the above section, the EM algorithm alternates between the E and M steps until convergence. Local convergence is guaranteed under certain conditions. A study of the behaviour of the algorithm near degenerated solutions when some of these conditions are not met is provided by Biernacki and Chrétien, 2003.

### 2.4.1 Parametrization

However, GMMs can require the estimation of a large number of parameters. For example, for a $J$-variate GMM fitted for $K$ components (clusters) requires a total of $KJ$ parameters for the vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $K-1$ for the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $KJ(J+1)/2$ for the covariance matrices $(\Sigma_k)_k$. To reduce the number of parameters, the covariance matrix $\Sigma_k$ can be decomposed using eigenvalues and eigenvectors:

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{2.4.8}$$

where $\lambda_k \in \mathbb{R}$ determines the volume and is equal to $|\Sigma_k|^{1/J}$, $D_k \in \mathbb{R}^{J\times J}$ is the matrix of eigenvectors of $\Sigma_k$ and determines the orientation, and $A_k \in \mathbb{R}^{J\times J}$ is a diagonal matrix such that $|A_k| = 1$ with the normalized eigenvalues of $\Sigma_k$ and the determines the shape of the $k$-th cluster. Various restrictions on $\lambda_k$, $D_k$, and $A_k$ lead to fourteen different models, each with different geometric characteristics. Table 2.1 reports all the possible combinations of these restrictions with the corresponding geometrical characteristics.

| Eigen decomposition of $\Sigma_k$ | # of parameters for $(\Sigma_k)_k$ | Sphericity | Volume | Shape | Orientation | Model Name |
|---|---|---|---|---|---|---|
| $\lambda I$ | 1 | Spherical | Equal | Equal | None | EII |
| $\lambda_k I$ | $K$ | Spherical | Variable | Equal | None | VII |
| $\lambda A$ | $K+J-1$ | Diagonal | Variable | Equal | Coordinate axes | VEI |
| $\lambda A_k$ | $K(J-1)+1$ | Diagonal | Equal | Variable | Coordinate axes | EVI |
| $\lambda_k A_k$ | $KJ$ | Diagonal | Variable | Variable | Coordinate axes | VVI |
| $\lambda A$ | $J$ | Diagonal | Equal | Equal | Coordinate axes | EEI |
| $\lambda DAD^T$ | $J(J+1)/2$ | Elliptical | Equal | Equal | Equal | EEE |
| $\lambda_k DAD^T$ | $J(J+1)/2+K-1$ | Elliptical | Variable | Equal | Equal | VEE |
| $\lambda DA_k D^T$ | $J(J+1)/2+(K-1)(J-1)$ | Elliptical | Equal | Variable | Equal | EVE |
| $\lambda_k D_k A D_k^T$ | $KJ(J+1)/2-(K-1)J$ | Elliptical | Variable | Equal | Variable | EEV |
| $\lambda_k D_k A_k D_k^T$ | $KJ(J+1)/2$ | Elliptical | Variable | Variable | Variable | VVV |

Table 2.1: The fourteen models for parameterizations of the covariance matrix $\Sigma_k$ in GMMs.

This solution was first proposed by Banfield and Raftery, 1993 and Celeux and Govaert, 1995. These models offer different levels of flexibility and complexity, allowing practitioners to choose the most appropriate model for their data. As said above, each of these models in endowed with different geometric properties and interpretations. Figure 2.1 shows examples of contours of the component densities for the various models in the two-dimensional case with two mixture components.

## 2.5 Model Selection

In practice, the number of components $K$ and the best model among the fourteen GMM variants are often unknown. IN a model-based clustering context, we can leverage the information provided by the likelihood of the model and make use of model selection criteria, such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and the Integrated Classification

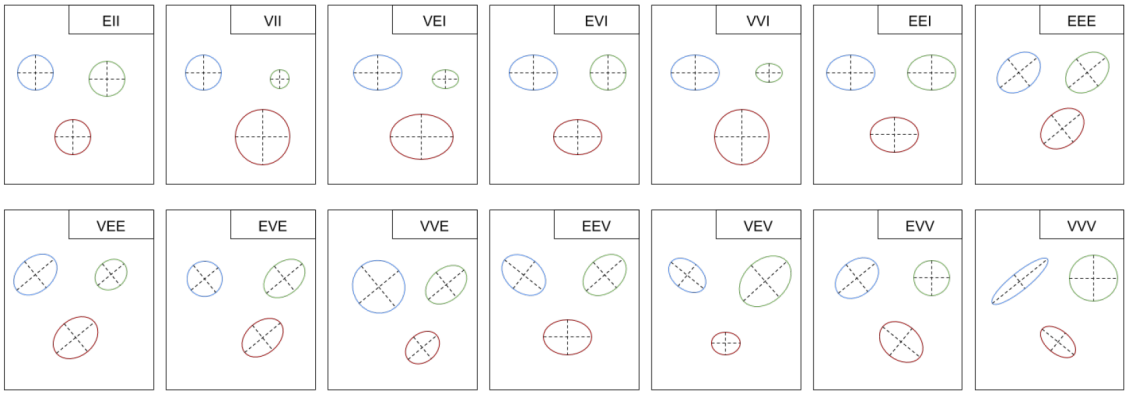



Figure 2.1: Graphical representations of the fourteen models for GMMs, in dimension $J = 2$. Source: Selosse, 2020

Likelihood (ICL) to choose the optimal model. In simpler terms, each criterion (BIC, ICL, AIC) is designed to approximate a different aspect of model evaluation: AIC aims to minimize the difference between the estimated model and the true model, BIC focuses on the overall likelihood considering parameter uncertainty, and ICL includes latent structure in its evaluation. A small description of each criterion follows.

**Akaike Information Criterion (AIC):** the AIC was introduced by Akaike, 1974, from whom it derives its name. The AIC goal is to minimize the Kullback-Leibler (KL) divergence(Kullback and Leibler, 1951) between the estimated model and the true model, thus minimizing the estimated information loss. For a mixture model with $K$ components, the AIC is defined as:

$$\text{AIC}_K = -2l_o(y, \hat{\boldsymbol{\theta}}) + 2\nu_k \tag{2.5.1}$$

where $l_o(y, \hat{\theta}_k)$ is the observed log-likelihood of the model given the data and the estimated parameters for the number of clusters $k$, and $\nu_k$ is the number of parameters in the model. The AIC is particularly useful for comparing models with different numbers of parameters, as it helps to avoid overfitting by penalizing models with too many parameters. However, AIC may still select overly complex models, especially in small sample sizes, because its penalty term is relatively mild. When the AIC is computed as in Equation 2.5.1, the model with the lowest AIC value is generally preferred. However, is worth noting that some authors may prefer to employ a different sign for the log-likelihood term, leading to change this rule and select the model which maximizes the criterion.

**Bayesian Information Criterion (BIC):** the BIC aims to maximize the posterior distribution of the model given the data. For equal priors on the model, this amounts to maximizing the integrated observed likelihood of the model, which is the likelihood of the data given the model, averaged (integrated) over the prior distribution of the model parameters. Introduced by Schwarz, 1978, for a mixture model with $K$ clusters is defined as:

$$\text{BIC}_K = -2l_o(y, \hat{\boldsymbol{\theta}}) + \nu_K \log(N) \tag{2.5.2}$$

where the only new term with respect to Equation 2.5.1 is the logarithm of the sample size $N$. As for the AIC, and subject to the same sign choice conditions, the model with the lowest BIC value is selected. Unlike the AIC, BIC penalizes model complexity more heavily, which helps to avoid overfitting, especially in small sample sizes. Moreover, the BIC is consistent in selecting the true model as the sample size increases (Claeskens and Hjort, 2008), making it reliable for model selection. However, BIC's stronger penalty can sometimes lead to the selection of overly simple models, which may not capture all the nuances of the data.

**Integrated Classification Likelihood (ICL):** introduced by Biernacki, Celeux, and Govaert, 2000, the ICL can be thought of as a variant of the BIC, since it maximizes the integrated complete likelihood, which considers not only the data but also the latent structure, providing a more comprehensive evaluation of the model. Indeed, it extends the BIC by incorporating an entropy term that accounts for the uncertainty in the clustering or classification of the data. For a mixture model with $K$ components, it is defined as:

$$\mathrm{ICL}_K = \mathrm{BIC}_K - \sum_{k=1}^{K}\sum_{i=1}^{N} p(\ell_{ik} = 1 \mid y; \hat{\theta}_k) \log p(\ell_{ik} = 1 \mid y; \hat{\theta}_k) \tag{2.5.3}$$

The ICL is often preferred when the complete-data log-likelihood is easier to compute. The introduction of the estimated mean entropy penalizes overlapping clusters and this makes ICL prefer models with well-separated clusters.

### 2.5.1 Conclusion

Model-based clustering using Finite Mixture Models is a powerful framework for density estimation and clustering, allowing for the use of statical techniques to address challenges and questions that often arise in clustering analysis. However, the high dimensionality of parameters can be a challenge. Solutions like different parametrization choices can help reduce the computational burden and can help mitigate the curse of dimensionality. For instance, thanks to the probabilistic framework, the computation of the likelihood of the clustering model is possible, allowing for the use of information criteria to select the optimal number of clusters when this is unknown. The concepts introduced in this section will be further developed and applied in subsequent chapters. These foundational ideas are essential for understanding and implementing model-based clustering in various applications.

# Chapter 3

# Model-based Clustering for Longitudinal Mixed-type Data: a Survey

## 3.1 Introduction

As we saw in Chapter 1, longitudinal data of mixed-type are increasingly common in many areas of science, but the analysis of such kind of data is a current research problem in the fields of statistics and machine learning. Concerning mixed-type data, practitioners often tends to transform the data and treat them as continuous because of the scarcity of existing adequate solutions. Such an approach is not satisfying since it leads either to the introduction of a bias or to an important information loss. The second scientific obstacle is the modelling of the temporal evolution. Currently, it is often the case for the analyses to be conducted independently at each temporal phase, to then trying *a posteriori* to find links between these different analyses. The ideal way would be a model able to describe the temporal evolution. Thus, the analysis will exhibit typical temporal evolution behaviours.
This chapter aims at providing a summary on the main methods existing in the literature regarding model-based clustering for longitudinal mixed-type data. Nevertheless, given the limited literature available on such methods, we deemed useful for the interested reader to include also a survey on the main approaches to (probabilistically) cluster mixed-type data and longitudinal ones. While the literature on clustering longitudinal mixed-type data is relatively limited, when considered in conjunction with the broader research on clustering longitudinal and mixed-type data independently, its volume becomes significant, especially since much of the existing literature tends to focus on methodological details or minor adjustments. Moreover, we want to focus on methods that can be implemented using open-source softwares, specifically the ones available through the R language (R Core Team, 2023). This is why, while we aim at providing a comprehensive survey to the reader, we will focus just on subset of the literature that is interesting to this thesis and that satisfies the software requirements.

## 3.2 Chapter organization

In the following, we will first present the kind of data we aim at treating in our analysis in Section 3.3, that is longitudinal mixed-type data, and introduce the notation that we will use throughout this work, if not differently specified. Then, a brief overlook on the main methods to cluster mixed-type data (Section 3.4), longitudinal data (Section 3.5) and finally longitudinal mixed-type data (Section 3.6) is given at the beginning of each section, while we will provide details on the main approaches in the rest of each section. In Section 3.7 we will present some useful softwares to implement the exposed models and finally we will draw some conclusions in Section 3.8.

## 3.3 Longitudinal mixed-type data: notation

As the reader can imagine, a variety of notations has been employed by different authors throughout their work, and they are sometimes difficult to standardize. Nonetheless, we will try to keep a coherent notation and express similar concepts by similar notation to make this work easier to follow. Let denote by $y_{ijt}$ the observation of the $j$-th ($j = 1, .., J$) variable for the $i$-th ($i = 1, \dots, N$) unit at time $t$ ($t = 1, \dots, T$), that is: imagine to observe $N$ units and measuring $J$ different mixed variables $T$ times throughout the course of the study. Consider $D$ different types of features, such that $J = \sum_{d=1}^{D} J_d$. $J_d$ is the number of features (variables) of the $d$-th data-type and $j_d$ indicates the general $j_d$-th variable of that type with $j_d = 1, \dots, J_d$.
Moreover, being our intent to perform clustering, let denote with $\ell_i$ the general cluster allocation variable for the $i$-th units, which will have different distribution or structure depending on the model to keep the notation simple. So, for example, depending on the model $\ell_i$ can be a categorical variable taking values $\ell_i \in \{1, \dots, K\}$ or a vector such that $\ell_i \in \{0, 1\}^K$. The probability of belonging to a cluster $k$ is $\pi_k$, where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Again, sometimes, we will abuse this notation and specify a different structures of $\pi_k$ depending on the model, but it will always relate to the probability of belonging to a cluster $k$. For the same principle, we will make use of the notation $\mathbb{P}(x)$ to indicate its density function $f(x)$ when the variable is continuos and $\mathbb{P}(X = x)$ when it is discrete. We will also use the more generic $\mathbb{P}(x)$ when the type is not specified.

Models can have different level of granularity. In order for us to adequate our notation accordingly but to keep it compact, if not differently specified, $y_i = (y_{i11}, \dots, y_{i1T}, \dots, y_{iJ1}, \dots, y_{iJT})^\top$, $y_j = (y_{1j1}, \dots, y_{1jT}, \dots, y_{Nj1}, \dots, y_{NjT})^\top$, $y_t = (y_{1jt}, \dots, y_{1jt}, \dots, y_{NJt}, \dots, y_{NJt})^\top$ and also allowing for the possible combinations, for example $y_{ij} = (y_{ij1}, \dots, y_{ijT})^\top$. Similarly, when one of the three dimensions is just reduced to be equal to one, as is the case of cross-sectional data ($t = 1$), we would have $y_j = (y_{1j1}, \dots, y_{Nj1})^\top$ and $y_{ij} = y_{ij1}$ would be just a scalar. As we want to avoid making the notation unnecessarily heavy, in these cases we will avoid to write down the unitary dimension, that is for example we will just write $y_{ij}$. The same notation will be used for latent variables when they are assumed.
It is often the case for models to include covariates: when these covariates will influence the outcome, they will be defined as $x$, while when they will influence the probability of belonging to one cluster we will use the symbol $\xi$. Again, the covariates can have different structures and vary or not in time and also be different for different outcomes: we will use the same notation described in the previous paragraph and their characteristics will be noted in the description of each model. In addition, models have parameters that we will denote with $\theta$, and $\theta_k$ when they are associated with the belonging to the $k$-th cluster.

Finally, the bold characters will identify collections, such as $\mathbf{y} = \{y_i\}_{i=1}^N$, where units are assumed to be independent, $\boldsymbol{\ell} = \{\ell_i\}_{i=1}^N$, $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^K$, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$.

### 3.3.1 Problems at hand

Two main problems arise when dealing with longitudinal mixed-type data, each due to one of the two natures of them:

#### Mixed-type data

When dealing with different types of data, it is often the case to deal with different distributions. In model-based clustering, the most common approach in this case is to assume that the different variables pertaining to different distributions are independent conditionally on the cluster belonging, which is assumed to explain all the possible relationship among them. Under this assumption of conditional independence, the joint probability of two variables of different type is given by the product of the two different distribution densities, that may not share the same support. This creates a problem, since the variables with the wider support could "overwhelm" the variables with stricter support, implicitly creating an importance ranking among variables.

#### Temporal Dynamics

Accounting for the temporal dynamics in statistical task is not an easy task, nor it straightforward, and even more so in clustering. Common longitudinal models often tend to cluster taking little into account the information provided by the time evolution trajectories, resulting in an information loss. For parametric approaches, the dynamical component is often introduced via appropriate modelling of the parameter's dynamics through potential smoothness or carefully designed jump processes when necessary.

## 3.4 Literature on cross-sectional mixed-type data

### 3.4.1 Overlook

Although many data sets contain mixed-type data, few mixture models can manage these data (Hunt and Jorgensen, 2011) due to the shortage of multivariate distributions able to handle them. Clustering with mixed-type data have received a large attention in the last decade from researchers in statistics and machine learning. The Latent Class Model (Everitt, 1984) is frequently used, and it assumes that the variables are conditionally independent upon the cluster membership. Consequently, the joint probability distribution function (pdf) of the features of different types is obtained by the product of the pdfs of each individual feature. However, when the variables are inherently correlated in a cluster, this model is not suitable. To overcome this issue, the authors of Marbac, Biernacki, and Vandewalle, 2017 wanted to conserve standard marginal distributions but also try to loosen the conditional independence on the variables. For this purpose, they use a copula, which allow definition of both the dependence model and the type of marginal distributions. The proposed model relies on the main assumption that each cluster follows a Gaussian copula. However, the authors note that model complexity increases promptly with the number of variables, which is not

suitable in a big-data context. Moreover, it is not easily interpretable by non-statistician practitioners. More recently, Hermes, Heerwaarden, and Behrouzi, 2024 proposed a similar approach by using copulas in the context of graphical models, which were already extended for use for mixed-type data by Cheng, Li, Levina, and Zhu, 2017. In Selosse, Jacques, and Biernacki, 2020, another model-based approach for ordinal, nominal, integer and continuous data is proposed, on the basis of conditional independence assumption and with the particularity of creating clusters of features as well as clusters of individuals (co-clustering).
Another way to address the issues of mixed-type data is to see some variables as the manifestation of latent variables. For example, in McParland and Gormley, 2016, the clustMD model considers continuous and categorical data (nominal and ordinal) and assumes that a categorical variable is the representation of an underlying latent continuous variable. Then, it is assumed that the continuous variables (observed and unobserved) follow a multivariate Gaussian mixture model. This model is further developed to address sparsity by Choi, Ahn, and Kim, 2023.

### 3.4.2 Latent Class Model

In the Latent Class Model (LCM) (Everitt, 1984), within each latent class, the observed variables are assumed to be statistically independent. The model assumes that there is a categorical latent variable $\ell_i$ with $K$ classes, which by consistency we will describe as a vector, such that $\ell_i \in \{0,1\}^K$ as described before. The observed variables $y_i = (y_{i1}, y_{i2}, \dots, y_{iJ})^\top$ are conditionally independent given the latent vector $\ell_i$ and each class $k$ has its own set of parameters governing the distribution of $y$.

Given the latent class variable $\ell_i$, the observed variables $y_1, y_2, \dots, y_J$ are independent, therefore $\mathbb{P}(y_{i1}, y_{i2}, \dots, y_{iJ} \mid \ell_i = k) = \prod_{j=1}^{J} \mathbb{P}(y_j \mid \theta_{jk})$. This means that the marginal probability of the observed data is:

$$\mathbb{P}(y_i) = \sum_{k=1}^{K} \mathbb{P}(y_i, \ell_{ik} = 1) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \mathbb{P}(y_{ij} \mid \theta_{jk}). \tag{3.4.1}$$

So, for a dataset with $N$ observations, the complete likelihood is:

$$\mathcal{L}_C(\boldsymbol{y}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[ \pi_k \prod_{j=1}^{J} \mathbb{P}(y_{ij} \mid \theta_{jk}) \right]^{\ell_{ik}}, \tag{3.4.2}$$

where $\theta_{jk}$ represents the parameters of the distribution of the $j$-th variable when the $i$-th unit belongs to the $k$-th cluster. An individual probability of belonging to latent class $k$ is modelled using a multinomial logistic regression according to covariates $\xi$

$$\mathbb{P}(\ell_{ik} = 1 \mid \xi_i) = \frac{\exp(\beta_{0k} + \xi_i^\top \beta_{1k})}{\sum_{k'=1}^{K} \exp(\beta_{0k'} + \xi_i^\top \beta_{1k'})},$$

where $\xi_i$ are some covariates and for identifiability $\beta_{0K} = \beta_{1K} = 0$. An EM algorithm is used for inference. For a through review on LCMs and their extensions, we suggest Proust-Lima, Séne, Taylor, and Jacqmin-Gadda, 2014.

### 3.4.3 clustMD

One of the most famous algorithms, and consequently R package, to deal with mixed data is the one provided by `clustMD` (McParland and Gormley, 2016). It is assumed that a latent variable, following a mixture of Gaussian distributions, generates the observed data of mixed type, and it can deal with any combination of continuous, binary, ordinal or nominal variables.
Let the observed $J$ variables be of mixed type. The model assumes that each observation vector $y_i = (y_{i1}, \ldots, y_{iJ})$ is a manifestation of an underlying latent continuous vector, $z_i$, which follows a Gaussian mixture distribution. For each variable type, the model assumes a different link between the latent variable and the observed one. When the $j$-th variable type is continuos, then an identity link is established, such that $y_{ij} = z_{ij}$. When the $j$-th variable type is ordinal, then for the observed response $y_{ij}$ with $C_j$ levels let $\gamma_j$ denote a $C_{j+1}$ vector of thresholds that partition the real line. The threshold parameters are constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \ldots \leq \gamma_{j,C_j} = \infty$. If the latent $z_{ij}$ is such that $\gamma_{j,c_j-1} < z_{ij} < \gamma_{j,c_j}$ then the observed ordinal response, $y_{ij} = c_j$. Thus the probability of observing level $c_j$ can be expressed as the difference between two standard Gaussian cumulative distribution functions, expressed as $\mathbf{\Phi}$:

$$\mathbb{P}(y_{ij} = k) = \Phi\left(\frac{\gamma_{j,c_j} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,c_{j-1}} - \mu_j}{\sigma_j}\right).$$

When the variable type is binary, then this variable can be thought of as an ordinal variable with two levels. Nominal variables are more difficult to model since the set of possible responses is unordered. For nominal variable $j$ with $C_j$ possible responses, the continuous underlying variable has $C_{j-1}$ dimensions, i.e. $z_{ij} = (z_{ij}^1, \ldots, z_{ij}^{C_{j-1}}) \sim \mathcal{N}_{C_{j-1}}(\mu_j, \Sigma_j)$. The observed nominal response $y_{ij}$ is a manifestation of the values of the elements of $z_{ij}$ relative to each other and to a threshold, assumed to be 0.

$$y_{ij} = \begin{cases} 1 & \text{if } \max_s\{z_{ij}^s\} < 0; \\ \\ c_j & \text{if } z_{ij}^{K_j-1} = \max_s\{z_{ij}^s\} \ \textit{and} \ z_{ij}^{c_j-1} > 0 \ \text{ for } s = 2, \ldots, C_j. \end{cases}$$

Thus the joint vector of observed and latent continuous data is assumed to follow a multivariate Gaussian distribution $z_i \sim \mathcal{N}_P(\mu, \Sigma)$, where $P = C + O + \sum_{j=C+O+1}^{J}(C_j - 1)$, since more than one latent dimension is required to model each nominal variable.
Finally, the joint model for mixed data is embedded in a finite mixture model, facilitating the clustering of mixed data: it is assumed that $z_i$ follows a mixture of $K$ Gaussian distributions:

$$z_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}_P(\mu_k, \Sigma_k),$$

where $\Sigma_k$ is assumed to be diagonal, that is the variable are assumed independent conditionally on the cluster allocation. The complete data likelihood under the clustMD model is given by:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k \cdot \mathcal{N}_C(z_i^\alpha | \mu_k^\alpha, \Sigma_k^\alpha) \cdot \mathcal{N}_{(P-C)}(z_i^\beta | \mu_k^\beta, \Sigma_k^\beta)\right]^{\ell_{ik}}, \tag{3.4.3}$$

where $z_i^\alpha$ represents the observed continuous variables for observation $i$, while $z_i^\beta$ corresponds to the latent variables associated with the categorical data, and $z_i = [z_i^\alpha, z_i^\beta]^\top$. An EM algorithm is used

for inference if nominal variables are not present, otherwise the model relies on an Monte Carlo EM algorithm, since a numerical approach is needed to approximate the expectation of the latent structure for nominal variables. The product between the two multivariate normal distribution in Equation 3.4.3 follows directly from the assumption of $\Sigma_k$ be diagonal, which implies of conditional independence.

### 3.4.4 Multiple Latent Block Model

The Multiple Latent Block Model (MLBM) (Selosse, Jacques, and Biernacki, 2020) is an extension of the Latent Block Model (LBM) (Govaert and Nadif, 2003, Govaert and Nadif, 2005), designed to handle mixed-type data by partitioning observations and features into clusters. The classical LBM is a co-clustering algorithm, and therefore it is assumed that both row-clusters and column-clusters exist. However, it can also be used for simple clustering. For mixed-type data, the observed matrix $Y$ of dimension $N \times J$ is assume to be composed of $D$ different sets of features such that $J = \sum_{d=1}^{D} J_d$. So, the matrix $y$ is partitioned into $D$ column subsets $y = (y^1, \dots, y^D)$, where each subset $y^d$ contains features of the same type. For each subset $y^d$, $H_d$ column clusters are defined. Let $\boldsymbol{w} = \{\boldsymbol{w}^d\}_{d=1}^{D} = \{w_{j_d}^d\}_{d=1}^{D}$ denote the column partitions of the d-th matrix with $w_{j_d}^d$ be a $H_d$ dimensional vector with $w_{j_d h_d}^d = 1$ is the $j_d$-th variable belongs to the $h_d$-th cluster, and let $\boldsymbol{\rho} = \{\rho^d\}_{d=1}^{D}$, where $\rho^d = (\rho_1^d, \dots, \rho_{H_d}^d)$ are the corresponding mixing proportions. Likewise, let $\boldsymbol{\ell}$ denote now row partitions and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ the corresponding mixing proportions for the rows. The joint likelihood for mixed data is:

$$\mathbb{P}(y|\boldsymbol{\ell}, \boldsymbol{w}; \boldsymbol{\theta}) = \prod_{d=1}^{D} \mathbb{P}(y^d|\boldsymbol{\ell}, \boldsymbol{w}^d; \boldsymbol{\theta}^d), \quad \text{where} \quad \mathbb{P}(y^d|\boldsymbol{\ell}, \boldsymbol{w}^d; \boldsymbol{\theta}^d) = \prod_{i,j,k,h_d} \mathbb{P}(y_{ij}^d; \theta_{kh_d}^d)^{\ell_{ik} w_{jh_d}^d}, \tag{3.4.4}$$

with $\boldsymbol{\theta}^d = (\theta_{kh_d}^d)_{k,h_d}$ be the distribution parameters of block $(k, h_d)$ of the of the matrix $y^d$. The first equation implies that the $D$ matrices data are conditionally independent of the row and column partition and the second that the univariate random variables $y_{ij}^d$ are conditionally independent given row partitions $\ell$ and column partitions $w^d$. Moreover, the latent variables $\ell, w^1, \dots, w^D$ are assumed to be independent:

$$\mathbb{P}(\boldsymbol{\ell}, \boldsymbol{w}; \boldsymbol{\pi}, \boldsymbol{\rho}) = \prod_{i,k} \pi_k^{\ell_{ik}} \prod_{d} \prod_{j,h} \rho_h^{d\, w_{jh_d}^d}.$$

Finally, the complete likelihood can be expressed as:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{(\boldsymbol{\ell}, (\boldsymbol{w}^d)_{d=1}^{D})} \prod_{i,k} \pi_k^{\ell_{ig}} \prod_{d} \rho_h^{d\, w_{jh_d}^d} \prod_{i,j,k,h_d} \mathbb{P}(y_{ij}^d; \theta_{kh_d}^d)^{\ell_{ik} w_{jh_d}^d}. \tag{3.4.5}$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\theta})$ are the model parameters. Because of the model complexity, a Stochastic Expectation-Maximization (SEM) algorithm with Gibbs sampling is used for inference.

Block-specific distribution $\mathbb{P}(y_{ij}^d; \theta_{kh_d}^d)^{\ell_{ik} w_{jh_d}^d}$ depends on the data type: continuous data modelled by a Gaussian distribution, while categorical nominal data are modelled by a multinomial distribution, count data by a Poisson distribution and ordinal by BOS (Binary Ordinal Search) distribution (Biernacki and Jacques, 2016).

This framework effectively models and clusters data with mixed types. However, the co-clustering is performed in such a way that features of different types cannot be part of a same column-cluster, and therefore may not capture associations among features. Additionally, its computational cost grows significantly with the number of rows, columns, and feature types, making it challenging to scale.

## 3.5 Literature on longitudinal data

### 3.5.1 Overlook

Modelling longitudinal data poses a different kind of challenge than mixed-type data, as the grouping has to account for the similarity of individual trajectories which disrupt the independence assumption among observations. Additionally, this kind of data often introduces the issue of dealing with sparse observations in time - meaning observations that are collected infrequently or irregularly over time - that makes unsuitable the use of models coming from the domains such as functional data, time series and Gaussian processes. In order to bypass these problems, some authors preferred to focus on geometric non-parametric clustering algorithms, as done by Bruckers, Molenberghs, Drinkenburg, and Geys, 2016 with an idea based on K-means clustering and by Zhou, Zhang, and Tu, 2023 with hierarchical clustering, among others. For parametric methods, a well established manner to model longitudinal data is through mixed-effects models. This research domain is well-established and vast. We refer to Gad and Kholy, 2012 for an overview and to the related work section of Hui, Dang, and Maestrini, 2024 for the most recent advancements. The main issues with this kind of models are the over-parametrization and the computational burden that often arises with it.
A different strategy is to employ Latent Markov Models (LMM), designed to analyse longitudinal data by identifying unobserved (latent) states that evolve over time. These models assume that the observed data are generated by an underlying Markov process, where transitions between latent states follow specified probabilities. They capture both within-state homogeneity and the temporal dynamics of transitions between states. Seminal works in this area include Langeheine and Van De Pol, 1990, which introduced LMMs for categorical data, and Collins and Wugalter, 1992, who extended the framework to include covariates affecting state transitions. More recently, Bartolucci, Farcomeni, and Pennoni, 2012, 2019 provide comprehensive treatments of LMMs, including computational techniques and applications. A hidden Markov latent variable model was designed for analysing multivariate longitudinal data with application to cocaine use among individuals undergoing treatment by Song, Xia, and Zhu, 2017. An R package, `LMest` has been developed to estimate latent Markov models for longitudinal categorical data.

Another approach to clustering longitudinal data that gained traction in the last decade 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-behaviour, 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 GMMs 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 MMN is provided in Section 2.1 of Wang and Melnykov, 2020). In addition,

in the context of linear mixed models with discrete individual random intercepts used to analyse longitudinal continuous data, Anderlucci and Viroli, 2015 proposed Covariance Pattern Mixture Model (CPMM) which, by leveraging three-way data structures, does not require the usual local independence assumption. This model can be seen as an extension of the proposal of McNicholas and Murphy, 2010 in the multivariate context. 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 can suffer from over-parametrization that leads to estimation issues. To overcome this issue a more parsimonious model (Sarkar, Zhu, Melnykov, and Ingrassia, 2020a) and a new R package (Zhu, Sarkar, and Melnykov, 2022) has been proposed. In addition, Cappozzo, Casa, and Fop, 2024 proposed a lasso-type penalization to account for sparsity. Despite their efficacy, up to now these methods have generally only been applied to continuous data.

### 3.5.2 Latent Markov Models

The first mention of a methodology employing Markov models for longitudinal data can be traced back to Kalbfleisch and Lawless, 1985. More recently, Bartolucci, Farcomeni, and Pennoni, 2012, 2014, Zucchini, MacDonald, and Langrock, 2017 have provided comprehensive summaries of latent (or hidden) Markov models used to analyse longitudinal or time series and developed an R package, `LMest` (Bartolucci, Pandolfi, and Pennoni, 2017).
A Latent Markov Model (LMM) is a statistical model that describes the evolution of a system over time using a set of hidden (latent) states, that form a Markov chain. Latent Markov models are designed for analysing longitudinal multivariate data where a latent process $\ell_i = (\ell_{i1}, \ell_{i2}, \ldots, \ell_{iT})$ ($\ell_{it} \in \{1, \ldots, K\}$) evolves over time according to a first-order Markov chain. The observed responses $y_{ijt}$, measured at time $t$ for individual $i$ and variable $j$, depend on the latent state $\ell_{it}$ and covariates $x_{it}$. The local independence assumption states that the observed responses are conditionally independent given the latent process:

$$\mathbb{P}(y_i|\ell_i, x_i) = \prod_{t=1}^{T}\prod_{j=1}^{J}\mathbb{P}(y_{ijt}|\ell_{it}, x_{it}).$$

The latent process is characterized by:

$$\pi_k := \mathbb{P}(\ell_{i1} = k), \quad \pi_{t,k|k'} := \mathbb{P}(\ell_{it} = k|\ell_{i(t-1)} = k'), \text{ for } t = 2, \ldots, T,$$

for time non-homogeneous case. For time homogeneous case, we have $\pi_{tk|k'} = \pi_{k|k'}$ for $t = 2, \ldots, T$. These probabilities are typically parametrized using multinomial logits to capture transitions between latent states.

The probability of an observed sequence is derived as:

$$\mathbb{P}(y_i) = \sum_{k=1}^{K}\pi_k\prod_{t=2}^{T}\pi_{t,k|k'}\prod_{j=1}^{J}\mathbb{P}(y_{ijt}|\ell_{it}, x_{it}), \tag{3.5.1}$$

where $\mathbb{P}(y_{ijt}|\theta_k)$ depends on the measurement model, i.e. the model according to which observations are related to hidden states. For continuous data and under the conditional independence assumption, the simplest measurement model assumes that $y_{ijt}$ follows a normal distribution:

$$y_{ijt}|\ell_{it} = k, x_{it} \sim \mathcal{N}(\mu_{jk}, \sigma^2_{jk}),$$

where, $\mu_j = x_{it}^\top \beta_j$, with $x_{it}^\top$ being a vector of possibly observed covariates and $\beta_{kj}$ are the parameters for variables $j$. For binary or ordinal data, link functions such as the logit or cumulative logit are used. For example, for categorical data:

$$\text{logit}\left(\mathbb{P}(y_{ijt} = c|\ell_{it} = k, x_{it})\right) = x_{it}^\top \beta_{kj}.$$

Then, the complete likelihood for $N$ observations can be expressed as:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{t=1}^{T} \left[ \pi_{t,k|k'} \prod_{j=1}^{J} \mathbb{P}(y_{ijt} \mid \theta_{jk}, x_{it}) \right]^{\mathbf{1}(\ell_{it}=k)}, \tag{3.5.2}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $\pi_{t,k|k'} = \pi_k$ when $t = 1$.
In the case of multivariate data, to avoid assuming that the observed variables are conditionally independent given the latent state process as done in Equation 3.5.1, one can adopt a single measurement model for all the multivariate outcomes of the same type $d$, such as multivariate Gaussian for continuous data. However, the above approach is difficult to adopt when outcomes are of mixed nature, or when all outcomes are continuous but not conditionally Gaussian. For this reason, it is in generally assumed that the conditional independence assumption holds. For an illustrative example on using Markov models for clustering mixed categorical and continuous values time series see Ghassempour, Girosi, and Maeder, 2014.
Parameter and transition probabilities estimation is performed via the EM algorithm. The forward-backward algorithm, which is an adaption of the the Viterbi algorithm (Viterbi, 1967, Juang and Rabiner, 1991), is used for decoding, i.e. the process of determining the most likely sequence of latent states for each unit over time, given the observed data. The algorithm proceeds through a forward-backward recursion of a complexity similar to the recursions adopted for maximum likelihood estimation within the EM algorithm, so that global decoding may be even performed for long sequences of data.
Note that, depending on the formulation, covariates can be used on the measurement model, as described above, or they can also contribute to the determination of the initial and transition probabilities through regression models using a multinomial logistic (softmax) link. However, it is suggested suggest to avoid that covariates simultaneously affect the distribution of the latent process and the conditional distribution of the response variables given this process. In fact, the two formulations have different interpretations and the resulting model would be difficult to interpret and estimate.

Finally, a note should be added to the clustering philosophy behind the LMM. Within this framework, it is assumed that the observed data is generated by a hidden (latent) Markov process, where the system transitions between a finite number of states over time. Therefore, clusters (states) are held to be constant in time, and units transition between them. This makes not possible to detect common trajectories among units, but just common "behaviour" at each time point. This is because LMMs assume homogeneity across individuals, meaning that the transition dynamics are the same across units. This can be limiting if there is significant heterogeneity in the population. The solution to this problem is presented in the next section. It is worth noticing that despite seeming an easy extension, the package `LMest` does not provide a way to cluster mixed-type data through LMM.

### 3.5.3 Mixed Hidden Markov Models

The Mixed Hidden Markov Models (MHMM) (Bartolucci, Farcomeni, and Pennoni, 2012) is an extension of the standard LMMs framework, designed to account for additional sources of time-invariant dependence in data and to introduce heterogeneity into the model, so to relax the homogeneity assumption by means of random effects. Before we dive into these models, a digression may be needed. The terms Hidden Markov Models and Latent Markov Models are often used as synonyms in many contexts, but there can be subtle differences in how they are applied, depending on the field of study or the specific literature. LMMs can be seen as a broader class of models that include HMMs as a special case, as in some literature LMMs may allow for more complex latent structures. They are often used in social sciences, psychology, and other fields where the latent structure may be more complex than a simple Markov process. On the other hand, HMMs are based on the same assumptions and estimation methods of the LMMs, but the structure of the data it aims to analyse is often different Bartolucci, Farcomeni, and Pennoni, 2012 and they are used in applications like speech recognition, bioinformatics, and time-series analysis. For the purpose of this work, we will use the two terms interchangeably.

MHMMs assume that individuals can be grouped into a finite number of latent classes (clusters), each representing a subgroup of individuals with similar trajectories among states. This is possible thanks to the addition of individual-specific random effects to account for dependence between longitudinal observations and unobserved heterogeneity. Van de Pol and Langeheine, 1990 first proposed the extension of the latent Markov approach to include random effects, while Humphreys, 1997, 1998 suggests a LMM where the transition probabilities matrix depends on subject-specific random effects. In a more recent key paper, Altman, 2007 broadly developed this class of models. Different parametrizations give rise to different possible models: these effects may be assumed to have a continuous or a discrete (Maruotti, 2011) distribution and, as the individual covariates, they may be included in the measurement model or in the latent model. Here, we will limit ourselves to the simplest model and we will focus on the use of random effects for clustering, since there lies our interest.

Let $u_i$ be discrete latent variable representing an additional time-invariant latent cluster membership within the population, $u_i \in \{1, \dots, K_1\}$. The probability of belonging to the latent cluster $k_1$ is denoted as $\lambda_{k_1}$, with $\sum_{k_1=1}^{K_1} \lambda_{k_1} = 1$, where $K_1$ is the number of latent classes. As before, let introduce the latent process $\ell_i = (\ell_{i1}, \dots, \ell_{iT})$ that follows a first-order Markov chain, this time conditional on $u_i$. The initial state probabilities $\pi_{k|k_1}$ and transition probabilities $\pi_{k|k',k_1}$ are given by:

$$\pi_{k|k_1} = \mathbb{P}(\ell_{i1} = k | u_i = k_1), \quad \pi_{t,k|k',k_1} = \mathbb{P}(\ell_{it} = k | \ell_{i(t-1)} = k', u_i = k_1). \tag{3.5.3}$$

for non-homogeneous case. For time homogeneous case, we have $\pi_{tk|k',k_1} = \pi_{k|k',k_1}$ for $t = 2, \dots, T$. These probabilities are typically parametrized using multinomial logits to capture transitions between latent states.

The distribution of the observed responses, $y_i$, is derived as:

$$\mathbb{P}(y_i) = \sum_{k_1=1}^{K_1} \lambda_{k_1} \sum_{k=1}^{K} \pi_{k|k_1} \prod_{t=2}^{T} \pi_{t,k|k',k_1} \prod_{j=1}^{J} \mathbb{P}(y_{ijt} | \ell_{it} = k, u_i = k_1), \tag{3.5.4}$$

Where, as for LMMs, the measurement $\mathbb{P}(y_{ijt}|\ell_{it} = k, u_i = k_1)$ can take different forms depending on the outcomes' data-types. The inference can be carried out by the EM algorithm similarly to Section 3.5.2. MHMMs allow model parameters such as transition probabilities and emission distributions to vary across groups, hence providing a new way to perform clustering of trajectories. IN this way, individuals belonging to the same cluster $k_1$ will have the same initial and transition probabilities. Additionally, MHMMs can handle over-dispersion and mixed-type data more effectively, since the random effect variable $u_i$ can act as a shared latent variable influencing variables of different data-types within each state, both as a discrete and continuous random variable.

### 3.5.4 Growth Mixture Models

Growth Mixture Modelling (GrMM) (Muthén et al., 2002, Jung and Wickrama, 2008, Ram and Grimm, 2009b) stands at the intersection of latent growth curve modelling and finite mixture modelling (Petras and Masyn, 2010) and can be described in a univariate setting where the multivariate Gaussian outcome $y_i \in \mathbb{R}^T$ is observed for individual $i$ at times $t = 1, \dots, T$. Conditional on the $k$-th cluster belonging, the observed outcomes are modelled as:

$$
\begin{aligned}
y_i &= \Lambda_k \eta_{ik} + \varepsilon_i k, \\
\eta_{ik} &= \alpha_k + \Gamma_k x_i + u_i,
\end{aligned}
$$

where $\Lambda_k$ is a $T \times m$ factor loading matrix linking the latent growth factors to the observed outcomes, $\eta_{ik} = (\eta_{0ik}, \eta_{1ik}, \dots, \eta_{m_i k})^\top$ is vector of $m$ latent growth factors (e.g., intercepts and slopes) and $\varepsilon_{ik} \sim \mathcal{N}_T(\mathbf{0}, \Phi_k)$ is a vector of residual errors. In the second equation, $\alpha_k \in \mathbb{R}^{m \times 1}$ and $\Gamma_k \in \mathbb{R}^{m \times q}$ are cluster-specific parameters for the growth factors representing the factors trajectories and its link to (possible) time-invariant covariates collected in the $q$-dimensional vector $\boldsymbol{x}_i$ and $\boldsymbol{u}_{ik}$ is vector of random-effect/residuals following $u_{ik} \sim \mathcal{N}_m(\mathbf{0}, \Sigma_k)$ capturing individual-specific deviations from the class-level mean and account for the relationship among the $m$ factors. Finally $\mathrm{Cov}(\varepsilon_{ik}, u_{ik}) = 0$ is assumed.
More specifically, the latent growth factors $\eta_{ik}$ for individual $i$ describe the trajectory of an individual's data over time. Typically, it includes components like an intercept $\eta_{0im}$, representing the baseline level of the observed outcomes, a linear slope $\eta_{1ik}$ describing the rate of change over time and optional higher-order growth terms (e.g., quadratic or cubic slopes), which are collected in the factor loading matrix. The assumption is that an individual has a certain trajectory class membership that does not change over time.

So, at the end, the vector $y_i$, conditionally on the cluster belonging and the covariates, follows a multivariate normal distribution:

$$
y_i|\ell_{ik} = 1, x_i \sim \mathcal{N}_T(\mu_{ik}, \Psi_k) \tag{3.5.5}
$$

where

$$
\begin{aligned}
\mu_{ik} &= \Lambda_k(\alpha_k + \Gamma_k x_i), \\
\Psi_i &= \Lambda_k \Sigma_k \Lambda_k^\top + \Phi_k.
\end{aligned}
$$

As noted by Muthén and Kaplan, 2004, the flexibility of GrMM allows modelling of both within-class variation and between-class heterogeneity, making it particularly effective for exploring multivariate trajectories over time. Of course, the cluster belonging is not known in advance and therefore it is modelled as mixture.
Additionally, it is easy to see as the covariance matrix $\Phi_k$ accounts for the residual variability among times while $\Sigma_k$ relates to the variability among growth factors.
Grouping as $\boldsymbol{\theta}$ the ensemble of parameters to be estimated, then the complete likelihood function becomes:

$$\mathcal{L}_C(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\ell}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}_T(y_{it} \mid \mu_{ik}, \Psi_k)\right]^{\ell_{ik}} . \tag{3.5.6}$$

The model can be estimated by maximum-likelihood using EM algorithm (Muthén and Shedden, 1999). However, as pointed out in Gilthorpe et al., 2014, EM convergence for GrMMs can be difficult when there are too many freely estimated parameters, over-parametrization may occur and lead to convergence issues. Constraining variance parameters across classes or over time can solve convergence issues but can also bias estimates.
It should be noted that in the literature is also easy to encounter some models called Latent Class Growth Models (LCGMs) Andruff et al., 2009. These are essentially a very restrictive version of a GrMM, with fixed intercept and slope terms. The difference between the two models lies in that GrMM allows for variation across individuals within the same group while LCGMs assumes individuals within groups are homogenous.

Finally, Asparouhov and Muthen, 2008, Muthen and Asparouhov, 2008 also introduce some change in the measurement model to take into account non-continuous outcome variables: instead of linking the growth factors directly to the observed outcome, it is linked to a continuous latent variable, which is in turn linked to the mixed-type observed outcomes through a link function depending on the type of outcomes. This allows to extend the GrMM to possibly mixed-type longitudinal data. Indeed, the most common software to fit this models, the Mplus software (Muthen and Muthen, 2017), describe the possibility of fitting such models also for combinations of outcomes of different types in its user's guide. However, we found that mathematical formulations for mixed-type outcomes in GrMMs are rarely provided in a comprehensive form in the abundant academic literature of the software's creators. Moreover, GrMM, extended for mixed-type data or not, can be viewed as a special case of the model presented in Section 3.6.3. For this reason, we will limit ourselves to treat GrMMs in the context of longitudinal data model only and refer the reader to the cited Section. Cited above, the Mplus software is a licensed statistical software that has the most complete library to fit GrMMs. However, more recently some open source alternatives have been proposed by Wardenaar, 2020.

### 3.5.5 Mixture of Matrix-Normals

Finite Mixture of Matrix-Normals (MMN), as introduced in Viroli, 2011a, can be a useful tool to cluster time-dependent data. For this purpose, we need to slightly change our notation to reorganize our data in a random-matrix form. Let $\mathbf{Y} = \{Y_i\}_{i=1}^N$ be a sample of $J \times T$-variate matrix observations (i.e. $Y_i \in \mathbb{R}^{J \times T}$), arose from studies with $J$-variate vector observations measured repeatedly over $T$ time points, as in a longitudinal study case. Assume that each $Y_i$ follows a

matrix-normal distribution, $Y_i \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Omega)$, 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 $\Omega \in \mathbb{R}^{J\times J}$ is the covariance matrix containing the variance and covariances of the $J$ variables. The matrix-normal pdf is given by:

$$\mathcal{MN}_{(J\times T)}(Y_i|M,\Phi,\Omega) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Omega|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Omega^{-1}(Y_i-M)\Phi^{-1}(Y_i-M)^\intercal]\right\}. \quad (3.5.7)$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Y_i \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Omega)$, then $\text{vec}(Y_i) \sim \mathcal{N}_{JT}(\text{vec}(M), \Phi \otimes \Omega)$, where vec(.) is the vectorization operator and $\otimes$ denotes the Kronecker product. Then, the mean and the variance of the matrix-normal distribution are:

$$\mathbb{E}(\text{vec}(Y_i)|M,\Phi,\Omega) = \text{vec}(M)$$
$$\mathbb{V}(\text{vec}(Y_i)|M,\Phi,\Omega) = \Phi \otimes \Omega$$

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 and Nagar, 2000). The separability condition of the covariance matrix has the twofold advantage of allowing the modeling of the temporal pattern of interest directly on the covariance matrix $\Phi$ and of representing a more parsimonious solution than that of the unrestricted $\Phi \otimes \Omega$.
The pdf of the MMN model is given by

$$f(Y|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi^{(J\times T)}(Y|M_k, \Phi_k, \Omega_k), \quad (3.5.8)$$

where $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k = \{\pi_k, M_k, \Phi_k, \Omega_k\}$. An EM algorithm is used to infer the parameters $\Theta$. The complete likelihood is then given by:

$$\mathcal{L}_C(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{\ell}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k \, \mathcal{MN}_{(J\times T)}(Y_i \mid M_k, \Phi_k, \Sigma_k)\right]^{\ell_{ik}}. \quad (3.5.9)$$

The model was also developed in a Bayesian way in Viroli, 2011b. A matrix-variate regression model was developed in Viroli, 2012.

## 3.6 Literature on longitudinal mixed-type data

Lastly, looking at mixed-type longitudinal data, the main methodology to deal with such data lies in the framework of discrete (time-constant or varying) random effects (Gelman and Hill, 2007). This approach enables to model heterogeneity and to handle multivariate and mixed outcomes, as done by Proust-Lima, Amieva, and Jacqmin-Gadda, 2013 for linear models and by Ram and Grimm, 2009a. These approaches are similar in that they model the change over time at both the population level and the individual level using random effects (or latent variables). In Komarek and Komárková, 2013, Komárek and Komárková, 2014 the authors rely on a multivariate extension of the classical generalized linear mixed model where a mixture distribution is additionally assumed for random

effects. However, only binary or count outcomes are considered other than continuous. Vávra and Komárek, 2021 extend this model to ordinal data. However, nominal variables are not directly taken into account in neither of the papers. Besides, the inference for such over-parametrized models has to be done in a Bayesian fashion, where the specification of priors and the use of associated computational methods can compensate the shortcomings frequentist approaches, especially for insufficient sample sizes. This work is expanded and improved in Vávra, Komárek, Grün, and Malsiner-Walli, 2024. In Cagnone and Viroli, 2018 the authors extended the latent class model to take into account time evolution by means of latent Markov variable (Bartolucci, Farcomeni, and Pennoni, 2014) to model longitudinal binary and ordinal data on alcohol use disorder. In a model-based clustering perspective, De la Cruz-Mesia, Quintana, and Marshall, 2008 proposed a mixture of hierarchical non-linear models for describing non-linear relationships across time. Manrique-Vallier, 2014 introduced a clustering strategy based on a mixed membership framework for analysing discrete multivariate longitudinal data.
Another, maybe simpler, approach consists into using generalized additive models (Hastie and Tibshirani, 1990) and splines (Boor, 2001) to model the longitudinal dimension and assuming conditional independence in a mixture model framework, as done by Grün and Leisch, 2008.

It is finally worth pointing out that latent variable models and random effects models both involve unobserved variables to explain variability in the data, but their purposes and applications differ. Latent variable models introduce unobserved variables to represent abstract or hidden constructs. These models focus on explaining relationships among observed variables by modelling shared influences through the latent variables. Random effects models, on the other hand, are designed for hierarchical or longitudinal data where observations are nested within groups or subjects. Random effects capture subject-specific deviations from population-level trends, such as personalized intercepts and slopes in repeated measures data.

### 3.6.1 Mixture of Generalized Additive Models

Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986, 1990, Wood, 2017) extend Generalized Linear Models (GLMs) (McCullagh, 1989) by incorporating non-linear relationships between predictors and the response variable. While GLMs already provide a flexible framework for modeling data with different types of response variables, GAMs further enhance this flexibility by utilizing smooth functions, typically splines, to capture complex patterns in the data that linear models cannot. Splines (Boor, 2001) are piecewise polynomial functions that are smooth at the points where the pieces connect, known as knots. They provide a flexible way to model non-linear trends over continuous variables, such as time.
The general form of a GAM is given by:

$$h(\mathbb{E}(y)) = \beta_0 + \sum_{i=1}^{p} f_i(x_i)$$

where $h$ is the link function and $f_i(x_i)$ are smooth functions of the predictors $x_i$. The different possible link functions allows to treat mixed-type data. GAMs are extend to perform clustering through mixtures of GAMs, as done in the R package `flexmix` introduced by Grün and Leisch, 2008.

Multivariate variables $y$ are assumed to be dividable into $D$ independent subsets. Each subset represent a specific data-type and it allows each type of response variable to be modelled independently using a distribution that is best suited to its characteristics. In the context of clustering,

a mixture of GAMs is developed. Then, the density for the $k$-th component, $f_k$, is given by a product over $D$ cluster-specific densities, $f_{kd}$ which are defined for the subset variables $y_d$ and their associated covariates $x_d$. The pdf of a mixture of GAMs is then:

$$\mathbb{P}(y \mid x, w, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(w, \alpha) \prod_{d=1}^{D} f_{kd}(y_d \mid x_d, \theta_{kd}) \tag{3.6.1}$$

where $\pi_k(\xi, \alpha)$ is the mixing probability for component $k$, which can depend on concomitant[1] variables $\xi$ and parameters $\alpha$, $f_{kd}(y_d \mid x_d, \theta_{kd})$ is the density function for subset $d$ in component $k$ of parameters $\theta_{kd}$. Within each function $f_{kd}$, splines can be used to model the effect of predictors, such as time, on the response variable. The EM algorithm is used for parameter estimation.

Because of the independence assumption within each component of the mixture, and since the model does not handle random effects, it cannot model direct relationships between two variables of different type. However, the mixing probabilities and the component-specific parameters can indirectly reflect dependencies between variables, as they influence the clustering and the distribution of the data within each component. While mixture of GAMs is able to capture non-linear trajectories when time is used as a covariate, and the generalized framework allows for mixed-type data, it must be noticed that the model becomes easily fairly complex, as it requires careful specification of models for each subset, leading it to be computational demanding on large datasets. More than this, the integration of multiple models and splines can make interpretation challenging, especially for practitioners with weak statistical background.

### 3.6.2 Mixture of Multivariate Generalized Linear Mixed Models

The challenge of clustering and analysing together longitudinal mixed data is addressed by Komarek and Komárková, 2013 by means of the Mixture of Multivariate Generalized Linear Mixed Models (MMGLMM), implemented in the R package `mixAK` (Komárek and Komárková, 2014). This model can be described as a generalized linear mixed model with a normal mixture for the random effects distribution. Thanks to the hierarchical structure of the model induced by random effects, the model design allows the introduction introduce a new notation: for each observed unit $i$, we can define the generic time observation $t_i \in \{1, \ldots, T_i\}$, meaning that for each observation the number and the time of the occasions a variable is measured can vary across subjects. The model first express the conditional mean of each response profile using a standard Generalized Linear Mixed Model (GLMM):

$$\mathbb{E}(y_{ijt_i} | \beta_j, u_{i,j}) = h_j(x_{i,j,t_i}^{\top} \beta_j + s_{i,j,t_i}^{\top} u_{i,j}), \tag{3.6.2}$$

where $h_j$ is the link function used to model the mean of the $j$-th variable. Furthermore, $\beta_j \in \mathbb{R}^{p_j}$ is a vector of unknown regression coefficients (fixed effects) and $u_{i,j} \in \mathbb{R}^{q_j}$ is a vector of random effects for the $j$-th response specific for the $i$-th subject, and $x_{i,j,t_i} \in \mathbb{R}^{p_j}$ and $s_{i,j,t_i} \in \mathbb{R}^{q_j}$ are vectors of known covariates such that $\sum_{j=1}^{J} p_j = p$ and $\sum_{j=1}^{J} q_j = q$, respectively associated to fixed and random effects. Moreover, being within the GLMM framework, we assume that the conditional distribution $\mathbb{P}(y_{ijt} | \theta_j, \beta_j, u_{i,j})$ follows a distribution from the exponential family, with the $\theta_j$ as unknown dispersion parameter and the location parameter for the link function specified as

[1] i.e. covariates that influence the mixing probabilities.

in Equation 3.6.2.

Further, let $u_i = (u_{i,1}^\top, \dots, u_{i,J}^\top)$ be a joint vector of random effects for the $i$-th subject. Dependence between the $J$ longitudinal variables of a particular subject $i$, represented by the response vectors $y_{i,1}, \dots, y_{i,J}$, is taken into account by assuming a joint distribution for the random effect vector $u_i$. It is assumed that the $i$-th subject belongs to one of a fixed number of $K$ clusters, each with a probability $\pi_k = \mathbb{P}(\ell_i = k)$, where $\ell_i \in 1, ..., K$ is the $i$-th subject allocation.
It is further assumed that the random effect vector $u_i$ follows a multivariate normal distribution with cluster-dependent unknown covariance matrix and mean, denoted as $\Sigma_k$ and $\mu_k$, respectively. Let then define $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K\}$, that is the ensemble of unknown parameters related to the distribution of random effects, all cluster dependent. Overall, a multivariate normal mixture in the distribution of random effects is assumed:

$$u_i|\boldsymbol{\theta} \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}_q(\mu_k, \Sigma_k).$$

Therefore, the observed likelihood $\mathcal{L}_O(\boldsymbol{\theta}; \boldsymbol{y})$ is

$$\mathcal{L}_O(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{j=1}^{J} \prod_{t_i=1}^{T_i} \mathbb{P}(y_{i,j,t_i}|\theta_j, \beta_j, u_{i,j}) \sum_{k=1}^{K} \pi_k \mathcal{N}_q(u_i \mid \mu_k, \Sigma_k) du_i. \tag{3.6.3}$$

Principally for computational reasons, the inference of the model is not carried out through an EM algorithm, but through a Bayesian approach based on Markov Chain Monte Carlo (MCMC) approximations, where random effects $u_i$ and allocation variables $\ell_i$ are considered additional parameters, with their distributions and priors. This is the reason why Equation 3.6.3 provides the observed likelihood, which is used to compute the posterior distribution.

In bio-statistical applications, as well as in other research areas, the longitudinal data are typically irregularly sampled, i.e., having in general different values of numbers of visits, which this model is able to handle. The dependence among different variables is induced by non-diagonal covariance matrix $\Sigma_k$ of the random effects vector. The time structure in the model is accounted for in two ways, that is the random slopes, that allow to capture subject-specific rates of change over time and through the covariance matrix $\Sigma_k$, that models the association among random effects for intercepts and slopes. In this model, the random effects serve as latent variables that link the variables: shared random effects ensure that such dependencies are not ignored.

### 3.6.3 Latent Class Linear Mixed Models

The Latent Class Linear Mixed Models (LCLMMs) are an extension of the linear mixed model (Laird and Ware, 1982, Molenberghs and Verbeke, n.d.) that handles heterogeneity through latent classes of trajectory (Proust-Lima, Séne, Taylor, and Jacqmin-Gadda, 2014). Here, we will follow Proust-Lima, Philipps, and Liquet, 2017, whose model is able to handle different types of outcomes, specifically: continuous Gaussian, continuous non-Gaussian or ordinal. As in the previous section, measurement times can vary across individuals, thanks to random effects. This can happen for each unit and for each measured variables, so that $t_{ij}$ will indicate the time for unit $i$ and variables $j$. Membership in these classes is represented by a discrete latent variable $\ell_i \in \{1, \dots, K\}$. A latent process $z_{ijt_{ij}}$, also called structural process, is assumed to represent the underlying trajectory of

the outcome $j$ for each individual $i$ at time $t_{ij}$, whose parameters are however not variable-specific. Then, the relationship between the latent process and each observed variable $j$ is modelled through the measurement model and a link function, conditional to the belonging to a latent class. Hence, the observed value $y_{ijt_{ij}}$, conditional on the latent class $k$, is modelled as:

$$y_{ijt_{ij}}|_{\ell_i=k} = h_j(z_{ijt_{ij}}|_{\ell_i=k} + x_{Yit_{ij}}^\top \alpha_j + b_{ji} + \varepsilon_{ijt_{ij}}\,;\,\eta_j), \tag{3.6.4}$$

where

$$z_{ijt_{ij}}\,|_{\ell_i=k} = x_{L1it_{ij}}^\top \beta + x_{L2it_{ij}}^\top \nu_k + s_{it_{ij}}^\top u_{ik} + w_{it_{ij}} \tag{3.6.5}$$

Here, each variable $j$ is associated to a link function $h_j(\cdot\,;\eta_j)$ where where $\eta_j$ represents the parameters of the link function. In the latent process represented by Equation 3.6.5, the covariates are split between $x_{L1it_{ij}} \in \mathbb{R}^{p_1}$ associated with common fixed effects $\beta \in \mathbb{R}^{p_1}$ shared across the whole population and $x_{L2it_{ij}} \in \mathbb{R}^{p_2}$ associated with class-specific fixed effects $\nu_k \in \mathbb{R}^{p_2}$. In addition, $s_{it_{ij}} \in \mathbb{R}^q$ are covariates associated with random effects $u_{ik} \in \mathbb{R}^q$, which follow $u_{ik} \sim \mathcal{N}_q(0, \sigma_k^2\Sigma)$ where $\Sigma$ is an unspecified variance-covariance matrix and $\sigma_k^2$ is a proportional coefficient ($\sigma_K^2 = 1$ for identifiability) allowing for a class-specific intensity of individual variability, and $w_{it_{ij}}$ is a zero-mean autocorrelated Gaussian stochastic process. Then, the measurement model is given by the latent process plus some covariates $x_{Yit_{ij}} \in \mathbb{R}^{p_3}$ associated with variable specific fixed-effects $\alpha_j \in \mathbb{R}^{p_3}$, where $\sum_{j=1}^J \alpha_j = 0$ and $\alpha_j$ is called contrast, because it captures the differential effect of covariates on each variable relative to the overall mean effect. In addition, a random intercept $b_{ji}$ is considered to capture subject-specific deviations for variable $j$ that are not explained by the covariates, and finally $\varepsilon_{ijt_{ij}} \sim \mathcal{N}(0, \sigma_{\varepsilon_j}^2)$ are Gaussian errors. The error term captures unaccounted variability between the latent process and the observed outcome. In this case, $z_{ijt_{ij}}$ itself does not have an explicit random distribution, as any variability in the process is introduced through the random effects $b_{ik}$, the stochastic process $w_{it_{ij}}$ and the measurement error $\varepsilon_{ijt_{ij}}$ that links $z_{ijt_{ij}}$ to the observed data $y_{ijt_{ij}}$.

The ability to handle different data-types come from the link function $h_j(\cdot)$. For continuous Gaussian data, the link function is assumed be the identity function. For ordinal data, it is a probit cumulative link function such that for an ordinal variable with $C_j$ levels, $P(y_{ijt_{ij}} = c_j \mid u_i;\gamma) = \Phi(\gamma_{c_j-1} - z_{ijt_{ij}}) - \Phi(\gamma_{c_j-1} - z_{ijt_{ij}})$, with $\Phi$ the cumulative distribution function (cdf) of a standard Gaussian variable and the additional parameter $\eta$ is represented by $\gamma$, that is a vector of thresholds. In the paper and in the R package, the model is also made capable of handling continuous non-Gaussian data through the link function, by using functions such as rescaled Beta cdf, I-splines or a linear transformation.

The probability $\pi_{ik} := \mathbb{P}(\ell_i = k)$ of individual $i$ belonging to class $k$ is modelled as:

$$\pi_{ik} = \frac{\exp(\beta_{0k} + \xi_i^\top \beta_k)}{\sum_{k'=1}^K \exp(\beta_{0k'} + \xi_i^\top \beta_{k'})}, \tag{3.6.6}$$

where $\beta_{0k}$ is the intercept, $\beta_k$ are the parameters for covariates $\xi_i$, and for identifiability, $\beta_{0K} = 0$ and $\beta_{1K} = 0$.

As noted in Lu, Ahmadiankalati, and Tan, 2023, LCLMM enhances modelling flexibility by allowing non-linear relationships and serial correlations, e.g., through basis splines or Gaussian

processes. This makes it robust for analysing heterogeneous longitudinal data with multiple features. The difference with a standard linear mixed model is that both fixed effects and the distribution of the random effects can be class-specific. The likelihood of the LCLMM for $N$ subjects is:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_{ik} \int \prod_{j=1}^{J} \prod_{t_{ij}=1}^{T_{ij}} \mathbb{P}(y_{ijt_{ij}} \mid z_{ijt_{ij}}, u_{ik}; \theta_{jk}) \, \mathcal{N}_q(u_{ik}|\theta_{jk}) \, du_{ik}, \tag{3.6.7}$$

where for the sake of brevity $\mathbb{P}(y_{ijt_{ij}} \mid z_{ijt_{ij}}, u_{ik}; \theta_k)$ is the distribution for feature $j$ at time $t_{ij}$ depending on the link function and $\theta_k$ in the ensemble of parameters depending on the class belonging or not.
The inference process is carried out through maximum likelihood estimation. More precisely, an iterative Marquardt algorithm, which belongs to the Newton-Raphson family (Proust and Jacqmin-Gadda, 2005), is used to maximize the likelihood in Equation 3.6.7.
Lastly, as already said in Section 3.5.4, GrMM, also extended to mixed-type data, can be viewed as a special case of LCLMM, which extends this framework to accommodate more complex data structures and outcomes. Both models incorporate random effects to account for within-class variability, but LCLMM offers greater flexibility in modelling diverse longitudinal patterns and outcomes.

### 3.6.4 Bayesian Consensus Clustering for longitudinal data

The Bayesian Consensus Clustering model for longitudinal data (BCClong) (Lu and Lou, 2022), is the Bayesian extension of the consensus clustering approach. The authors of the paper introducing the model developed an R package as well, the `BCClong` library (Tan, Shen, and Lu, 2022).
In its simplest form, Consensus Clustering (CC) (Monti, Tamayo, Mesirov, and Golub, 2003) involves repeated sub-sampling and clustering of data to evaluate the stability of clusters. The consensus matrix records the proportion of times pairs of items are grouped together, providing a quantitative measure of cluster stability. Given $M$ clustering results with connectivity matrices $C^{(m)}$, where $C^{(m)}_{ii'} = 1$ if units $i$ and $i'$ are clustered together in the $m$-th clustering, the consensus matrix $\bar{C}$ is calculated as:

$$\bar{C} = \frac{1}{M} \sum_{m=1}^{M} C^{(m)}$$

Each entry $\bar{C}_{ij}$ indicates the proportion of times units $i$ and $i'$ are grouped together.
However, Bayesian methodology allows this to be achieved through two interconnected components: local clustering and global clustering. Local clustering focuses on identifying patterns within individual variables, allowing each variable to reveal its unique trajectories and subgroups. For this model, measurement times can vary across individuals and across variables, so that $t_{ij} \in \{1, \ldots, T_{ij}\}$ will indicate the time for unit $i$ and variables $j$. Let $y_{ij} = (y_{ij1}, \ldots, y_{ijT_{ij}})$ denote the vector of all the measurement times for the unit $i$ and variable $j$. Let $\ell_{ij} \in \{1, \ldots, K\}$ represents the local cluster label for subject $i$ and variable $j$, and $\ell_i \in \{1, \ldots, K\}$ represents the overall (global) cluster label for subject $i$. Note that the local and the global clusters are the same. For the sake of simplicity, we will define $k_j$ the generic local cluster allocation for variable $j$. The joint distribution of

the $J$ features for individual $i$ is modelled as:

$$\mathbb{P}\left(y_{i1},\ldots,y_{iJ}\right)=\sum_{k_1,\ldots,k_J}\pi_{k_1,\ldots,k_J}\mathbb{P}\left(y_{i1},\ldots,y_{iJ}\mid \ell_{i1}=k_1,\ldots,\ell_{iJ}=k_J\right) \tag{3.6.8}$$

where $\pi_{k_1,\ldots,k_J}=\mathbb{P}\left(\ell_{i1}=k_1,\ldots,\ell_{iJ}=k_J\right)$ and $\mathbb{P}\left(y_{i1},\ldots,y_{iJ}\mid \ell_{i1}=k_1,\ldots,\ell_{iJ}=k_J\right)$ will be addressed in the following. The probability of a subject belonging to a particular local cluster $k_j$ for variable $j$ is given by:

$$\mathbb{P}(\ell_{ij}=k_j\mid y_{ij},\ell_i,\beta_{k_j},u_{ik_j})\propto a(k_j,\ell_i,\alpha_j)f_{k_j}(y_{ij}\mid\beta_{k_j},u_{ik_j}) \tag{3.6.9}$$

Here, $a(k_j,\ell_i,\alpha_j)$ is the dependence function adjusted by the adherence parameter $\alpha_j$. It represents the relationship between the local cluster labels and the global cluster labels, specifically represent the degree to which the local cluster labels $\ell_{ik}$ adhere to the global cluster label $\ell_i$, measure by the adherence parameter $\alpha_j$. It indicate the contribution of each variable to the global clustering. The function $f_{k_j}$ represents the distribution from the exponential family with a mean function:

$$h_{k_j}^{-1}(\mathbb{E}(y_{ij}\mid\beta_{k,j},u_{ik_j}))=x_{ij}^{\top}\beta_{kj}+s_{ij}^{\top}u_{ikj}, \tag{3.6.10}$$

where $h_{k_j}^{-1}$ is the inverse link function, $x_{ij}\in\mathbb{R}^{p_j}$ is a vector of predictor associated with fixed effect $\beta_{k_j}\in\mathbb{R}^{p_j}$ and $s_{ij}\in\mathbb{R}^{q_j}$ is a vector of predictors associated with random effects $u_{ik_j}\in\mathbb{R}^{q_j}$, where $u_{ik_j}\sim\mathcal{N}_{q_j}(\mathbf{0},\Sigma_{k_j})$.

Global clustering, on the other hand, seeks to integrate these local insights into an overarching clustering structure that reflects the combined influence of all variables. The joint distribution of local clusterings is expressed in terms of the global clustering:

$$\mathbb{P}(\ell_{i1},\ldots,\ell_{iJ})=\sum_{k=1}^{K}\pi_k\,\mathbb{P}(\ell_{i1},\ldots,\ell_{iJ}\mid\ell_i=k), \tag{3.6.11}$$

where $\pi_k:=\mathbb{P}(\ell_i=k)$. Finally, the complete likelihood can be written as:

$$\mathcal{L}\left(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{x},\boldsymbol{s}\right)=\prod_{i=1}^{N}\sum_{k_1,\ldots,k_J}\left(\sum_k\pi_k\prod_{j=1}^{J}a(\ell_{ij}=k_j,\ell_i=k,\alpha_j)\right)\prod_{j=1}^{J}f_{k_j}(y_{ij}\mid\beta_{k_j},u_{ik_j};x_{ij},s_{ij}). \tag{3.6.12}$$

As one can notice by Equation 3.6.12, the BCClong model assumes conditional independence given the local and the global cluster labels. However, it still captures dependencies among the variables through the cluster labels themselves. The trajectories of each longitudinal variable are modelled thanks to to the linear predictor which includes both fixed effects and and random effects. The inference is performed in a Bayesian fashion, and the posterior computation is carried out using a Gibbs sampling scheme coupled with the Metropolis-Hastings algorithm. The Bayesian formulation and inference naturally leads to computationally intensive inference, along with the model's complexity in specifying priors and ensuring convergence. Moreover, the assumption of conditional independence among variables given cluster labels may oversimplify data relationships.

## 3.7 Softwares

To facilitate the implementation of the models discussed in this chapter, we collect in this section the several software packages that are available and that we presented throughout the survey. Below is a summary of the key software packages, their associated models, and the primary functions used to apply these models. The table is divided into three categories: mixed-type data, longitudinal data, and longitudinal mixed-type data, to help readers identify the most relevant tools for their specific needs.

When used in R, the process involves specifying separate models for each type of response variable using the `FLXMRmgcv` function, which incorporates GAMs and splines. The mixture model is then fitted using the `flexmix` function, which estimates the parameters and cluster assignments.

### 3.7.1 Mixed-Type Data

| **Software** | **Associated Model** | **Key Functions** |
|---|---|---|
| clustMD | clustMD | `clustMD()` |
| mixedClust | Multiple Latent Block Model | `mixedCoclust()` |
| lcmm | Latent Class Model | `hlme()` |

### 3.7.2 Longitudinal Data

| **Software** | **Associated Models** | **Key Functions** |
|---|---|---|
| LMest | Latent Markov Model for categorical and continuos outcomes, Mixed Hidden Markov Model. | `lmest()`,`lmestCont()`, `lmestMixed()` |
| lcmm | Growth Mixture Model. | `lcmm()`, with time as covariate |
| MatTransMix | Mixture of Matrix-Normals. | `MatTrans.EM()` |

### 3.7.3 Longitudinal Mixed-Type Data

| Software | Associated Models | Key Functions |
|---|---|---|
| flexmix | Mixture of Generalized Additive Model. | `FLXMRmgcv()`, `flexmix()` |
| mixAK | Mixtures of Multivariate Generalized Linear Mixed Models. | `GLMM_MCMC()` |
| BCClong | Bayesian Consensus Clustering for longitudinal data. | `BCC.multi()` |
| lcmm | Latent Class Linear Mixed Models. | `multlcmm()` |
| longmixr | Consensus Clustering with mixtures of GAMs [2]. | `longmixr()` |

## 3.8 Conclusions

This survey has provided an overview of the main methodologies for model-based clustering of longitudinal mixed-type data, highlighting the challenges and solutions proposed in the literature. The analysis of such data is complex due to the simultaneous presence of temporal dynamics and mixed data types, which often leads to biases or information loss when treated independently or transformed into continuous forms. The ideal approach involves models that can jointly address both the temporal evolution and the heterogeneity of data types, but the literature on this topic remains relatively limited.

Traditional clustering methods often fail to capture the dependencies between observations over time, leading to suboptimal results or suffer from over-parametrization and computational complexity. For instance, Latent Markov Models (LMMs) and their extensions, like Mixed Hidden Markov Models (MHMMs), offer a way to model temporal evolution through latent states. However, LMMs assumes trajectory homogeneity across individuals, which may not hold in heterogeneous populations. MHMMs address this by introducing a further latent variable to account for individuals or groups-specific trajectories, but this increases computational demands and interpretative complexity.
The other significant issue is the handling of mixed-type data. Models like the Latent Class Model (LCM) and clustMD assume conditional independence between variables of different types given the cluster membership, which simplifies the modelling process but may overlook important dependencies between variables. Similarly, the Multiple Latent Block Model (MLBM), while providing a flexible co-clustering at the cost of increased computational demand, does not allow clusters of variables to be of different type.

[2] Not described above.

The integration of longitudinal and mixed-type together data poses further challenges. Models like the Mixture of Multivariate Generalized Linear Mixed Models (MMGLMM) and Latent Class Linear Mixed Models (LCLMMs) offer promising solutions by combining random effects and latent variables to capture both temporal and mixed-type dependencies. However, these models often require complex inference methods, which can be computationally intensive. Additionally, the interpretation of results from such complex models remains a challenge for practitioners without a strong statistical background, as the relationships between variables and clusters may not be straightforward.
Software availability is another critical aspect. While several R packages provide implementations of numerous models, they often focus on specific types of data or require advanced statistical knowledge for effective use. The development of more user-friendly and comprehensive software tools could significantly enhance the accessibility and applicability of these models in practical settings.

In conclusion, while significant progress has been made in developing models for longitudinal mixed-type data in the last decade, there may be still a need for more parsimonious, interpretable, and computationally efficient solutions. Future research should focus on addressing the issues of over-parametrization, computational complexity, and the interpretation of results. Additionally, the development of user-friendly software tools could greatly enhance the adoption of these models in various fields, enabling practitioners to effectively include clustering longitudinal mixed-type data in their workflow.

# Chapter 4

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

This chapter was published in April 2024 in the *Statistics and Computing* journal, volume 34 (Amato, Jacques, and Prim-Allaz, 2024). We have reproduced the entire article as published, expect for some small changes to notation to keep it consistent with the rest of the thesis. For this reason, some concepts may be repeated, particularly concerning Section 4.3.2 with regard to Chapter 2 and Sections 4.2.1 and 4.1.1 to Chapter 3. The paper addresses the problem of clustering longitudinal ordinal data by introducing the Mixture of Ordinal Matrices (MOM) model, and represents the first extension of the mixture of matrix-normals model to non-continuos data.

---

**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 an 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, and approaches to deal with some arising computational challenges are outlined. 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.

---

## 4.1 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, Madden, and Firtle, 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-Gélie, 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, Celeux, Murphy, and Raftery, 2019b).

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.

### 4.1.1 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 and Magidson, 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 and Kruschke, 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 and Zavoina, 1975, Winship and Mare, 1984, Becker and Kennedy, 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 and Piccolo, 2005, that introduces the

CUB model, later developed through the R package `CUB` (Iannario and Piccolo, 2016), to Giordan and Diana, 2011 and more recently Ranalli and Rocci, 2016; Fernandez, Arnold, and Pledger, 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 and Jacques, 2016 and extended for co-clustering by Jacques and Biernacki, 2018. A mixture of item response models was developed to for ordinal response data in the Bayesian framework by McParland and Gormley, 2013, to be later expanded in the frequentist paradigm and to handle mixed data in McParland and Gormley, 2016. More recently, Corneli, Bouveyron, and Latouche, 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 and Murphy, 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 and Viroli, 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` (Komárek and Komárková, 2014) the basis for clustering is a mixture of multivariate generalized linear mixed models. In Vávra and Komárek, 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 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, 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 MMN is provided in §2.1 of Wang and Melnykov, 2020). Anderlucci and Viroli, 2015 extends on the work of McNicholas and Murphy, 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, Bulut, and Arslan, 2016, Gallaugher and McNicholas, 2018 and Melnykov and Zhu, 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, Zhu, Melnykov, and Ingrassia, 2020a). A new comprehensive R package to apply this family to clustering continuous three-way data (Zhu, Sarkar, and Melnykov, 2022) has been proposed, endeavoring the creation of a `mclust` (Scrucca, Fop, Murphy, and Raftery, 2016) for three-way continuous data.

### 4.1.2 The 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 and Viroli, 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 4.2 and 4.3 we will detail our model and the EM algorithm to perform inference. In Section 4.4 the results on synthetic data are presented to assess the performance of the model. Finally, in Section 4.5 an application on real data concerning grocery shopping preferences by a French sample during the Covid-19 pandemic period is outlined.

## 4.2 Model

### 4.2.1 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

$$f(Z|M,\Phi,\Sigma) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^\intercal]\right\}. \tag{4.2.1}$$

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 $\text{vec}(Z) \sim \mathcal{N}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$, 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:

$$\mathbb{E}(\text{vec}(Z)|M,\Phi,\Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M,\Phi,\Sigma) = \Sigma \otimes \Phi. \tag{4.2.2}$$

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 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 $\Phi$. 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

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

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

### 4.2.2 The Mixture of Ordinal Matrices model

Let denote by $y_{ijt}$ 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 $Y_i = (y_{ijt}) \in \mathbb{N}^{J\times T}$. 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 $y_{ijt}$ is the manifestation of an underlying latent continuous variable $z_{ijt}$ which follows a Gaussian distribution, as done in the clustMD model (McParland and Gormley, 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 $\gamma_j$ 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 $z_{ijt}$ is such that $\gamma_{j,c-1} < z_{ijt} < \gamma_{j,c}$ then the observed ordinal response, $y_{ijt} = c$.

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, $\ell_i = (\ell_{i1}, \ldots, \ell_{iK})$, 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 all possible ordinal matrices of size $J \times T$ whose general row $j$ takes values in $\{1, \ldots, C_j\}$. Each element of $\mathcal{O}^{J\times T}$ is called a response pattern, that is each element of the set represents one of the possible configuration (pattern) of the $J \times T$ ordinal matrix, given the levels $C_j$. 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 $\Omega_r$ of the latent space $\mathbb{R}^{J\times T}$ according to thresholds $\boldsymbol{\gamma} := \{\gamma_j\}_{j=1}^J$. Let the binary vector $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iR})$ 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. 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:

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

we get:

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

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 $\xi_i$ 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 $Z_i, \ell_i$ and of the parameters, respectively. Finally, let $\tilde{\mathbf{Y}} := \{\tilde{Y}_i\}_{i=1}^N$ be the collection of the observed response pattern vectors $Y_i$.

## 4.3 Inference

### 4.3.1 Thresholds

#### Identifiability

A key point is of course the choice of the thresholds $\boldsymbol{\gamma}$. Imagine to observe a sample of ordinal categories $c = 1, \ldots, C_j$ for variable $j$ and to work in the same framework as Section 4.2. Let consider each variable separately in an univariate case for the sake of simplicity. Then, assume that each observation derive from the discretization of an underlying continuous variable following a normal distribution with parameters $(\mu_j, \sigma_j^2)$, and consider the $C_j - 1$ dimensional thresholds vector $\gamma_j$ 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 identifiable 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 $(p_{\theta_1} = p_{\theta_2})$.

This simple example shows that we cannot aim at estimating the thresholds and the latent distribution parameters at the same time without incurring in some identifiability issues. Different strategies come to mind to overcome this problem.
Indeed, one solutions is to fix either the thresholds or the parameters $\boldsymbol{\Theta}$. In our case, being clearly the parameters of the mixture the quantity of interest, we decided to fix the thresholds as outlined in Section 4.3.1. However, it is also possible to go for a "mixed strategy", partially fixing some of the distribution parameters and of the thresholds, to then estimate the rest, as done in as done in Millsap and Yun-Tein, 2004.

#### Choice of thresholds

As written in Section 4.1.1, assuming underlying continuous variables categorized according to some thresholds is not new and there are several ways of specifying such thresholds.
In McParland and Gormley, 2016 the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^{J}$ are fixed 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 is less than or equal to level $c$ and $\varphi$ is the standard normal cumulative distribution function. With this assumption, the ordinal distribution of clusters will have the same global shape, not necessarily uni-modal, which makes clusters interpretation harder.

On the other hand, in Corneli, Bouveyron, and Latouche, 2020 thresholds are fixed arbitrarily (keeping equidistant the classes) as $\gamma_j = (1.5, 2.5, \dots, C_j - 0.5)$ and $C_j$ is assumed to be equal for all variables, proposing a scale conversion pre-processing algorithm (Gilula, McCulloch, Ritov, and Urminsky, 2019) for cases when this does not hold true. The advantages of such an approach is that an 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.
Finally, it is also worth noting that the thresholds are fixed and do not change over time.

### 4.3.2 EM-algorithm

The EM algorithm (Dempster, Laird, and Rubin, 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.

### 4.3.3 Complete Likelihood

The complete log-likelihood can be written as

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

where for $\tilde{Y}_{ir}\log(\mathbf{1}_{\Omega_r}(Z_i))$ we can apply the convention, frequently used in computer science and information theory, according to which $0\log(0) = 0$, to avoid the indeterminateness of $\log(0)$.

### 4.3.4 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:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(s)}) := \mathbb{E}(\log \mathcal{L}_C(\boldsymbol{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \boldsymbol{\ell})|\hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) =$$

$$\mathbb{E}\left(\sum_{i=1}^{N}\left\{\sum_{r=1}^{R}\tilde{Y}_{ir}\mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K}\ell_{ik}\left[\log(\hat{\pi}_k^{(s)}) - \frac{TJ}{2}\log(2\pi) - \frac{J}{2}\log(|\hat{\Phi}_k^{(s)}|) - \frac{T}{2}\log(|\hat{\Sigma}_k^{(s)}|) - \frac{1}{2}\mathrm{tr}[\hat{\Sigma}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)}) \times \hat{\Phi}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})^\intercal]\right]\right\}\middle|\hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}\right) = \tag{4.3.1}$$

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\tilde{Y}_{ir}\,\mathbb{E}(\mathbf{1}_{\Omega_r}(Z_i)|\hat{\boldsymbol{\pi}}^{(s)}, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) + \tag{4.3.2}$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}(\ell_{ik}|\hat{\boldsymbol{\pi}}^{(s)}, \hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) \times \left[\log(\hat{\pi}_k^{(s)}) - \frac{TJ}{2} \times \log(2\pi) - \frac{J}{2}\log(|\hat{\Phi}_k^{(s)}|) - \frac{T}{2}\log(|\hat{\Sigma}_k^{(s)}|)\right] - \tag{4.3.3}$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{1}{2}\,\mathbb{E}(\ell_{ik}\mathrm{tr}[\hat{\Sigma}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)}) \times \hat{\Phi}_k^{-1(s)}(Z_i - \hat{M}_k^{(s)})^\intercal]|\hat{\boldsymbol{\Theta}}^{(s)}, \tilde{\mathbf{Y}}) \tag{4.3.4}$$

We can treat each of the three expectations separately, and we get for (4.3.2)

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

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

For (4.3.3), we can write

$$\begin{aligned}\mathbb{E}(\ell_{ik}|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) &= \mathbb{P}(\ell_{ik}=1|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) \\ &= \frac{\mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)})\mathbb{P}(Y_{ir}^{R}=1|\ell_{ik}1,\hat{\boldsymbol{\Theta}}^{(s)})}{\mathbb{P}(Y_{ir}^{R}=1|\hat{\boldsymbol{\Theta}}^{(s)})} \\ &= \frac{\pi_k^{(s)}\int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ}{\sum_{k=1}^{K}\pi_k^{(s)}\int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ} =: \boldsymbol{\tau}_{ik}^{(s+1)},\end{aligned} \tag{4.3.5}$$

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, (4.3.4) is less straightforward, and we will need some tricks to deal with it. As done in McParland and Gormley, 2016, we can break down as

$$\begin{aligned}&\mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)},\tilde{\mathbf{Y}})\times \\ &\mathbb{E}(tr[\hat{\Sigma}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})\times\hat{\Phi}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})^{\intercal}]|\ell_{ik}=1,\tilde{\mathbf{Y}},\hat{\boldsymbol{\Theta}}^{(s)}).\end{aligned} \tag{4.3.6}$$

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

$$\begin{aligned}&\hat{\Sigma}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})^{\intercal} = \\ &\hat{\Sigma}_k^{-1(s)}Z_i\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}-\hat{\Sigma}_k^{-1(s)}Z_i\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal}-\hat{\Sigma}_k^{-1(s)}\hat{M}_k^{(s)}\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}+\hat{\Sigma}_k^{-1(s)}\hat{M}_k\hat{\Phi}_k^{-1(s)}\hat{M}_k^{\intercal}.\end{aligned} \tag{4.3.7}$$

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 see in Section 4.3.5, 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

$$\mathbb{E}(z_i|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: m_{ik}^{(s+1)} \tag{4.3.8}$$

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 $S_{ik}^{(s+1)} := \mathbb{E}(z_iz_i^{\intercal}|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)})\in\mathbb{R}^{JT\times JT}$, 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 $\mathbb{E}(Z_i\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1)$ by computing it element-by-element. In order to do that, we can define $D_{ik}^{(s+1)} := \mathbb{E}(Z_i\hat{\Phi}_k^{-1(s)}Z_i^{\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=$

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

$$\begin{aligned}
D_{ik}^{(s+1)} &:= \mathbb{E}(Z_i\hat{\Phi}_k^{-1(s)}Z_i^\intercal|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1)) \\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T}\sum_{g=1}^{T} z_{i,hg}\hat{\varphi}_{k,gd}^{(s)}z_{i,td}\Big)_{h,t}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1\right) \\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{T}\sum_{g=1}^{T} z_{i,hg}z_{i,td}\hat{\varphi}_{k,gd}^{(s)}\Big)_{h,t}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1\right) \\
&= \Big(\sum_{d=1}^{T}\sum_{g=1}^{T} S_{ik,[(g-1)J+h,(d-1)J+t]}^{(s+1)}\hat{\varphi}_{k,gd}^{(s)}\Big)_{h,t}, \qquad (4.3.9)
\end{aligned}$$

where in 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

$$\begin{aligned}
C_{ik}^{(s+1)} &:= \mathbb{E}(Z_i^\intercal\hat{\Sigma}_k^{-1(s+1)}Z_i|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1) \\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{J}\sum_{g=1}^{J} z_{i,gh}\hat{\sigma}_{k,gd}^{(s+1)}z_{i,dt}\Big)_{h,t}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1\right) \\
&= \mathbb{E}\left(\Big(\sum_{d=1}^{J}\sum_{g=1}^{J} z_{i,gh}z_{i,dt}\hat{\sigma}_{k,gd}^{(s+1)}\Big)_{h,t}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\tilde{Y}_{ir}=1\right) \\
&= \Big(\sum_{d=1}^{T}\sum_{g=1}^{T} S_{ik,[(h-1)J+g,(t-1)J+d]}^{(s+1)}\hat{\sigma}_{k,gd}^{(s+1)}\Big)_{h,t}. \qquad (4.3.10)
\end{aligned}$$

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

- $\mathbb{E}(\ell_{ik}|\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: \tau_{ik}^{(s+1)}$,
- $\mathbb{E}(z_i|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: m_{ik}^{(s+1)}$,
- $\mathbb{E}(z_iz_i^\intercal|\ell_{ik}=1,\tilde{Y}_{ir}=1,\hat{\boldsymbol{\Theta}}^{(s)}) =: S_{ik}^{(s+1)}$, whose elements are required for the computation of $D_{ik}^{(s+1)}$ and $C_{ik}^{(s+1)}$.

### 4.3.5 M-step

By taking the first derivatives of Equation 4.3.1, the maximum likelihood estimators of the parameters are given by

$$\hat{\pi}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}}{N} \tag{4.3.11}$$

$$\hat{M}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} \hat{M}_{ik}^{(s+1)}}{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}} \tag{4.3.12}$$

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 $D_{ik}^{(s+1)}$. The updating of the covariance matrices is obtained through:

$$\hat{\Sigma}_k^{(s+1)} = \frac{\sum_{i=1}^N \tau_{ik}^{(s+1)} [D_{ik}^{(s+1)} - \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} M_{ik}^{\intercal(s+1)} - M_{ik}^{(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\intercal(s+1)} + \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\intercal(s+1)}]}{T \sum_{i=1}^N \tau_{ik}^{(s+1)}}, \tag{4.3.13}$$

$$\hat{\Phi}_k^{(s+1)} = \frac{\sum_{i=1}^N \tau_{ik}^{(s+1)} [C_{ik}^{(s+1)} - \hat{M}_k^{\intercal(s+1)} \hat{\Sigma}_k^{-1(s+1)} M_{ik}^{(s+1)} - M_{ik}^{\intercal(s+1)} \Sigma_k^{-1(s+1)} \hat{M}_k^{(s+1)} + \hat{M}_k^{\intercal(s+1)} \hat{\Sigma}_k^{-1(s+1)} \hat{M}_k^{(s+1)}]}{J \sum_{i=1}^N \tau_{ik}^{(s+1)}}. \tag{4.3.14}$$

It is worth to remark that the computation of $\hat{\Sigma}_k^{(s+1)}$ and $\hat{\Phi}_k^{(s+1)}$ relies on $D_{ik}^{(s+1)}$ and $C_{ik}^{(s+1)}$, respectively. The two quantities in turn rely on the elements of $\hat{\Phi}_k^{(s)}$ and $\hat{\Sigma}_k^{(s+1)}$, as shown in Equation 4.3.9 and Equation 4.3.10. This means that in the algorithm one needs to compute first $D_{ik}^{(s+1)}$, then $\hat{\Sigma}_k^{(s+1)}$, and $C_{ik}^{(s+1)}$ and $\hat{\Phi}_k^{(s+1)}$ subsequently. The updating order of the parameters can be exchanged, but it is important to use the updated parameters coherently.

### 4.3.6 Initialization

To find the initial values of $\hat{\boldsymbol{\Theta}}^{(0)}$ mentioned in Section 4.3.2, our proposal is the following. Identity matrices are chosen for the initialization of the covariance matrices $\Phi_k$ and $\Sigma_k$, while $\pi_k = 1/K$. For the initialization of $M_k$, two solutions are proposed and tested in Section 4.4.2. The first is a Kmeans++ (Arthur and Vassilvitskii, 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 4.4.2, 5 random initialization proved to be enough, but for more complex setting a higher number might be needed.

### 4.3.7 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 $\nu_k$ is the total number of model parameters:

$$\nu_k := k[1 + JT + J(J+1)/2 + T(T+1)/2] - 1, \tag{4.3.15}$$

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

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

To select the model with the optimal $K$, the algorithm needs to be executed for every $k = 1, ..., K$ and the model with the lowest $\text{BIC}_k$ is chosen.

### 4.3.8 Classification

Finally, a criterion for the classification of the units must be established. The criterion we use is the maximum conditional allocation probability. Defining with the superscript $c$ the step at which the convergence has been reached or the maximum number of iterations attained, the observation $i$ will be allocated to the cluster $h = \arg\max_h \tau_{ih}^{(c)}$.

## 4.4 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.
The algorithm has been implemented in R.

### 4.4.1 Simulation Setup

100 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 4.3.1.Concerning the distributions' parameters, identity matrices were chosen for matrices $\Phi_k$ and $\Sigma_k$ for every cluster, while the mean matrices $M_k$ were selected so that there would be a partial overlap among the clusters, in order to avoid triviality. However, estimating theoretically the overlapping area in such a setting is complex endeavour. That is why we evaluate an approximated "optimal" Adjusted Rand Index (ARI) (Rand, 1971), by comparing the classification obtained using the true model parameters with the known groups. Thus, the mean matrices $M_k$ are chosen so that this estimated optimal ARI would be around 0.85. Note that we would expect the study to show convergence to this number as the sample size increases. This setting led to the choice of $M_1 = 1.75 \cdot \mathbf{1}_5\mathbf{1}_5^\top$, $M_2 = 2.5 \cdot \mathbf{1}_5\mathbf{1}_5^\top$ 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 proportion $\tau$ of units using a uniform distribution on levels $C_j$, allocated to the three clusters proportionally to the clusters' size: 0 (scenario 1), 0.1 (scenario 2), 0.2 (scenario

3).
The two different kinds of initialization described in Section 4.3.6 have been tested.

Finally, we use a difference between observed log-likelihood at step $(s+1)$ and $(s)$ as stopping criterion, setting this difference to be lower than 0.001 as stopping rule.

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 seconds for $N = 3000$.

### 4.4.2 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 5.1 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 ($\approx 0.85$) 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, probably steaming from the smaller sample size.

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 $\theta$ it is expressed through the formula:

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

where $\hat{\theta}_i$ is the estimated parameter and $\theta_i$ 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 5.2. There seems not to be a clear difference between the two initializations.
Concerning 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.

Globally, there not seems to be a significant difference in terms of performance results for the two initialization procedures regarding the partitioning capabilities. The only biggest difference seems to be the slightly lower variability of the estimates produced by the random initialization. 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

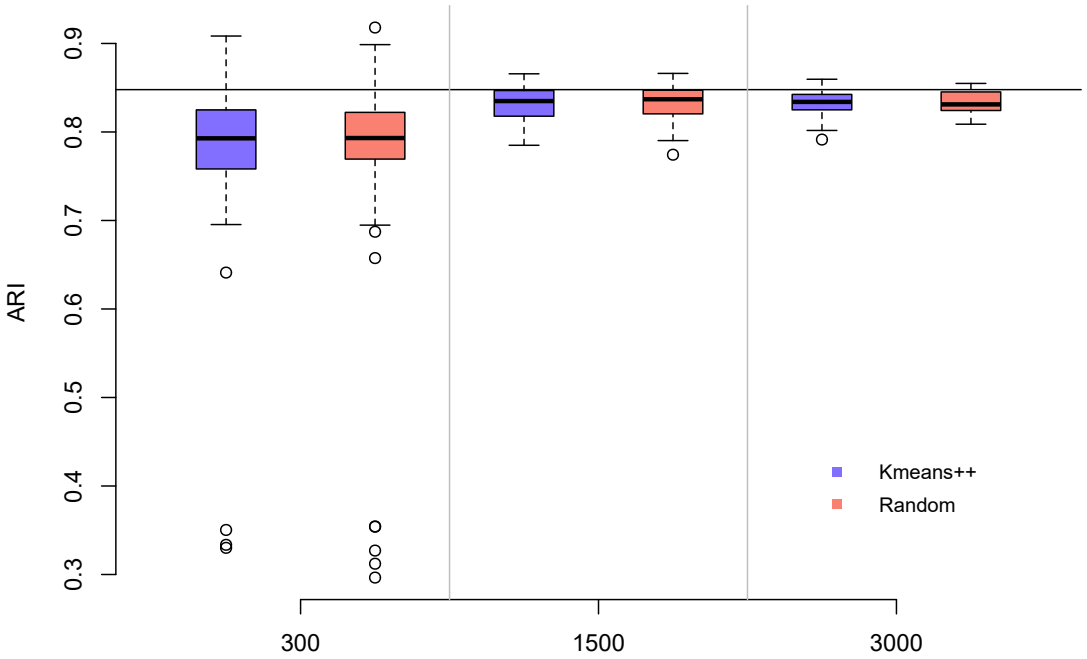



Figure 4.1: Influence of initialization. The horizontal line represents the estimated optimal ARI.

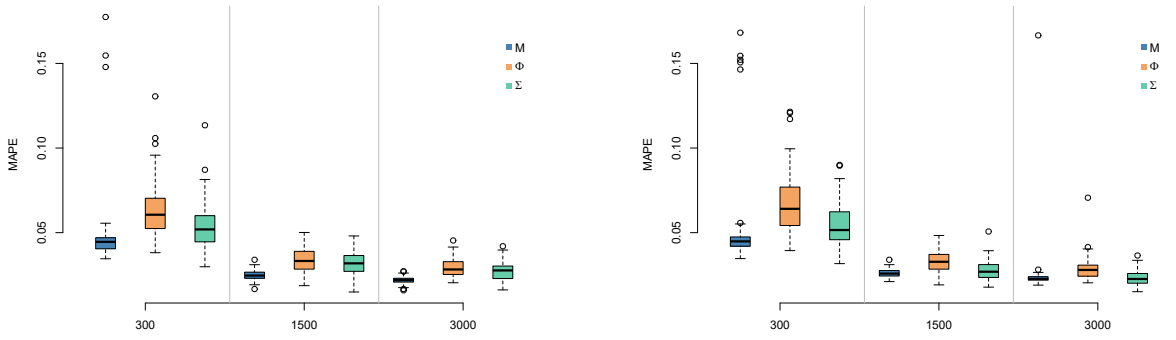
time-consuming and computationally costly.
In the following, given the similarities in performance and the computational advantages, we will carry out most of the analysis using only the Kmeans++ initialization.

### 4.4.3 Robustness to noise

As written in Section 4.4.1, we also simulated some noisy data to study the behaviour of MOM when the underlying normality assumption is not fully respected. ARIs for different noise proportions were measured and the results are visible in Figure 5.3. 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 horizontal black line in the graph. For $N = 300$, the noise disturbs the model estimate, and we do not get an ARI as close to the optimal one as for bigger samples, but still overall better for non-noisy data. The clustering of matrix-normally




(a) Kmeans++ init

(b) Random init.

Figure 4.2: MAPE for increasing N

distributed data therefore seems a bit disturbed by noise when $N$ is small, but it corrects when $N$ increases. This may be due to the fewer non-noisy units left to the model to infer the parameters from.

### 4.4.4 Model selection

Following the setup described in Section 4.4.1, 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 $N$, 100 data sets have been drawn. Model selection has been performed through BIC, as described in Section 4.3.7. The results are shown in Table 5.1.

For $N = 300$, all the simulated data sets yield 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. In this context, it is worth recalling that the BIC is asymptotically consistent. Therefore, one may not be surprised to the fact that for small sample sizes it encounters some issues in selecting the true model.

| 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 | 100 | 0 | 0 | 0 | 0 | 300 | 0 | 100 | 0 | 0 | 0 | 0 | 300 | 0 | 100 | 0 | 0 | 0 | 0 |
| 1500 | 0 | 0 | 100 | 0 | 0 | 0 | 1500 | 0 | 0 | 100 | 0 | 0 | 0 | 1500 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3000 | 0 | 0 | 100 | 0 | 0 | 0 | 3000 | 0 | 0 | 100 | 0 | 0 | 0 | 3000 | 0 | 0 | 100 | 0 | 0 | 0 |

Table 4.1: Frequency of selection of each model K by MOM through BIC among the 100 simulated data sets, for increasing N. The actual value for K is 3. Kmeans++ initialization.

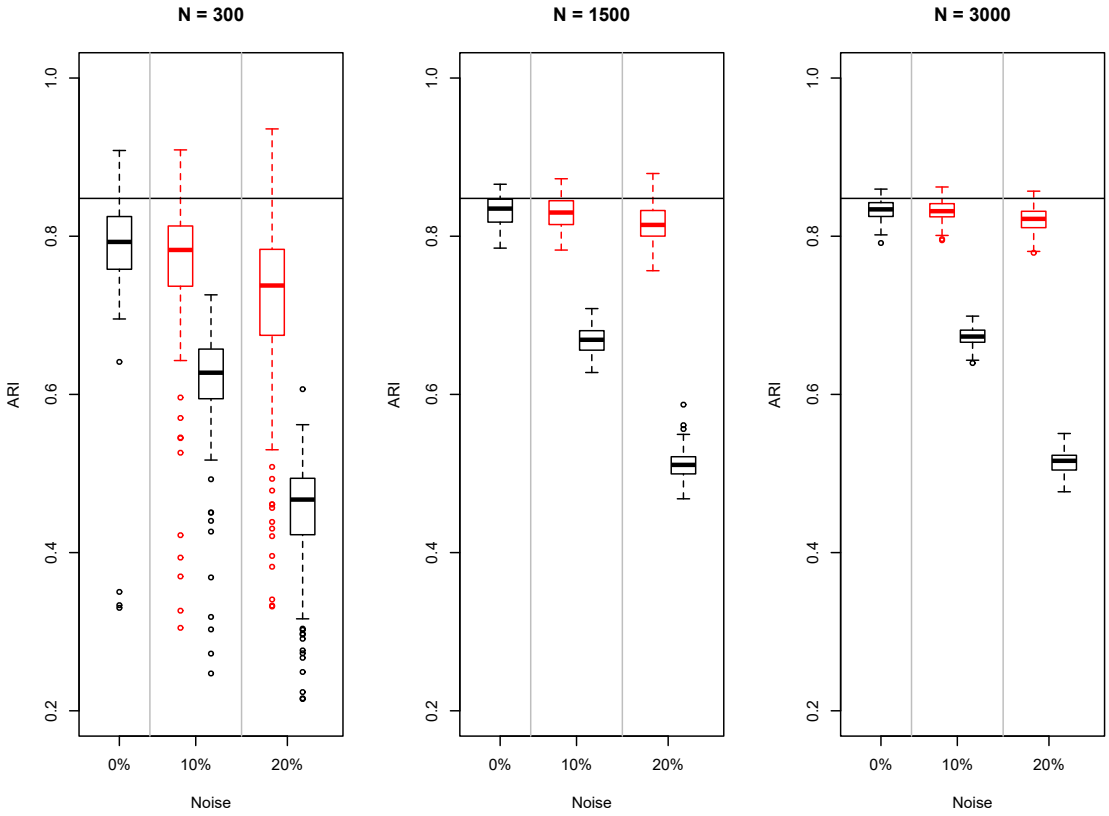



Figure 4.3: ARI for increasing noise proportions and increasing N.
The red (left) box plots is for non-noisy units (0.1 and 0.2 of noise), the black (right) for all units.

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.

### 4.4.5 Comparison with competitors

Finally, we compared the results obtained for the MOM model to the ones given by its continuous version, the Mixture of Matrix-Normals (MMN) (Viroli, 2011a), mentioned in Section 4.1.1, by treating our ordinal data as continuous ones, as frequently done by practitioners. Moreover, we compared our model against a plain mixture of multivariate normal distributions as well, applied on the vectorized version of the data. To do so, we used the R package `mclust` (Scrucca, Fop,

Murphy, and Raftery, 2016).
The hyper-parameters of the competitors have been set to be similar to the one of the MOM in terms of convergence and covariance matrix parametrization. Hence, in both cases the stopping rule is given by the absolute difference of two consecutive log-likelihoods being less than $1 \times 10^{-3}$ and the two covariance matrices for MMN and the single one for `mclust` are fully parametrized. Moreover, we think it is worth mentioning that we tried to perfom the comparison also with the package `clustMD`, by again running the algorithm on the vectorized version of the data. However, the algorithm was not able to produce any meaningful result. We believe this may be due mainly to the different way the package chooses its thresholds, resulting in computational issues by `clustMD` for data generated as described in Section 4.4.1.

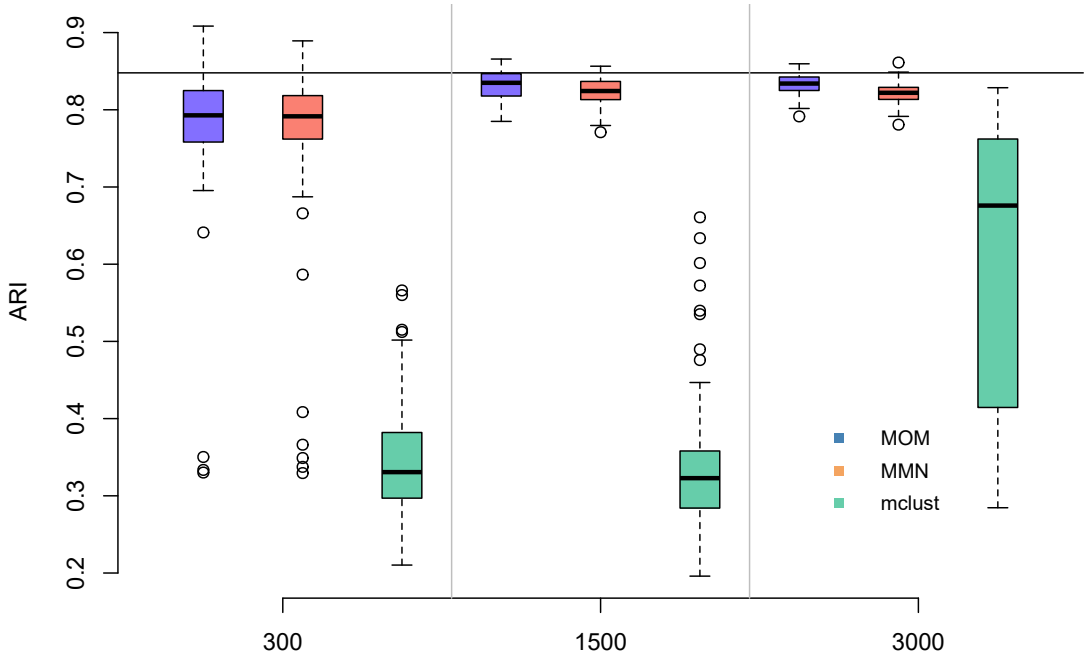
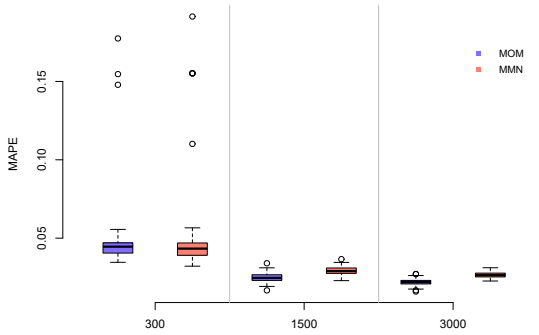



Figure 4.4: ARI for MOM, MMN and mclust. Kmeans++ initialization for MOM and MMN.

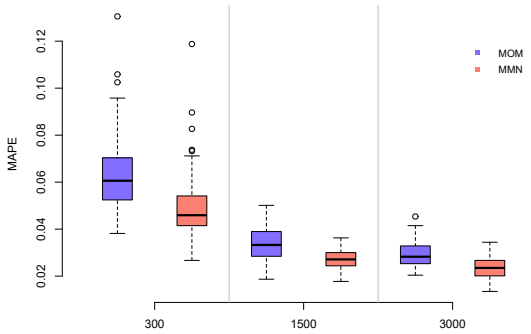
In Figure 4.4 the results for the partitioning task are shown. The difference in the ARI measurement is negligible for $N = 300$ for the two matri-variate model, but increases as $N$ increases. On the other hand, mclust is outperformed consistently.

The difference between MOM and MMN is clearer when comparing the MAPE values for the parameters estimation. As shown in Figure 5.5, the distance in error increases as $N$ increases for

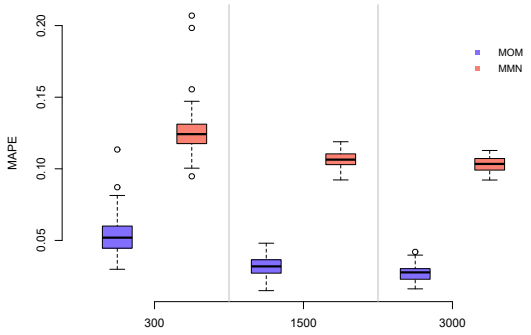
$M$ and $\Sigma$, but the same does not happens for the diagonal of $\Phi$, for which the MMN method seems to perform better, even if the difference dims as the sample size increases.




(a) MAPE for $M$




(b) MAPE for $\Phi$




(c) MAPE for $\Sigma$

Figure 4.5: MAPE results for parameter matrices. MOM vs MMN. Kmeans++ init. Note the difference in the scales.

## 4.5 Real Data

### 4.5.1 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 (François-Lecompte, Innocent,

Kréziak, and Prim-Allaz, 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 $1^{st}$ lockdown); April 30 - May 11, 2020 (end of the $1^{st}$ lockdown); June 9 - June 16, 2020 (post-lockdown); October 28 - November 9, 2020 (beginning of the $2^{nd}$ lockdown); March 5 - March 25, 2021 (just before the $3^{rd}$ 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:

- 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.

### 4.5.2 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 B1). 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 A1 for the mean $M$, Table A3 for the time covariances $\Phi$ and in Table A5 for the variable covariances $\Sigma$. To gain interpretability, covariances matrices have been transformed in correlation matrices in Table A2 for $\Phi$ and in Table A4 for $\Sigma$. In the tables the questions are named using their codes. Moreover, the correlation matrices $\Phi$ and $\Sigma$ are represented by correlation plots in Figures 4.8 and 4.9, respectively.

Figure 4.6 represents the 337 units (individuals) using a non-metric MDS (Venables and Ripley, 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 $11 \times 5$-dimensional matrices to 55-dimensional vectors. Each individual is represented by a circle whose color depends on its cluster.

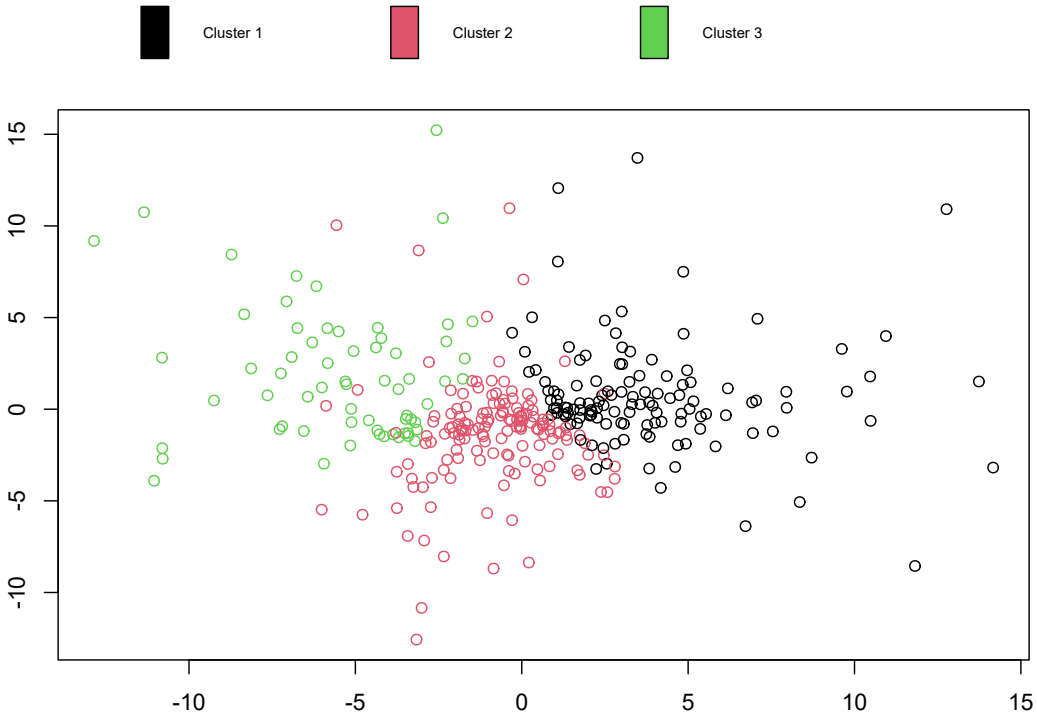



Figure 4.6: Units represented through isoMDS and colored by cluster allocation.

Figure 4.7 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.

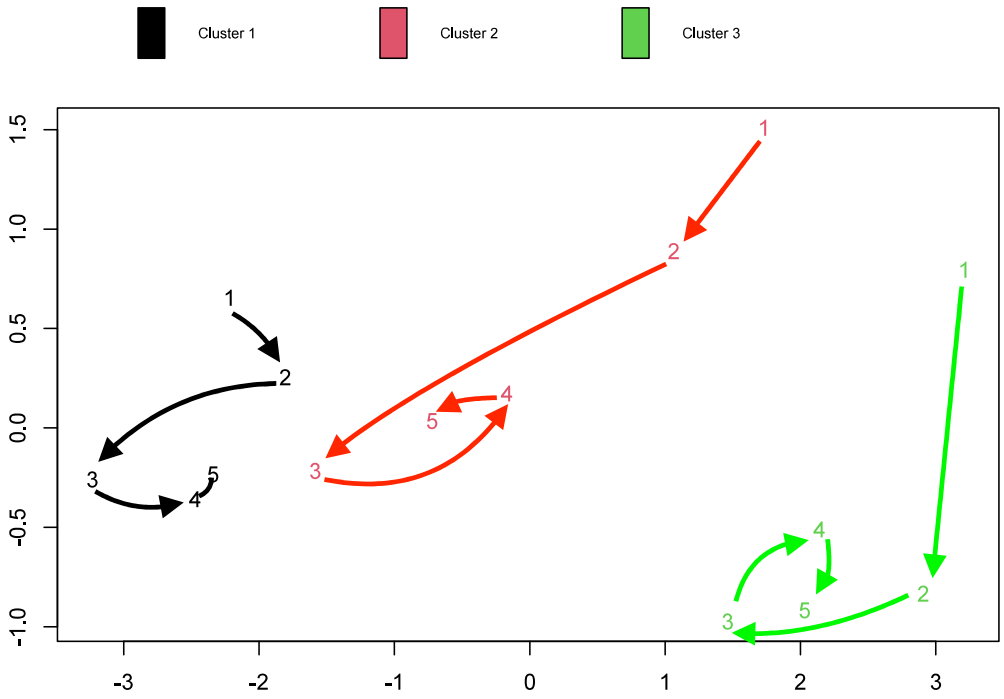



Figure 4.7: Evolution in time of cluster means. Representation through isoMDS. Numbers represent the time and the colors indicate the clusters.

### 4.5.3 Interpretation

Even if data are represented by means of a dimensionality reduction technique, discarding the temporal structure of data, we can see on Figure 4.6 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 A1. Moreover, from Figure 4.7 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 Cluster 2, which could be seen as an intermediary cluster between the other two.

- **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 4.5.1, 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 A1, 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

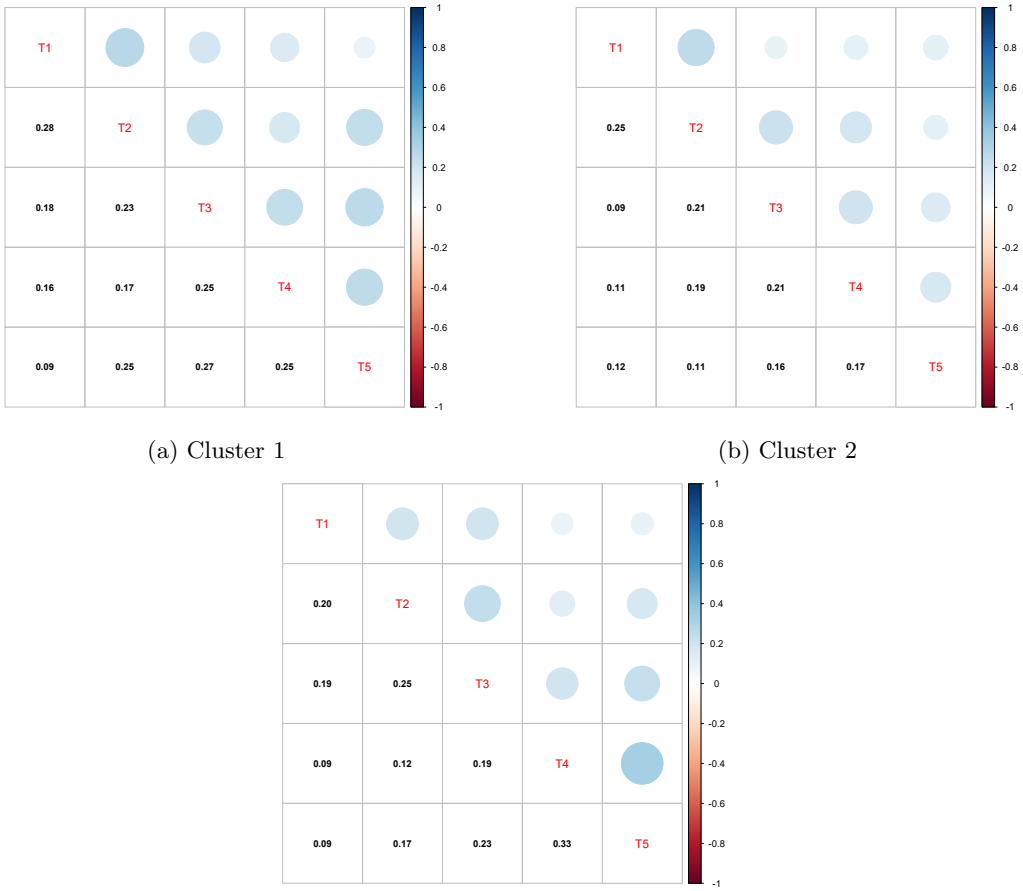
3 has remarkably high values here, meaning that this group of people really found that the pandemic period was stimulating a reflection on our lifestyle. As it turns out, this opinion fades as we advance towards T3 to then re-approach higher levels. It is interesting to observe that T3 corresponds to 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, there are some comments to be made about Q8(1) and intra-group correlations. As said in Section 4.5.1, Q8(1) is an inverse item, and it 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.

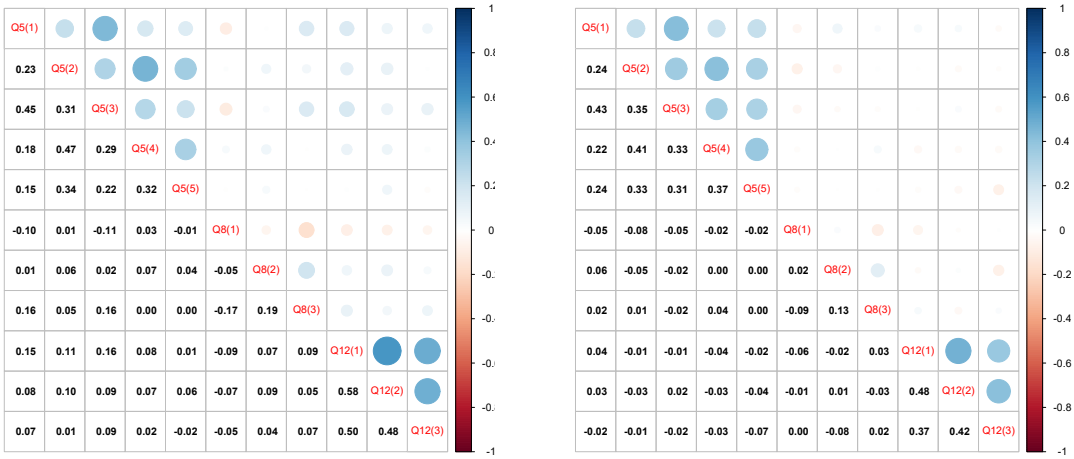



(a) Cluster 1

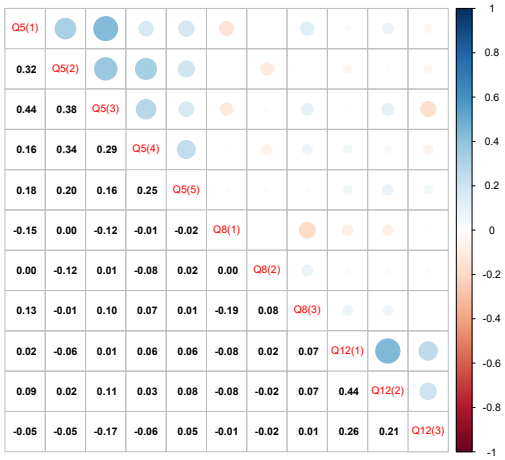
(b) Cluster 2

(c) Cluster 3

Figure 4.8: Clusters' corr-plots among time.




(a) Cluster 1 (b) Cluster 2




(c) Cluster 3

Figure 4.9: Clusters' corr-plots among variables.

## 4.6 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 4.4.4, 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 and Viroli, 2015 and in Sarkar, Zhu, Melnykov, and Ingrassia, 2020a. Besides, by applying a modified Cholesky decomposition on the time-related covariance matrix, one would obtain new matrices whose elements can be interpreted as generalized auto-regressive parameters and innovation variances, as shown by McNicholas and Murphy, 2010. 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 and Gormley, 2016.
Finally, one could as well think of implying, with proper adjustments, different underlying continuous distributions, such as heavy-tailed (Tomarchio, Punzo, and Bagnato, 2020), skewed (Gallaugher and McNicholas, 2018, Melnykov and Zhu, 2018) or t-student (Doğru, Bulut, and Arslan, 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 François-Lecompte, Innocent, Kréziak, and Prim-Allaz, 2020 for sharing their data. We would also like to thank Brendan Murphy for his invaluable inputs and support throughout the research process. His insights and expertise were instrumental in shaping the direction of this project.

# Appendices

# .1 Tables

Table A1: Clusters' means over time. The estimated parameter $\hat{\pi}$ = (0.37,0.44,0.19)

| | Cluster 1 | | | | | Cluster 2 | | | | | Cluster 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Questions | T1 | T2 | T3 | T4 | T5 | T1 | T2 | T3 | T4 | T5 | T1 | T2 | T3 | T4 | T5 |
| Q5(1) | 3.99 | 4.16 | 4.20 | 4.22 | 4.17 | 3.80 | 4.08 | 4.22 | 4.21 | 4.18 | 4.27 | 5.04 | 4.92 | 4.59 | 4.85 |
| Q5(2) | 3.60 | 3.77 | 4.02 | 4.02 | 4.15 | 3.72 | 3.79 | 4.10 | 4.07 | 4.13 | 3.83 | 4.36 | 4.48 | 4.35 | 4.47 |
| Q5(3) | 3.89 | 4.22 | 4.19 | 4.42 | 4.35 | 3.73 | 4.03 | 4.35 | 4.30 | 4.25 | 4.49 | 5.43 | 5.16 | 5.23 | 5.28 |
| Q5(4) | 3.51 | 3.78 | 3.95 | 3.98 | 4.03 | 3.49 | 3.78 | 3.99 | 3.97 | 3.98 | 3.53 | 4.08 | 4.26 | 4.34 | 4.44 |
| Q5(5) | 3.32 | 3.64 | 3.86 | 4.14 | 4.03 | 3.37 | 3.61 | 3.96 | 4.00 | 4.11 | 3.69 | 3.78 | 4.21 | 4.30 | 4.39 |
| Q8(1) | 3.36 | 3.42 | 3.61 | 3.66 | 3.64 | 3.30 | 3.49 | 3.70 | 3.70 | 3.57 | 2.15 | 2.26 | 2.55 | 3.03 | 2.74 |
| Q8(2) | 4.06 | 4.17 | 4.08 | 4.10 | 4.03 | 4.04 | 4.23 | 3.99 | 4.05 | 3.97 | 4.12 | 4.00 | 4.06 | 4.15 | 4.11 |
| Q8(3) | 4.12 | 4.16 | 4.15 | 4.19 | 4.09 | 4.05 | 4.27 | 4.07 | 4.14 | 4.08 | 4.35 | 4.76 | 4.53 | 4.49 | 4.64 |
| Q12(1) | 4.30 | 4.53 | 3.73 | 4.10 | 4.23 | 6.69 | 6.15 | 4.66 | 5.14 | 4.92 | 7.20 | 7.10 | 6.36 | 6.76 | 6.59 |
| Q12(2) | 4.13 | 4.50 | 3.53 | 3.94 | 4.15 | 6.69 | 6.38 | 4.49 | 5.56 | 5.18 | 7.22 | 6.93 | 6.08 | 6.61 | 6.65 |
| Q12(3) | 4.38 | 4.41 | 3.67 | 4.10 | 4.02 | 6.49 | 6.07 | 4.70 | 5.69 | 5.29 | 7.32 | 6.72 | 6.04 | 6.48 | 6.24 |

Table A2: Clusters' time correlation

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

Table A3: Clusters' time covariances

| | Cluster 1 | | | | | Cluster 2 | | | | | Cluster 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T / T | T1 | T2 | T3 | T4 | T5 | T1 | T2 | T3 | T4 | T5 | T1 | T2 | T3 | T4 | T5 |
| T1 | 1.34 | 0.36 | 0.20 | 0.17 | 0.09 | 1.17 | 0.30 | 0.09 | 0.10 | 0.12 | 1.50 | 0.32 | 0.27 | 0.13 | 0.14 |
| T2 | 0.36 | 1.25 | 0.25 | 0.18 | 0.26 | 0.30 | 1.22 | 0.21 | 0.19 | 0.11 | 0.32 | 1.78 | 0.38 | 0.19 | 0.28 |
| T3 | 0.20 | 0.25 | 0.88 | 0.21 | 0.23 | 0.09 | 0.21 | 0.84 | 0.17 | 0.13 | 0.27 | 0.38 | 1.33 | 0.26 | 0.33 |
| T4 | 0.17 | 0.18 | 0.21 | 0.87 | 0.22 | 0.10 | 0.19 | 0.17 | 0.81 | 0.14 | 0.13 | 0.19 | 0.26 | 1.42 | 0.49 |
| T5 | 0.09 | 0.26 | 0.23 | 0.22 | 0.85 | 0.12 | 0.11 | 0.13 | 0.14 | 0.86 | 0.14 | 0.28 | 0.33 | 0.49 | 1.48 |

Table A4: Clusters' variables correlation

| **Cluster 1** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| J / J | 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** | | | | | | | | | | | |
| J / J | 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** | | | | | | | | | | | |
| J / J | 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 A5: Clusters' variables covariances

| **Cluster 1** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| J / J | 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** | | | | | | | | | | | |
| J / J | 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** | | | | | | | | | | | |
| J / J | 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 |

## .2 Figures

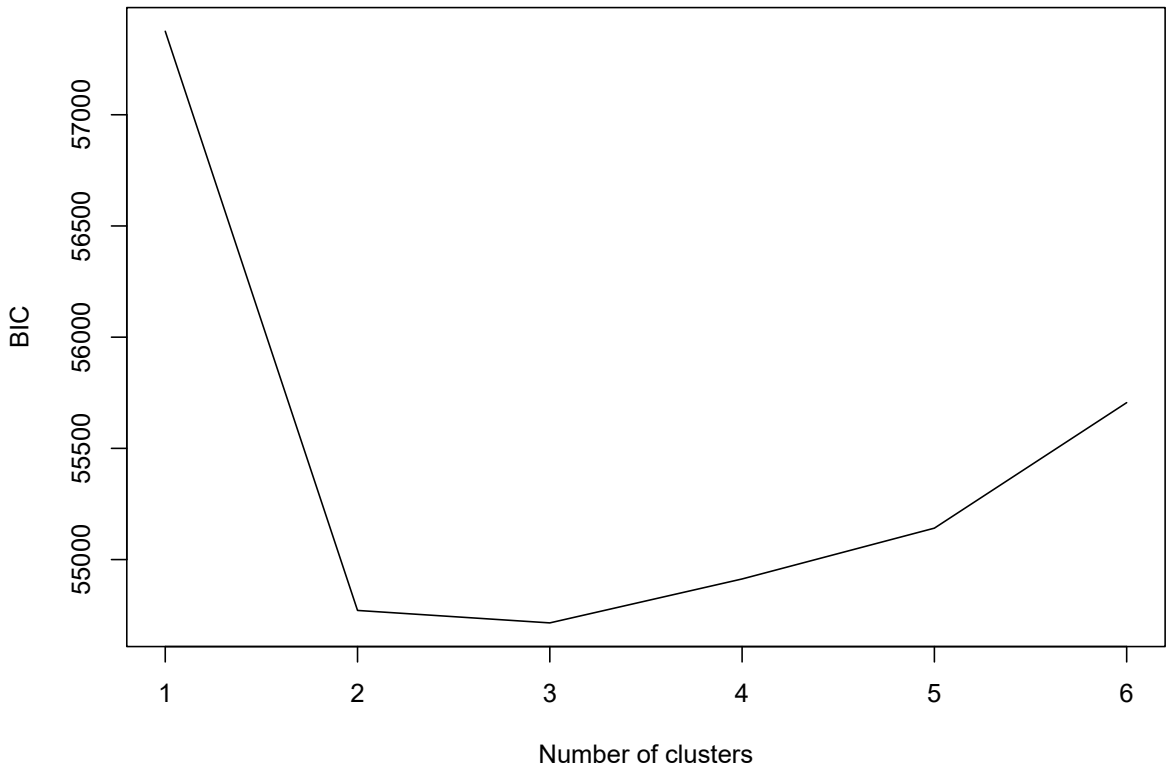



Figure B1: Visualization of BIC for K as results of application on real data. Kmeans++ initialization.

# Chapter 5

# MMM: Clustering Multivariate Longitudinal Mixed-type Data

This chapter has been submitted to the *Journal of Computational and Graphical Statistics* and is currently under review. A preprint has been released (Amato and Jacques, 2024). We have reproduced the entire preprint as released, expect for some small changes to the notation to keep it consistent with the rest of the thesis. For this reason, some concepts may be repeated, particularly concerning Section 5.3.1 with regard to Chapter 2 and Sections 5.1.2 and 5.1.1 to Chapter 3. The paper develops a model to cluster longitudinal mixed-type data by introducing the Mixture of Mixed Matrices (MMM) model.

**Abstract.** Multivariate longitudinal data of mixed-type are increasingly collected in many science domains. However, algorithms to cluster this kind of data remain scarce, due to the challenge to simultaneously model the within- and between-time dependence structures for multivariate data of mixed kind. We introduce the Mixture of Mixed-Matrices (MMM) model: reorganizing the data in a three-way structure and assuming that the non-continuous variables are observations of underlying latent continuous variables, the model relies on a mixture of matrix-variate normal distributions to perform clustering in the latent dimension. The MMM model is thus able to handle continuous, ordinal, binary, nominal and count data and to concurrently model the heterogeneity, the association among the responses and the temporal dependence structure in a parsimonious way and without assuming conditional independence. The inference is carried out through an MCMC-EM algorithm, which is detailed. An evaluation of the model through synthetic data shows its inference abilities. A real-world application on financial data is presented.

**Keywords.** Model-based clustering. Mixed-type multivariate longitudinal data. Three-way data. Mixture models. Matrix-variate Gaussians.

## 5.1 Context

Multivariate longitudinal data of mixed-type are increasingly collected in many science domains. For example, in social sciences studies are often based on questionnaires encompassing different type of answers completed by participants multiple times. In physical sciences, phenomena are often measured repeatedly with different types of measurements.
However, the statistical analysis of these data is far from simple, for several reasons. First, the collected data are often of different typology, ranging from continuous to count data. The analysis of such mixed-type data is a current research problem in the fields of statistics and machine learning (Ahmad and Khan, 2019). The second scientific obstacle is the modeling of the temporal trajectory. Currently, frequently the analyses are done independently at each temporal phase, then researchers try *a posteriori* to find links among times, by seeking from one phase to the other to find similar typical behavior. An example is Selosse, Jacques, Biernacki, and Cousson-Gélie, 2019 in the case of clustering of longitudinal ordinal data for an application in psychology.

In this work we aim at providing a tool to perform clustering on multivariate longitudinal mixed-type data. 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, Celeux, Murphy, and Raftery, 2019b).

### 5.1.1 Related work

While several approaches exist for the clustering of longitudinal and mixed-type data separately, literature is rather poor when they are to be dealt with simultaneously. In the following, we will present a brief overlook to the main methods to cluster mixed data, longitudinal data and mixed longitudinal data.

Although many data sets contain mixed-type data, few mixture models can manage these data (Hunt and Jorgensen, 2011) due to the shortage of multivariate distributions able to handle them. Clustering with mixed-type data have received a large attention in the last decade from the researcher in statistics and machine learning. The Latent Class Model (LCM) (Everitt, 1984) is frequently used, and it assumes that the variables are conditionally independent upon the cluster membership. Consequently, the joint probability distribution function (pdf) of the features of different types is obtained by the product of the pdfs of each individual feature. However, when the variables are inherently correlated in a cluster, this model is not suitable. To overcome this issue, the authors of Marbac, Biernacki, and Vandewalle, 2017 wanted to conserve standard marginal distributions but also tried to loosen the conditional independence on the variables. For this purpose, they used a copula, which allow definition of both the dependence model and the type of marginal distributions. The proposed model relies on the main assumption that each cluster follows a Gaussian copula. However, the authors note that model complexity increases promptly with the number of variables, which is not suitable in a big-data context. Moreover, it is not easily interpretable by non-statistician practitioners. More recently, Hermes, Heerwaarden, and Behrouzi, 2024 proposed a similar approach by using copulas in the context of graphical models, which were already extended for use for mixed-type data by Cheng, Li, Levina, and Zhu, 2017. In Selosse, Jacques, and Biernacki, 2020, another model-based approach for ordinal, nominal, integer and continuous

data is proposed, on the basis of conditional independence assumption and with the particularity of creating clusters of features as well as clusters of individuals (co-clustering).
Another way to address the issues of mixed-type data is to see some variables as the manifestation of latent variables. For example, in McParland and Gormley, 2016, the clustMD model considers continuous and categorical data (nominal and ordinal) and assumes that a categorical variable is the representation of an underlying latent continuous variable. Then, it is assumed that the continuous variables (observed and unobserved) follow a multivariate Gaussian Mixture Model (GMM). This model is further developed to address sparsity by Choi, Ahn, and Kim, 2023.

Modelling longitudinal data poses a different kind of challenge than mixed-type data, as the grouping has to account for the similarity of individual trajectories which disrupt the independence assumption among observations. Additionally, this kind of data introduces the issue of dealing with time, often with sparse observations that makes unsuitable the use of models coming from the domains such as functional data, time series and Gaussian processes. In order to bypass these problems, some authors preferred to focus on geometric non-parametric clustering algorithms, as done by Bruckers, Molenberghs, Drinkenburg, and Geys, 2016 with an idea based on K-means clustering and by Zhou, Zhang, and Tu, 2023 with hierarchical clustering, among others. For parametric methods, a well established manner to model longitudinal data is through mixed-effects models. This research domain is well-established and vast. We refer to Gad and Kholy, 2012 for an overview and to the related work section of Hui, Dang, and Maestrini, 2024 for the most recent advancements. The main issues with this kind of models are the over-parametrization and the computational burden that often arises with it.

Another approach to clustering longitudinal data that gained traction in the last decade 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-behaviour, 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 MMN is provided in §2.1 of Wang and Melnykov, 2020). In addition, in the context of linear mixed models with discrete individual random intercepts to analyze longitudinal continuous data, Anderlucci and Viroli, 2015 proposed Covariance Pattern Mixture Model (CPMM) which, by leveraging three-way data structures, does not require the usual local independence assumption. This model can be seen as an extension of the proposal of McNicholas and Murphy, 2010 in the multivariate context. 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, Zhu, Melnykov, and Ingrassia, 2020a) and a new R package (Zhu, Sarkar, and Melnykov, 2022) has been proposed. In addition, Cappozzo, Casa, and Fop, 2024 proposed a lasso-type penalization to account for sparsity. Despite their efficacy, up to now these methods have generally only been applied to continuous data.
More recently, Amato, Jacques, and Prim-Allaz, 2024 proposed a method to cluster longitudinal

ordinal data by assuming an underlying mixture of matrix-variate distributions.

Finally, looking at mixed-type longitudinal data, one main methodology to deal with such data lies in the framework of discrete (time-constant or varying) random intercepts for modeling heterogeneity, that includes mixture of random effect models for longitudinal data extended to deal with multivariate and mixed outcomes by Proust-Lima, Amieva, and Jacqmin-Gadda, 2013 and growth mixture models (Ram and Grimm, 2009a), where individuals are grouped in classes having a specific growth structure variability. These approaches are similar in that they model the change over time at both the population level and the individual level using random effects (or latent variables). In Komarek and Komárková, 2013 the authors rely on a multivariate extension of the classical generalized linear mixed model where a mixture distribution is additionally assumed for random effects. Vávra and Komárek, 2021 extend this model presenting a statistical model for joint modeling of mixed-type longitudinal data, while performing unsupervised clustering with respect to different covariate patterns. However, nominal (polytomous) variables are not taken into account in neither of the papers and time-dependent information is neglected. This work is expanded and improved in Vávra, Komárek, Grün, and Malsiner-Walli, 2024. In Cagnone and Viroli, 2018 the authors extended the latent class model to take into account time evolution by means of latent Markov variable (Bartolucci, Farcomeni, and Pennoni, 2019) to model longitudinal binary and ordinal data on alcohol use disorder.
In a model-based clustering perspective, De la Cruz-Mesia, Quintana, and Marshall, 2008 proposed a mixture of hierarchical nonlinear models for describing nonlinear relationships across time. Manrique-Vallier, 2014 introduced a clustering strategy based on a mixed membership framework for analyzing discrete multivariate longitudinal data.

### 5.1.2 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 variances and covariances of the $J$ variables. The matrix-normal probability density function is given by

$$f(Z|M,\Phi,\Sigma) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^\intercal]\right\}. \tag{5.1.1}$$

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 $\text{vec}(Z) \sim \mathcal{N}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$, where vec(.) is the vectorization operator, that is the function mapping from 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 matrix-normal distribution are:

$$\mathbb{E}(\text{vec}(Z)|M,\Phi,\Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M,\Phi,\Sigma) = \Psi. \tag{5.1.2}$$

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 $\Phi$. Second, it represents a more parsimonious solution than that of the unrestricted $\Psi$.

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

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

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

### 5.1.3 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 and Viroli, 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 Mixed Matrices (MMM) model, aiming at expanding the use to matrix-variate mixtures to ordinal data in an unsupervised learning context.
Our model expands the use of matrix-variate mixtures to mixed-type data, by building on the framework proposed by McParland and Gormley, 2016 and Choi, Ahn, and Kim, 2023 in the cross-sectional context.

In the following, in Sections 5.2 and 5.3 we will detail our model and the MCMC-EM algorithm to perform inference, respectively. In Section 5.4 some results on synthetic data are presented to assess the performance of the model. Finally, in Section 5.5 an real-world application concerning stock exchange data during the Covid-19 pandemic period is outlined.

## 5.2 The MMM model

Let denote by $y_{ijt}$ the observation of the $j$-th ($j = 1, .., J$) variable for the $i$-th ($i = 1, \dots, N$) unit at time $t$ ($t = 1, \dots, T$), 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, $O$ ordinal, binary and nominal ones and $G$ as count variables, such that $C + O + G = J$. We are going to put ordinal, binary and nominal variables together as we will treat them in the same way.
Let us reorganize this data in a random-matrix form such that we denote the observed record of the $i$-th subject as $Y_i \in \mathbb{R}^{J\times T}$. $\mathbf{Y} = \{Y_i\}_{i=1}^N$ is a sample of $J \times T$-variate matrix observations $Y_i \in [\mathbb{R}^{C\times T}, \mathbb{N}^{O\times T}, \mathbb{N}_0^{G\times T}]^\intercal$, $J = C+O+G$. The ordinal, binary and nominal classes are arbitrarily coded by non-negative integers such that each variable $O$ has levels $\{1, 2, \dots, C_o\} \in \mathbb{N}$ [1].

[1] In this work, we will consider zero not included in the set of natural numbers. We will use the notation $\mathbb{N}_0$ to indicate $\mathbb{N} \cup \{0\}$

Then, we assume that each variable $y_{ijt}$ is the manifestation of an underlying latent continuous variable $z_{ijt}$.

### 5.2.1 Modeling continuous variables

Let $c$ indicate the generic $c$-th continuous variable. We assume that the observed continuous variables $y_{ict}$ matches exactly the latent variable:

$$y_{ict} = z_{ict}$$

### 5.2.2 Modeling categorical ordinal variables

To map ordinal data, we follow Amato, Jacques, and Prim-Allaz, 2024. Let the generic ordinal $o$-th have $C_o$ levels. Let $\gamma_o$ 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 $z_{iot}$ is such that $\gamma_{o,c-1} < z_{iot} < \gamma_{o,c}$ then the observed ordinal response, $y_{iot} = c$.

Moreover, let define $\mathcal{O}^{O\times T}$ the set of ordinal matrices of size $J \times T$ whose elements 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 $\Omega_r$ 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 $\gamma_o = (-\infty, 1.5, 2.5, \ldots, C_o - 0.5, \infty)$.

### 5.2.3 Modeling categorical nominal variables

For categorical nominal 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.

### 5.2.4 Modelling count variables

For count data we consider a Matrix variate Poisson-log normal distribution (Silva et al., 2023). Let $g$ be the generic $g$-th count variable, then we assume that $y_{igt}$ follows a Poisson distribution with parameter $\exp(z_{igt})$, where $z_{igt}$ is a term of the $G \times T$ underlying latent matrix following a matrix normal distribution.

### 5.2.5 Joint model

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, followed by $O$ rows representing the categorical variables and the remaining $J - C - O = G$ rows, representing the count variables. Notice that the slicing happens

just over rows but not over columns. Then, we can write $Y_i = [Y_i^\alpha, Y_i^\beta, Y_i^\gamma]^\intercal$, 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 categorical ones (that we coded via integers) and the binary ones, and $Y_i^\gamma \in \mathbb{N}_0^{G\times T}$ is the block containing the count variables.

At this point, we can assume that each observed block of the matrix $Y_i$ is indeed the manifestation of the corresponding block of the latent random matrix $Z_i = [Z_i^\alpha, Z_i^\beta, Z_i^\gamma]^\intercal$, 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 $y_{ijt}$, as described previously.
Then, we assume a mixture of matrix-normal distributions on the latent space. We can consequently write

$$f\begin{pmatrix} Z_i^\alpha \\ Z_i^\beta \\ Z_i^\gamma \end{pmatrix} = \sum_{k=1}^{K} \pi_k \, \mathcal{MN}_{(J\times T)}\left(\begin{pmatrix} M_k^\alpha \\ M_k^\beta \\ M_k^\gamma \end{pmatrix}, \Phi_k, \begin{pmatrix} \Sigma_k^{\alpha\alpha} & \Sigma_k^{\alpha\beta} & \Sigma_k^{\alpha\gamma} \\ \Sigma_k^{\beta\alpha} & \Sigma_k^{\beta\beta} & \Sigma_k^{\beta\gamma} \\ \Sigma_k^{\gamma\alpha} & \Sigma_k^{\gamma\beta} & \Sigma_k^{\gamma\gamma} \end{pmatrix}\right). \tag{5.2.1}$$

From here, we can derive the joint model. To keep notation coherent, let define with $\tilde{Y}_i^\beta$ the one-hot encoding of the categorical part of $Y_i$ as described in Section 5.2.2. In addition to $Z_i$, we introduce a latent binary K-dimensional allocation 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. Recalling the links each kind of observed variables has with the latent ones, we can express our model through the following distributional assumptions:

$$\begin{aligned}
&\ell_i \sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} := (\pi_1, \dots, \pi_K) \\
&Z_i^\alpha | \ell_{ik} = 1 \sim \mathcal{MN}_{(C\times T)}(Z_i^\alpha | \Theta_k^\alpha),\ \Theta_k^\alpha := \{M_k^\alpha, \Phi_k, \Sigma_k^\alpha\}, \\
&Z_i^\beta | Z_i^\alpha, \ell_{ik} = 1 \sim \mathcal{MN}_{(O\times T)}(Z_i^\beta | \Theta_k^{\beta|\alpha}),\ \Theta_k^\beta := \{M_k^{\beta|\alpha}, \Phi_k, \Sigma_k^{\beta|\alpha}\}, \\
&Z_i^\gamma |, Z_i^\alpha, Z_i^\beta, \ell_{ik} = 1 \sim \mathcal{MN}_{(G\times T)}(Z_i^\gamma | \Theta_k^{\gamma|\alpha,\beta}),\ \Theta_k^\gamma := \{M_k^{\gamma|\alpha,\beta}, \Phi_k, \Sigma_k^{\gamma|\alpha,\beta}\}; \\
&\tilde{Y}_i^\beta | Z_i^\beta, \ell_{ik} = 1 \sim \mathcal{M}(1, \xi_i),\ \xi_i := (\mathbf{1}_{\Omega_1}(Z_i^\beta), ..., \mathbf{1}_{\Omega_R}(Z_i^\beta)), \\
&Y_{igt}^\gamma | Z_{igt}^\gamma \sim \mathcal{P}(\exp(Z_{igt}^\gamma)),
\end{aligned}$$

where $\mathcal{M}$ indicates the multinomial distribution and $\mathbf{1}_{\Omega_r}(Z_i^\beta)$ is the indicator function that equals 1 when the elements in $Z_i^\beta$ have values that determine the $r$-th pattern. Hence, when $\tilde{Y}_{ir}^\beta = 1$, the vector $\xi_i$ is a vector whose $r$-th element equals 1 and all the others equal 0.
Further, to avoid assuming the independence between the different blocks, to link the matrix latent distributions we resort to condition on one block to another by using the properties of matrix-variate normal distribution (Gupta and Nagar, 2000). Thus, $\Theta_k^{\gamma|\alpha,\beta} := \{M_k^{\gamma|\alpha,\beta}, \Phi_k, \Sigma_k^{\gamma|\alpha,\beta}\}$, more precisely $M_k^{\gamma|\alpha,\beta} = M_k^\gamma + \Sigma_k^{\gamma\cdot}\Sigma_k^{-1,\cdot\cdot}(Z_i^{\alpha,\beta} - M_k^{\alpha,\beta})$ and $\Sigma_k^{\gamma|\alpha,\beta} = \Sigma_k^{\gamma\gamma} - \Sigma_k^{\gamma\cdot}\Sigma_k^{-1,\cdot\cdot}\Sigma_k^{\cdot\gamma}$, and where $\Theta_k^{\beta|\alpha} := \{M_k^{\beta|\alpha}, \Phi_k, \Sigma_k^{\beta|\alpha}\}$, more precisely $M_k^{\beta|\alpha} = M_k^\beta + \Sigma_k^{\beta\alpha}\Sigma_k^{-1,\alpha\alpha}(Y_i^\alpha - M_k^\alpha)$ and $\Sigma_k^{\beta|\alpha} = \Sigma_k^{\beta\beta} - \Sigma_k^{\beta\alpha}\Sigma_k^{-1,\alpha\alpha}\Sigma_k^{\alpha\beta}$.

Lastly, Assuming that the observed value pattern of $\tilde{Y}_i^\beta$ is $r$ for sake of notation, we can compose the distribution of each observed mixed matrix as

$$f(Y_i) = \sum_{k=1}^{K} \pi_k \, \mathcal{MN}_{(C\times T)}(Z_i^\alpha | \Theta_k^\alpha) \cdot \int_{\Omega_r} \mathcal{MN}_{(O\times T)}(Z_i^\beta | \Theta_k^{\beta|\alpha}) dZ_i^\beta$$
$$\cdot \int_{\mathbb{R}} \prod_t^T \prod_g^G \mathcal{P}(y_{igt}^\gamma | \exp(z_{igt}^\gamma)) \cdot \mathcal{MN}_{(G\times T)}(Z_i^\gamma | \Theta_k^{\gamma|\alpha,\beta}) dZ_i^\gamma \,. \tag{5.2.2}$$

### 5.2.6 Likelihood

In the following, $\boldsymbol{Z} := \{Z_i\}_{i=1}^N$ , $\boldsymbol{\ell} := \{\ell_i\}_{i=1}^N$ will indicate the ensembles of $Z_i$ and $\ell_i$ respectively, and $\mathbf{Y} := \{Y_i\}_{i=1}^N$ be the collection of the observed matrices $Y_i$. Finally, the set of unknown parameters to be estimated is $\boldsymbol{\Theta} := \{\pi_k, M_k, \Phi_k, \Sigma_k\}_{k=1}^K$ .

The joint density of $Y_i^\gamma, Z_i^\gamma, \tilde{Y}_i^\beta, Z_i^\beta, Z_i^\alpha, \ell_i$ is:

$$f(Y_i^\gamma, Z_i^\gamma, \tilde{Y}_i^\beta, Z_i^\beta, Z_i^\alpha, \ell_i) = f(Y_i^\gamma | Z_i^\gamma, \tilde{Y}_i^\beta, Z_i^\beta, Z_i^\alpha, \ell_i) \cdot f(Z_i^\gamma | \tilde{Y}_i^\beta, Z_i^\beta, Z_i^\alpha, \ell_i) \cdot$$
$$f(\tilde{Y}_i^\beta | Z_i^\beta, Z_i^\alpha, \ell_i) \cdot f(Z_i^\beta | Z_i^\alpha, \ell_i) \cdot f(Z_i^\alpha | \ell_i) \cdot f(\ell_i).$$

We can therefore write the complete log-likelihood as:

$$\mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \prod_{i=1}^N \prod_{k=1}^K \Big[ \pi_k \cdot \Big( \prod_t^T \prod_g^G \mathcal{P}(y_{igt}^\gamma | \exp(z_{igt}^\gamma)) \Big) \cdot \mathcal{MN}_{(G\times T)}(Z_i^\gamma | \Theta_k^{\gamma|\alpha,\beta}) \cdot$$
$$\mathcal{MN}_{(O\times T)}(Z_i^\beta | \Theta_k^{\beta|\alpha}) \cdot \mathcal{MN}_{(C\times T)}(Z_i^\alpha | \Theta_k^\alpha) \cdot \prod_{r=1}^R \mathbf{1}_{\Omega_r}(Z_i^\beta)^{\tilde{Y}_{ir}^\beta} \Big]^{\ell_{ik}}. \tag{5.2.3}$$

We acknowledge the identity $Y_i^\alpha = Z_i^\alpha$ and the fact that the last term is non-stochastic since conditioning on $\boldsymbol{Z}$ implies that the value of $Z_i^\beta$ is known and we can discard it. Therefore, by using the notation of Equation 5.2.1, we can rewrite this equation compactly as the complete log-likelihood can be written as:

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^N \sum_{k=1}^K \ell_{ik} \Big[ C + \log(\pi_k) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) -$$
$$\frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \Big]. \tag{5.2.4}$$

where $C$ is a constant with respect to the set of parameters $\boldsymbol{\Theta}$.

On the other hand, we can define the observed likelihood as $\mathcal{L}_O(\boldsymbol{\Theta}; \boldsymbol{Y})$, that is:

$$\mathcal{L}_O(\boldsymbol{\Theta}; \boldsymbol{Y}) := \prod_{i=1}^{N} \Big\{ \sum_{k=1}^{K} \pi_k \, \mathcal{MN}_{(C\times T)}(Z_i^{\alpha} | \Theta_k^{\alpha}) \cdot \int_{\Omega_r} \mathcal{MN}_{(O\times T)}(Z_i^{\beta} | \Theta_k^{\beta|\alpha}) dZ_i^{\beta} \cdot \int_{\mathbb{R}} \prod_{t}^{T} \prod_{g}^{G} \mathcal{P}(y_{igt}^{\gamma} | \exp(z_{igt}^{\gamma})) \times \mathcal{MN}_{(G\times T)}(Z_i^{\gamma} | \Theta_k^{\gamma|\alpha,\beta}) dZ_i^{\gamma} \Big\}. \tag{5.2.5}$$

## 5.3 Inference

In our model, we are assuming two different latent (unobserved) variables. Therefore, we will use the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) to infer the MMM model's parameters. The EM algorithm is well-suited for situations involving latent variables or unobserved data, as it allows for the estimation of model parameters despite the incompleteness.

### 5.3.1 EM-algorithm

The EM algorithm is an iterative algorithm that 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.
For the MMM model, 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 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.

### 5.3.2 Initialization

To find the initial values of $\hat{\boldsymbol{\Theta}}^{(0)}$ mentioned in Section 5.3.1, our proposal is the following. Identity matrices are chosen for the initialization of the covariance matrices $\Phi_k$ and $\Sigma_k$.
For the initialization of $M_k$ and $\pi_k$, two solutions are proposed and tested in Section 5.4.3. The first is a Kmeans++ (Arthur and Vassilvitskii, 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 5.4.3, 5 random initializations proved to be enough, but a higher number might be needed for more complex settings.
Both the initialization techniques are applied on the latent space, meaning that for count data they are applied on the logarithm of the observed data.

### 5.3.3 E-step

As previously stated, 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.

We can expand Equation 5.2.4 as:

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \ell_{ik}\Bigg[C + \log(\pi_k) - \frac{J}{2}\log(|\Phi_k|) - \frac{T}{2}\log(|\Sigma_k|) - \frac{1}{2}tr[\Sigma_k^{-1} Z_i \Phi_k^{-1} Z_i^\intercal - \Sigma_k^{-1} Z_i \Phi_k^{-1} M_k^\intercal - \Sigma_k^{-1} M_k \Phi_k^{-1} Z_i^\intercal + \Sigma_k^{-1} M_k \Phi_k^{-1} M_k^\intercal]\Bigg]. \quad (5.3.1)$$

Then, from Equation 5.3.1, it is easy to see that the expected values to be computed are $\mathbb{E}(\ell_{ik}|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, $\mathbb{E}(\ell_{ik} Z_i|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ and $\mathbb{E}(\ell_{ik} Z_i \Phi_k^{-1(s)} Z_i^\intercal|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ or $\mathbb{E}(\ell_{ik} Z_i^\intercal \Sigma_k^{-1(s)} Z_i|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ by the cyclic property of the trace. As we will see in Section 5.3.4, we will need both.

We will proceed with their computation one by one. First, $\mathbb{E}(\ell_{ik}|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$ can be computed according to the Bayes' rule as

$$\mathbb{E}(\ell_{ik}|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}) = \frac{q_{ik}}{\sum_{h=1}^{K} q_{ih}} =: \hat{\tau}_{ik}^{(s+1)} \quad (5.3.2)$$

where

$$q_{ik} = \pi_k \, \mathcal{MN}_{(C\times T)}(Z_i^\alpha|\Theta_k^{(s),\alpha}) \cdot \int_{\Omega_r} \mathcal{MN}_{(O\times T)}(Z_i^\beta|\Theta_k^{(s),\beta|\alpha}) dZ_i^\beta \cdot \int_{\mathbb{R}} \prod_t^T \prod_g^G \mathcal{P}(y_{igt}^\gamma|\exp(z_{igt}^\gamma)) \cdot \mathcal{MN}_{(G\times T)}(Z_i^\gamma|\Theta_k^{\gamma|\alpha,\beta}) dZ_i^\gamma$$

where the first integral can be approximated through a Monte-Carlo approach applied on the vectorized reparametrization of the matrix-variate distribution and the second one can be approximated by using the estimated value for $Z_i^\gamma$ presented in the following.

For $\mathbb{E}(\ell_{ik} Z_i|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y})$, recalling the block structure of $Z_i$, we can write

$$\mathbb{E}(\ell_{ik} Z_i|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}) = \mathbb{P}(\ell_{ik} = 1|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}) \cdot \mathbb{E}\left(\begin{bmatrix} Z_i^\alpha \\ Z_i^\beta \\ Z_i^\gamma \end{bmatrix} \middle| \ell_{ik} = 1, \hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}\right) =$$
$$= \mathbb{P}(\ell_{ik} = 1|\hat{\boldsymbol{\Theta}}^{(s)}, \mathbf{Y}) \cdot \begin{bmatrix} Y_i^\alpha \\ \mathbb{E}(Z_i^\beta|M_k^{\beta|\alpha,(s)}, \Phi_k^{(s)}, \Sigma_k^{\beta|\alpha,(s)}) \\ \mathbb{E}(Z_i^\gamma|M_k^{\gamma|\alpha,\beta,(s)}, \Phi_k^{(s)}, \Sigma_k^{\gamma|\alpha,\beta,(s)}) \end{bmatrix} := \hat{\tau}_{ik}^{(s+1)} \cdot \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix}, \quad (5.3.3)$$

where the matrix-variate expectation related to count data can be computed by defining $z_i^\gamma \in \mathbb{R}^{GT\times 1}$ as the vectorized version of $Z_i^\gamma$ and computing its expectation $\hat{m}_{ik}^{\gamma,(s+1)} := \mathbb{E}(z_i^\gamma|\ell_{ik} = 1, \mathbf{Y}, \hat{\boldsymbol{\Theta}}^{(s)})$ by means of the sampler implemented in the R package `Rstan`, that is the R interface

to the `Stan` software (Stan Development Team, 2024).
The matrix-variate expectation related to categorical data can be computed by defining $z_i^\beta \in \mathbb{R}^{OT\times 1}$ as the vectorized version of $Z_i^\beta$ and computing its expectation $\hat{m}_{ik}^{\beta,(s+1)} := \mathbb{E}(z_i^\beta|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)})$ through the use of a Gibbs sampler to sample from a truncated multivariate normal distribution.

Finally, for $\mathbb{E}(\ell_{ik}Z_i\Phi_k^{-1}Z_i^\intercal|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$ we have:

$$\begin{aligned}\mathbb{E}(\ell_{ik}Z_i\Phi_k^{-1}Z_i^\intercal|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) &= \mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})\cdot\mathbb{E}(Z_i\Phi_k^{-1}Z_i^\intercal|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) = \\ &= \hat{\tau}_{ik}^{(s+1)}\cdot\begin{bmatrix} Y_i^\alpha\hat{\Phi}_k^{-1(s)}Y_i^{\alpha\intercal} & Y_i^\alpha\hat{\Phi}_k^{-1(s)}\hat{M}_{ik}^{\beta,\intercal(s+1)} & Y_i^\alpha\hat{\Phi}_k^{-1(s)}\hat{M}_{ik}^{\gamma,\intercal(s+1)} \\ \hat{M}_{ik}^{\beta,(s+1)}\hat{\Phi}_k^{-1(s)}Y_i^{\alpha\intercal} & \hat{D}_{ik}^{(s+1)} & \hat{M}_{ik}^{\beta,(s+1)}\hat{\Phi}_k^{-1(s)}\hat{M}_{ik}^{\gamma,\intercal(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)}\hat{\Phi}_k^{-1(s)}Y_i^{\alpha\intercal} & \hat{M}_{ik}^{\gamma,(s+1)}\hat{\Phi}_k^{-1(s)}\hat{M}_{ik}^{\beta,\intercal(s+1)} & \hat{B}_{ik}^{(s+1)}\end{bmatrix},\end{aligned} \tag{5.3.4}$$

where $\hat{D}_{ik}^{(s+1)} := \mathbb{E}(Z_i^\beta\Phi_k^{-1}Z_i^{\beta\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$ and $B_{ik}^{(s+1)} := \mathbb{E}(Z_i^\gamma\Phi_k^{-1}Z_i^{\gamma\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}))$.

To compute $D_{ik}^{(s+1)}$ we make use of the the elements of $\hat{S}_{ik}^{\beta,(s+1)} := \mathbb{E}(z_i^\beta z_i^{\beta\intercal}|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)})$. The samples generated to calculate the first moment $\hat{m}_{ik}^{\beta,(s+1)}$ can be reused to compute the matrix $\hat{S}_{ik}^{(s+1)}$ by calculating the mean of the inner product between them.
Similarly, for $\hat{B}_{ik}^{(s+1)}$, we make use of the the elements of $\hat{S}_{ik}^{\gamma,(s+1)} := \mathbb{E}(z_i^\gamma z_i^{\gamma\intercal}|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)})$. As before, the samples generated to calculate the first moment $\hat{m}_{ik}^{\gamma,(s+1)}$ can be reused to compute the matrix $\hat{S}_{ik}^{\gamma,(s+1)}$.

On the other hand, to compute $\mathbb{E}(\ell_{ik}Z_i^\intercal\Sigma_k^{-1}Z_i|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$:

$$\begin{aligned}\mathbb{E}(\ell_{ik}Z_i^\intercal\Sigma_k^{-1}Z_i|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) &= \mathbb{P}(\ell_{ik}=1|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})\cdot\mathbb{E}(Z_i^\intercal\Sigma_k^{-1}Z_i|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) = \\ &= \hat{\tau}_{ik}^{(s+1)}\cdot\Big(Y_i^{\alpha\intercal}\hat{\Sigma}_k^{-1,\alpha\alpha}Y_i^\alpha + Y^{\alpha\intercal}\hat{\Sigma}_k^{-1,\alpha\beta}\hat{M}_{ik}^{\beta,(s+1)} + Y^{\alpha\intercal}\hat{\Sigma}_k^{-1,\alpha\gamma}\hat{M}_{ik}^{\gamma,(s+1)} + \\ &\quad \hat{M}_{ik}^{\beta,(s+1)\intercal}\hat{\Sigma}_k^{-1,\beta\alpha}Y_i^\alpha + \hat{C}_{ik}^{(s+1)} + \hat{M}_{ik}^{\beta,(s+1)\intercal}\hat{\Sigma}_k^{-1,\beta\gamma}\hat{M}_{ik}^{\gamma,(s+1)} + \\ &\quad \hat{M}_{ik}^{\gamma,(s+1)\intercal}\hat{\Sigma}_k^{-1,\gamma\alpha}Y_i^\alpha + \hat{M}_{ik}^{\gamma,(s+1)\intercal}\hat{\Sigma}_k^{-1,\gamma\beta}\hat{M}_{ik}^{\beta,(s+1)} + \hat{A}_{ik}^{(s+1)}\Big),\end{aligned} \tag{5.3.5}$$

where $\hat{C}_{ik}^{(s+1)} := \mathbb{E}(Z_i^{\beta\intercal}\Sigma_k^{\beta\beta}Z_i^\beta|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$, $\hat{A}_{ik}^{(s+1)} := \mathbb{E}(Z_i^{\gamma\intercal}\Sigma_k^{\gamma\gamma}Z_i^\gamma|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$ and $\hat{\Sigma}_k^{-1,**}$ indicated the corresponding block of the inverted matrix $\hat{\Sigma}_k^{-1}$ with respect to the notation in Equation 5.2.1. Again, to compute $\hat{C}_{ik}^{(s)}$ we will make use of the elements of $\hat{S}_{ik}^{\beta,(s+1)}$, while for $\hat{A}_{ik}^{(s)}$ we will use the elements of $\hat{S}_{ik}^{\gamma,(s+1)}$.

Summing up, this means that computing $\mathbb{E}(\log\mathcal{L}_C(\boldsymbol{\Theta};\mathbf{Y},\mathbf{Z},\boldsymbol{\ell})|\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})$ requires to compute:

- $\mathbb{E}(\ell_{ik}|\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) =: \hat{\tau}_{ik}^{(s+1)}$,
- $\mathbb{E}(z_i^\beta|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) =: \hat{m}_{ik}^{\beta,(s+1)}$,
- $\mathbb{E}(z_i^\beta z_i^{\beta\intercal}|\ell_{ik},\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) =: \hat{S}_{ik}^{\beta,(s+1)}$, whose elements are required for the computation of $\hat{D}_{ik}^{(s+1)}$ and $\hat{C}_{ik}^{(s+1)}$,

- $\mathbb{E}(z_i^\gamma | \ell_{ik} = 1, \mathbf{Y}, \hat{\boldsymbol{\Theta}}^{(s)}) =: \hat{m}_{ik}^{\gamma,(s+1)}$,
- $\mathbb{E}(z_i^\gamma z_i^{\gamma\intercal} | \ell_{ik}, \mathbf{Y}, \hat{\boldsymbol{\Theta}}^{(s)}) =: \hat{S}_{ik}^{\gamma,(s+1)}$, whose elements are required for the computation of $\hat{B}_{ik}^{(s+1)}$ and $\hat{A}_{ik}^{(s+1)}$.

### 5.3.4 M-step

The updated for the parameters at step $(s+1)$ are given by

$$\hat{\pi}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}}{N} \quad , \quad \hat{M}_k^{(s+1)} = \frac{1}{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}} \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix}, \tag{5.3.6}$$

$$\begin{aligned} \hat{\Sigma}_k^{(s+1)} = \frac{1}{T \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}} \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} \times \\ \left( \begin{bmatrix} Y_i^\alpha \hat{\Phi}_k^{-1(s)} Y_i^{\alpha\intercal} & Y_i^\alpha \hat{\Phi}_k^{-1(s)} \hat{M}_{ik}^{\beta,\intercal(s+1)} & Y_i^\alpha \Phi_k^{-1(s)} \hat{M}_{ik}^{\gamma,\intercal(s+1)} \\ \hat{M}_{ik}^{\beta,(s+1)} \hat{\Phi}_k^{-1(s)} Y_i^{\alpha\intercal} & \hat{D}_{ik}^{(s+1)} & \hat{M}_{ik}^{\beta,(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_{ik}^{\gamma,\intercal(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \hat{\Phi}_k^{-1(s)} Y_i^{\alpha\intercal} & \hat{M}_{ik}^{\gamma,(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_{ik}^{\beta,\intercal(s+1)} & \hat{B}_{ik}^{(s+1)} \end{bmatrix} - \right. \\ \left. \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix}^\intercal - \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\intercal(s+1)} + \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\intercal(s+1)} \right) \end{aligned} \tag{5.3.7}$$

The update formulas of the two covariance matrices are interconnected:

$$\begin{aligned} \hat{\Phi}_k^{(s+1)} = \frac{1}{J \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}} \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} \Bigg( & Y_i^{\alpha\intercal} \hat{\Sigma}_k^{-1,\alpha\alpha} Y_i^\alpha + Y^{\alpha\intercal} \hat{\Sigma}_k^{-1,\alpha\beta} \hat{M}_{ik}^{\beta,(s+1)} + Y^{\alpha\intercal} \hat{\Sigma}_k^{-1,\alpha\gamma} \hat{M}_{ik}^{\gamma,(s+1)} + \\ & \hat{M}_{ik}^{\beta,(s+1)\intercal} \hat{\Sigma}_k^{-1,\beta\alpha} Y_i^\alpha + \hat{C}_{ik}^{(s+1)} + \hat{M}_{ik}^{\beta,(s+1)\intercal} \hat{\Sigma}_k^{-1,\beta\gamma} \hat{M}_{ik}^{\gamma,(s+1)} + \\ & \hat{M}_{ik}^{\gamma,(s+1)\intercal} \hat{\Sigma}_k^{-1,\gamma\alpha} Y_i^\alpha + \hat{M}_{ik}^{\gamma,(s+1)\intercal} \hat{\Sigma}_k^{-1,\gamma\beta} \hat{M}_{ik}^{\beta,(s+1)} + \hat{A}_{ik}^{(s+1)} - \\ & \hat{M}_k^{\intercal(s+1)} \hat{\Sigma}_k^{-1(s+1)} \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix} - \begin{bmatrix} Y_i^\alpha \\ \hat{M}_{ik}^{\beta,(s+1)} \\ \hat{M}_{ik}^{\gamma,(s+1)} \end{bmatrix}^\intercal \hat{\Sigma}_k^{-1(s+1)} \hat{M}_k^{(s+1)} + \\ & \hat{M}_k^{\intercal(s+1)} \hat{\Sigma}_k^{-1(s+1)} \hat{M}_k^{(s+1)} \Bigg) \end{aligned} \tag{5.3.8}$$

### 5.3.5 Convergence

Because of the MCMC use during the E-step, the property of monotone increase of the observed log-likelihood does not hold for our model (McLachlan and Krishnan, 2008, Ruth, 2024). Therefore,

to asses convergence we use moving average estimation on the observed log-likelihood.

Let $l_o^s$ the observed log-likelihood at step $s$, then our convergence criterion is

$$\left| \frac{\left(\frac{1}{w_1}\sum_{i=s-w_1+1}^{s} l_o^i\right) - \left(\frac{1}{w_2}\sum_{i=s-w_1-w_2+1}^{s-w_1} l_o^i\right)}{\frac{1}{w_2}\sum_{i=s-w_1-w_2+1}^{s-w_1} l_o^i} \right| < \varepsilon.$$

In the following, $\varepsilon = 1 \cdot 10^{-3}$ is chosen.

### 5.3.6 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

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

where $\nu_k$ is the total number of model parameters:

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

and $\mathcal{L}_O(\boldsymbol{\Theta}; \boldsymbol{Y})$ is the observed likelihood defined in Equation 5.2.5.
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.

## 5.4 Simulation study

This section presents numerical experiments on simulated data in order to illustrate the behavior of the proposed model. First, we aim at studying the influence of the initialization procedure and sample size in estimating the partition and the parameters. Secondly, the robustness to different noise ratio in the data concerning the clustering, the parameters estimation and the model selection. Finally, we compare the MMM model to a its continuous counterpart (MMN) when used on mixed data treated like continuous data.

### 5.4.1 Simulation Setup

20 different samples have been simulated for increasing number of units $N \in \{100, 500, 1000\}$, with number of clusters $K = 2$, number of variables $J = 4$, number of times $T = 3$ and cluster proportions $\pi = (0.6, 0.4)$. The $J$ variables are of mixed type, with the first variable being continuous, the second being ordinal, the third being binary and the fourth being a counting variable. The ordinal variable has 5 levels. Each sample has been drawn from a matrix-variate Gaussian and then transformed according to the model described in Section 5.2. The distributions parameters were chosen as following: identity matrices for the covariance matrices $\Phi_k$ and $\Sigma_k$ for each cluster, while mean matrices $M_k$ 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 of two mean matrices as described in Table C1.

Moreover, three scenarios are derived from this setting by adding some noise by adding to the underlying continuous latent matrix of a percentage $\tau$ of units a reasonable level of noise, generated according to a centered Gaussian with variance equal to 0.5, allocated to the two clusters proportionally to the clusters' size: 0% (scenario 1), 10% (scenario 2), 20% (scenario 3). The two different kinds of initialization described in Section 5.3.2 have been tested. 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. Concerning the simulation done via `stan`, we set the chain iterations to 500, of which half as burn-in, for 3 different chains.

### 5.4.2 Computational time

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 for $K = 1$ is about 5 seconds for $N = 100$ and about 30 seconds second for $N = 1000$.

### 5.4.3 Influence of initialization & sample size

We first aim at studying the ability of the algorithm to recover the simulated model depending on the type of initialization of the EM algorithm and on the size of the sample. Figure 5.1 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 ($\approx 0.85$) according to the simulation scheme is represented by a horizontal line. The boxplots show some small differences in the median values of the ARI measurements between the two initialization methods, with the random multistart initialization performing moderately better than its Kmeans++ counterpart, both in terms of median and of lower variability. When the sample size is sufficiently large, the result that stems from the multistart initialization almost attains the optimal ARI.

However, while the random multistart initialization seems to perform marginally better than the Kmeans++ from a partitioning point of view, it is noteworthy to consider the computational and temporal burden of the former compared to the latter. One might consider whether the tade-off is worthy on a case-by-case base.

In addition, we measure their performance also by computing the Mean Absolute Percentage Error (MAPE) on their estimation of the distribution parameters. We recall that 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 $\theta$ it is expressed through the formula:

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

where $\hat{\theta}_i$ is the estimated parameter and $\theta_i$ 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 5.2.

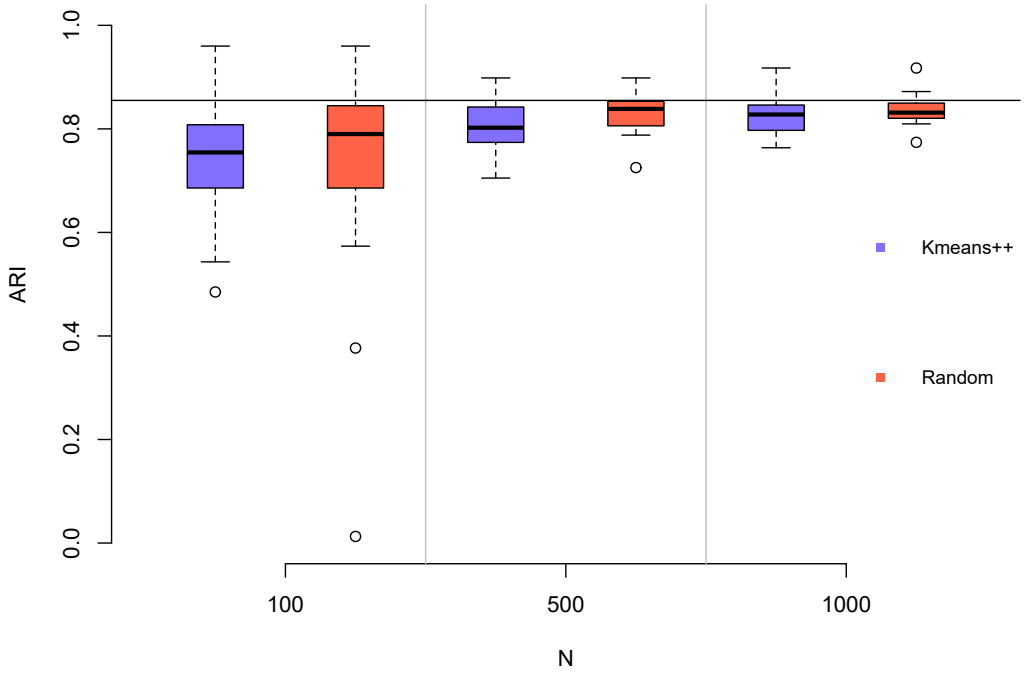
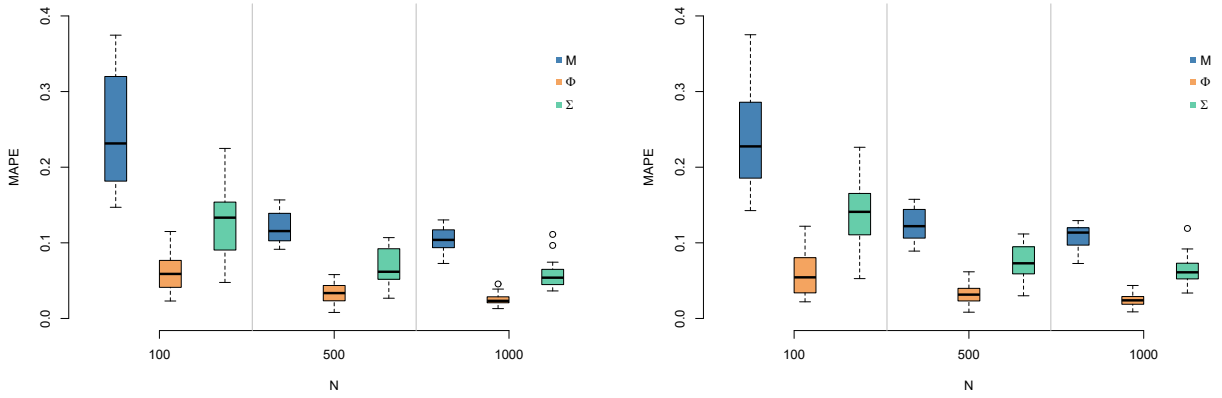



Figure 5.1: Influence of initialization and sample size. The horizontal line represents the estimated Bayesian error.

Regarding the parameters estimates with respect to the different initialization strategies, there is no clear difference in terms of MAPE, with the Kmeans having thinly better rendering.

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, and the median MAPE values appear to reach a stable value already for $N = 500$, while the values improve further for $N = 1000$, especially in terms of lower variability.

Overall, while the random multistart seems to provide a better partitioning, the difference is so tiny that for the rest of our experiments we will use just the Kmeans++ initialization strategy, which is less time consuming.
Last, while the general magnitude of the MAPE can seem important, it is important to recall that we use a convergence tolerance of $\varepsilon = 1 \cdot 10^{-3}$, as per Section 5.3.5. Better results can be found by reducing the $\varepsilon$, while making the execution more time-consuming.




(a) Kmeans++ init

(b) Random init.

Figure 5.2: MAPE for increasing sample size

### 5.4.4 Robustness to noise

As written in Section 5.4.1, we also simulated some noisy data to study the robustness of the model in presence of some noise. ARI for different noise proportions were measured and the results are visible in Figure 5.3. 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 slightly decreases as the level of noise increases, indicating that the model is actually disturbed by the noise.
Interestingly but somehow to be expected, when N increases the model is more disturbed by the noise, as there are more units affected by it. Moreover, the noise affects the allocation estimation of non-noisy units as well, and again this estimation seems to be more disturbed for a larger N.

### 5.4.5 Model selection

Following the setup described in Section 5.4.1, by varying $N \in \{100, 500, 1000\}$ 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 5.3.6. The results are shown in Table 5.1.

When the sample size increases, the model converges toward the true model. However, as clearly visible in the table, the model tend to underestimate the true number of clusters when the sample size is not sufficiently large, probably due to the insufficient number of units to estimate the parameters from. Interestingly, some noise actually helps the model to recover the true model.

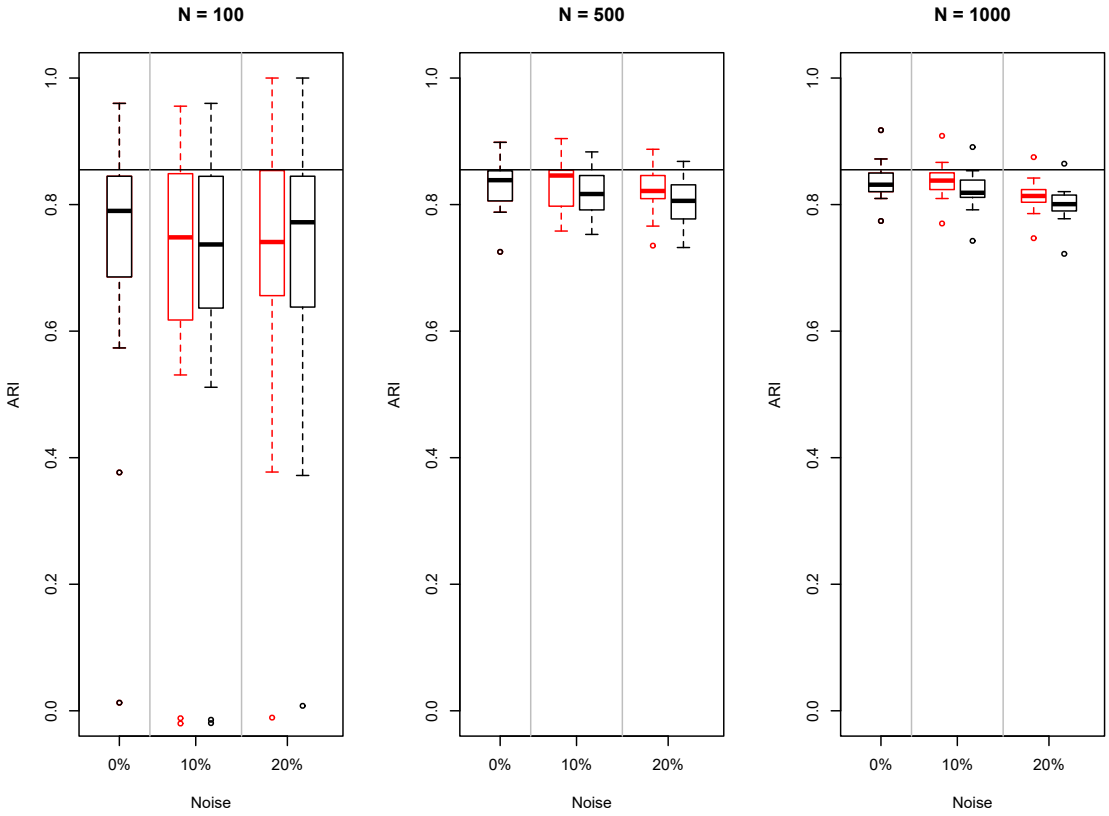



Figure 5.3: ARI for increasing noise proportions and increasing N.
In red the ARI for non-noisy units, in black for all of them.

### 5.4.6 Comparison with continuous conterpart

Finally, we compared the MMM model to the classical Mixture of Matrix-Normals (MMN) model, mentioned in Section 5.1.1, in a version implemented by us following Viroli, 2011a. Essentially, this means treating all the different data-type equally as continuous, as it is often done by practitioners, but keeping the advantages of the matrix-variate structure. The results of the partitioning is presented in Figure 5.4. The hyper-parameters of the competitors have been set to be similar to the one of the MMM in terms of initialization, convergence and covariance matrix parametrization. The MMM model clearly outperforms the MMN model, independently from the sample size.

In Figure 5.5, we compared the MAPE values for the parameters estimation between MMM and MMN models. The difference between the two is severe, especially for the matrix of means $M$ and the covariance matrix $\Sigma$, while moderate for $\Phi$, due to the constraint on its determinant.The important difference of the results of the MMN model against the MMM one with respect to $M$ and $\Sigma$ is probably due to the count data-type variable. Indeed, without assuming the latent log-normal distribution, the values likely become too out of scale compared to the others.

| $N/K$ | Scenario $\tau = 0$ | | | | Scenario $\tau = 0.1$ | | | | Scenario $\tau = 0.2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | **2** | 3 | 4 | 1 | **2** | 3 | 4 | 1 | **2** | 3 | 4 |
| 100 | **14** | 6 | 0 | 0 | **13** | 7 | 0 | 0 | **12** | 8 | 0 | 0 |
| 500 | 0 | **19** | 1 | 0 | 0 | **20** | 0 | 0 | 0 | **20** | 0 | 0 |
| 1000 | 0 | **17** | 3 | 0 | 0 | **17** | 2 | 1 | 0 | **18** | 2 | 0 |

Table 5.1: Frequency of selection of each model K by the model through BIC among the 20 simulated data sets, for increasing N. The actual value for K is 2. Kmeans++ initialization. In bold the true value for K and the most frequent K detected for each noise ratio and sample size.

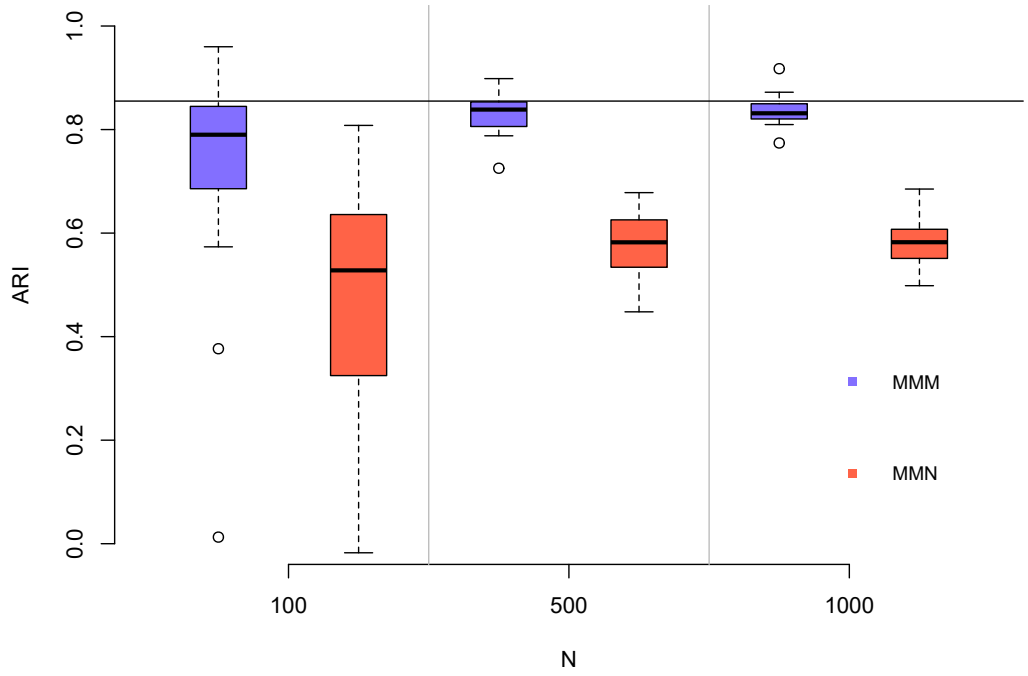
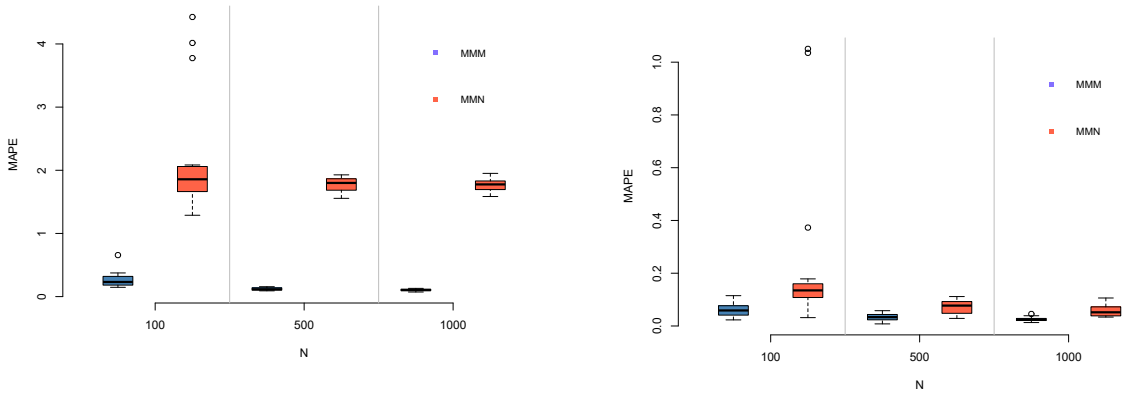


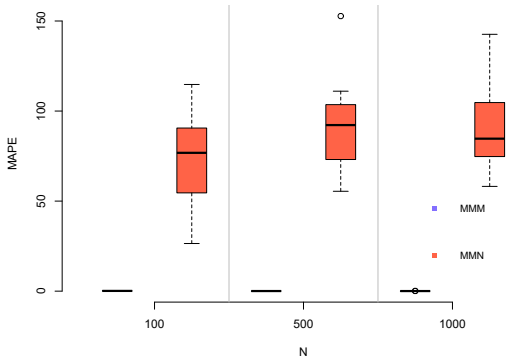

Figure 5.4: Comparison between the MMM and MOM models.

Globally, the experiments described above proved that the MMM model is able to retrieve the true partitioning and to infer the true parameters, even in presence of moderate noise. It is also able select the appropriate number of clusters through BIC when presented with enough sample units. We proved that our model outperforms its continuous matrix-variate counterpart when the latter is used to model mixed-type data, as often done by practitioners. We are now confident enough to apply the MMM model on real-world data.




(a) MAPE for $M$

(b) MAPE for $\Phi$




(c) MAPE for $\Sigma$

Figure 5.5: MAPE results for parameter matrices. MMM vs MMN. Kmeans++ init. Note the difference in the scales.

## 5.5 Real-world application

### 5.5.1 Data description

The S&P500 index is a stock exchange index tracking the stock performances of 500 of the largest companies listed on stock exchange market in the United States, where each company is weighted by its market capitalization. It is one of the most commonly followed equity indices and the companies included in the index represent 80% of the total market capitalization of U.S. public companies. While investors are commonly interested in the index in its entirety, it is often the case for them to be interested in the composing companies, reputed as the best ones to invest in, in order to create specific portfolios to be used for long-term investments and wealth management.

We collected data concerning companies composing the S&P500 stock market index. Specifically, we focused on the time period going to the beginning of 2019 to the end of 2023, hence encompassing the period immediately prior and the one immediately succeeding to the COVID-19 pandemic, which went from the 30th of January 2020 to the 5th of May 2023 according to the World Health Organization (WHO) (Sarker et al., 2023). The objective of our study is to cluster companies according to their stock behavior during the pandemic period, in order to discover similar patterns during a shock period and possibly adjust our stock portfolio accordingly.

For our analysis, we collected for each year and for each listed company the following variables:

- **LogReturns:** continuous variable. The logarithm of the yearly return of the stock. The return is computed as the relative percentage change in the stock adjusted closing price between the first trading day versus the last trading day of the year. In financial analysis, log-returns are often employed instead of the simple returns as log returns have an infinite support (compared to simple returns which are lower-bounded by -100) and as they take into account the compounding effect, making them more suitable for long-term analysis.

- **Grades:** ordinal variable. The investment grade of the stock expressed by institutional investment banks. Specifically, for this study we considered the grades given by "Bank of America", since it is the institution that releases them for most of the companies of the S&P500. The grades have three levels: "Underperform", "Neutral" and "Buy". Grades are given multiple times in a year and not all at the same time, so we considered their mode for each fiscal year.

- **Dividends:** binary variable. Whether the stock gave right to a dividend during the fiscal year or not, regardless of the amount .

- **Volume:** count variable. The total volume of stocks exchanged during the year. Because of the high amount of stocks that are traded during a year, we decided to count per millions of stocks exchanged. Therefore, each counted units will represent a million stocks traded. Generally, securities with higher volume are more liquid.

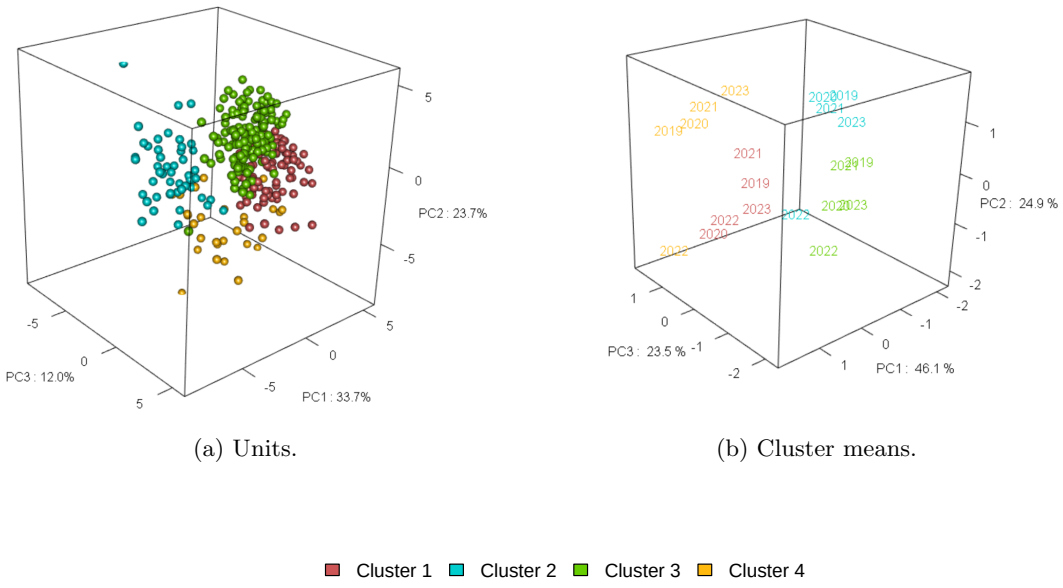
The data were collected making use of the Python package `pyfinance`, which downloads the data from the website "yahoo!finance".
However, grades were not released by Bank of America for all the S&P500 companies for the entirety of the time window of our study, but just on 330 of them. We decided to reduce our survey to them. So, overall our dataset is composed of four mixed variables (continuous, ordinal, binary and count) collected for 330 observations over 5 time points (years from 2019 to 2023 included). We reorganized these data into a list of matrices.

### 5.5.2 Results

After performing our clustering algorithm with a number of clusters $K$ ranging from 1 to 8 using Kmeans++ initialization, the model with the lowest BIC is the one with K = 4 (Fig. D2). The number of units in each cluster is respectively of 94, 50, 154 and 32.
The estimated parameters are reported in Table D1 for the mean $M$, Table D2 for the time covariances $\Phi$ and in Table D3 for the variable covariances $\Sigma$. In addition, the correlation matrices

are represented by correlation plots in Figs. 5.8 and 5.9, respectively. The tickers of the companies allocated to each cluster is reported at Table D4.
In Figure 5.7 the evolution for the observed outcomes for each cluster is showed.
Moreover, by using the "Global Industry Classification Standard" (GICS) industrial taxonomy developed by "Standard & Poor's" (S&P), we represented the sector composition of each cluster in Figure D1.

By performing a PCA the latent continuous embedding computed by the MMM model, we can represents the 330 units as in Figure 5.6a. A $3D$ representation is provided. For this representation, the temporal structure has been discarded and we have transformed our latent embedding for the units from $4 \times 5$-dimensional matrices to 20-dimensional vectors. On the other hand, Figure 5.6b represents cluster means at each of the 5 years. Such plot allows to visualize the time evolution of each cluster.




(a) Units.

(b) Cluster means.

Figure 5.6: Units and cluster means represented through PCA.

### 5.5.3 Interpretation

First, we can get a preliminary idea by looking at Figure 5.6a, where PCA was computed on the estimated units' latent space representation. As we can see, of the 4 clusters, units belonging to Cluster 1 and Cluster 3 are more concentrated and closer to each other in the latent space, while the ones belonging to Cluster 2 and 4 are more spread out. This is confirmed by looking at Figure 5.6b, where the clusters of Cluster 1 and Cluster 3 means occupy adjacent space regions.

In the following, we give a summary description for Cluster 4, which we deemed be the most

interesting. Interpretations for the other clusters can be found in Appendix .2.

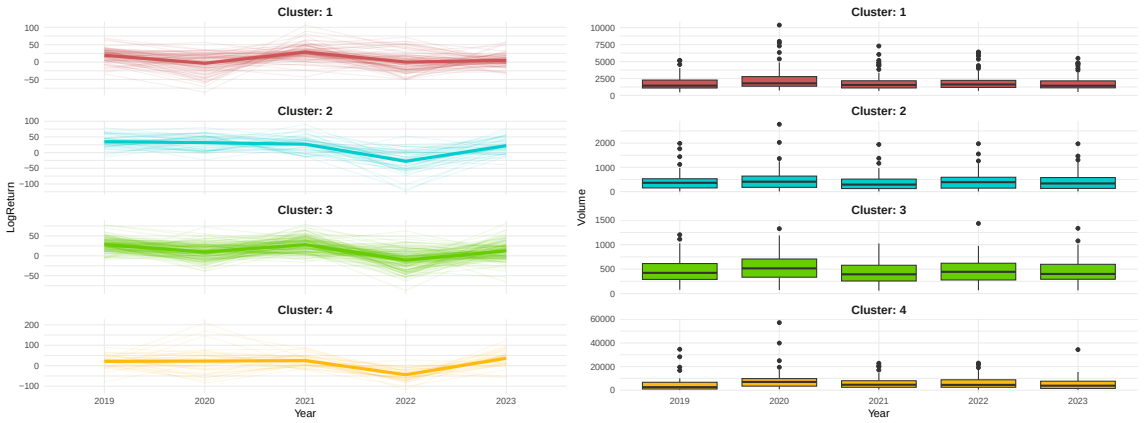
- **Cluster 4**: 32 units.
  - **Means**: the cluster is qualified by generally constant strong values for LogReturn, with the exepction of 2022, where the cluster has the lowest negative value. The cluster also has the second highest values for Grade and the highest values for Volume. The values for Dividend are small and fluctuate around zero in time suggesting heterogeneity in the cluster regarding this variable.
  - **Correlation in time**: the cluster is characterized the second strongest correlations overall.
  - **Correlation among variables**: the main feature of the cluster concerning variables correlation is the absence of a negative correlation between volume and LogReturn, while a weak negative correlation between Dividend and LogReturn is estimated.

  Cluster 4 is defined by its high value of the variable Volume compared to the others. The values of LogReturn are more stable in time, except for 2022. The value of Dividens float around zero, and Figure 5.7 shows us that the dividend distribution is almost evenly split for most of the years.
  It is also the only cluster to have a negative correlation between Dividend and LogReturn, implying that stocks with higher returns are also the ones with no dividends. This paradox can be explained by looking at the sector distribution in Figure 5.7 : a majority of the companies whose stocks are allocated to Cluster 4 belong to sectors such as "Technology" and "Consumer Cyclical", and when we look at Table D4 we realize it includes companies like Amazon, Tesla, Netflix, Nvidia, AMD and Moderna, that do not allocate dividends but prefer to reinvest their profit in R&D.

## 5.6 Conclusions

In this work we have presented a novel approach for modeling longitudinal mixed-type data with unobserved heterogeneity. The model presented does not require the conditional independence assumption. The matrix-variate structure allows for a more parsimonious modeling of multivariate longitudinal data than other models in the literature. Also, it can explicitly model the temporal structure and the association among the responses, that can vary among clusters. An MCMC-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 mixed-type data and to get clusters that are easy to interpret and to work with even by non-statisticians in a real-world example.

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, 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

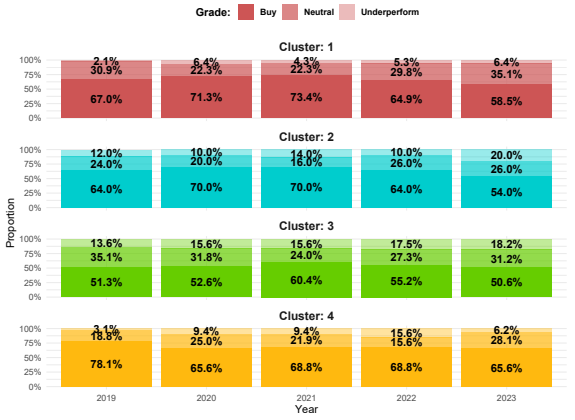



(a) LogReturns




(b) Volume


(c) Dividends




(d) Grades

Figure 5.7: Observed variables values for each cluster. Note that for graphical reason in plots (a) and (b) the company NVIDIA has been removed from the set, due to its out-of-scale values compared to the others companies.

and parsimonious models, as done for example in Anderlucci and Viroli, 2015 and in Sarkar, Zhu, Melnykov, and Ingrassia, 2020a. Another solution to this problem can be the one proposed by Cappozzo, Casa, and Fop, 2024.
Similarly, the matrix-variate structure is not just inherent to multivariate longitudinal data, but can actually be found in many other applications. The MMM model can be employed in such cases as well, with minimal adjustments required.
Moreover, EM algorithm can be leveraged to extend the model to deal with incomplete data under the missing at random (MAR).

Finally, one could as well think of employing, with proper adjustments, different underlying continuous distributions, such as heavy-tailed (Tomarchio, Punzo, and Bagnato, 2020), skewed


(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 2

Figure 5.8: Clusters' corr-plots among years.

(Gallaugher and McNicholas, 2018, Melnykov and Zhu, 2018) or t-student (Doğru, Bulut, and Arslan, 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).


(a) Cluster 1




(b) Cluster 2




(c) Cluster 3




(d) Cluster 4

Figure 5.9: Clusters' corr-plots among variables.

# Appendices

# .1 E-step computations

Here we will expand the computations presented in Section 5.3.1.

For Equation 5.3.3, the matrix-variate expectation related to count data can be computed by defining $z_i^\gamma \in \mathbb{R}^{GT\times 1}$ as the vectorized version of $Z_i^\gamma$ and computing

$$\hat{m}_{ik}^{\gamma,(s+1)} := \mathbb{E}(z_i^\gamma|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) = \tag{.1.1}$$
$$= \int_{\mathbb{R}} z_i^\gamma \cdot \frac{\prod_t^T\prod_g^G \mathbb{P}(y_{igt}^\gamma|\exp(z_{igt}^\gamma))\cdot \mathcal{MN}_{GT}(z_i^\gamma|\text{vec}(M_k^{(s),\gamma|\alpha,\beta}),\Sigma_k^{(s),\gamma|\alpha,\beta}\otimes\Phi_k^{(s)})}{\int_{\mathbb{R}}\prod_t^T\prod_g^G \mathbb{P}(y_{igt}^\gamma|\exp(z_{igt}^\gamma))\cdot \mathcal{MN}_{GT}(z_i^\gamma|\text{vec}(M_k^{(s),\gamma|\alpha,\beta}),\Sigma_k^{(s),\gamma|\alpha,\beta}\otimes\Phi_k^{(s)})dz_i^\gamma} dz_i^\gamma.$$

This integral does not have any close form solution, so we resort to numerically compute it through the No-U-Turn sampler implemented in the R package `Rstan`.

Then, $\hat{M}_{ik}^{\gamma,(s+1)} := \text{vec}^{-1}_{G\times T}(\hat{m}_{ik}^{\gamma,(s+1)})$, $\text{vec}^{-1}_{G\times T}$ being the inverse of the vectorization function, i.e. the function mapping from a $GT$-dimensional vector to a $O\times T$ matrix.

The matrix-variate expectation related to categorical data can be computed by defining $z_i^\beta \in \mathbb{R}^{OT\times 1}$ as the vectorized version of $Z_i^\beta$ and computing

$$\hat{m}_{ik}^{\beta,(s+1)} := \mathbb{E}(z_i^\beta|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) = \int_{\Omega_r} z_i^\beta \mathcal{MN}_{OT}(z_i^\beta|\text{vec}(M_k^{(s),\beta|\alpha}),\Sigma_k^{(s),\beta|\alpha}\otimes\Phi_k^{(s)})dz_i^\beta \tag{.1.2}$$

through the use of a Gibbs sampler to sample from a truncated multivariate normal distribution.

Then, as we did for count data; we map the estimated values back to a matrix form as $\hat{M}_{ik}^{\beta,(s+1)} := \text{vec}^{-1}_{O\times T}(\hat{m}_{ik}^{\beta,(s+1)})$.

For Equation 5.3.4, to compute $D_{ik}^{(s)}$, we start by defining $\hat{\varphi}_{k,gd}^{(s)}$ as the $(g,d)^{th}$ element of $\hat{\Phi}_k^{-1(s)}$. Then, the $(h,t)^{th}$ element of $Z_i^\beta\Phi_k^{-1}Z_i^{\beta\intercal}$ would be $\sum_{d=1}^T\sum_{g=1}^T z_{i,hg}^\beta\hat{\varphi}_{k,gd}^{(s)}z_{i,td}^\beta$ and we would get

$$\hat{D}_{ik}^{(s)} := \mathbb{E}(Z_i^\beta\Phi_k^{-1}Z_i^{\beta\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})) =$$
$$= \left(\sum_{d=1}^T\sum_{g=1}^T \hat{S}_{ik,[(g-1)O+h,(d-1)O+t]}^{\beta,(s+1)}\hat{\varphi}_{k,gd}^{(s)}\right)_{h,t}, \tag{.1.3}$$

where we make use of the the elements of

$$\hat{S}_{ik}^{\beta,(s+1)} := \mathbb{E}(z_i^\beta z_i^{\beta\intercal}|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) = \int_{\Omega_r} z_i^\beta z_i^{\beta\intercal}\mathcal{MN}_{OT}(z_i^\beta|\text{vec}(M_k^{(s),\beta|\alpha}),\Sigma_k^{(s),\beta|\alpha}\otimes\Phi_k^{(s)})dz_i^\beta. \tag{.1.4}$$

The samples generated to calculate the first moment $m_{ik}^{\beta,(s+1)}$ can be reused to compute the matrix $\hat{S}_{ik}^{(s+1)}$, that can be approximated by calculating the inner product of the vectors used to compute $m_{ik}^{\beta,(s+1)}$ then calculating the sample mean of these inner products.

Similarly, for $\hat{B}_{ik}^{(s)}$ the $(h,t)^{th}$ element of $Z_i^{\gamma}\Phi_k^{-1}Z_i^{\gamma\intercal}$ would be $\sum_{d=1}^{T}\sum_{g=1}^{T} z_{i,hg}^{\gamma}\varphi_{k,gd}z_{i,td}^{\gamma}$ and we would get

$$\begin{aligned}\hat{B}_{ik}^{(s)} &:= \mathbb{E}(Z_i^{\gamma}\Phi_k^{-1}Z_i^{\gamma\intercal}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y})) = \\ &= \left(\sum_{d=1}^{T}\sum_{g=1}^{T}\hat{S}_{ik,[(g-1)G+h,(d-1)G+t]}^{\gamma,(s+1)}\hat{\varphi}_{k,gd}^{(s)}\right)_{h,t}, \end{aligned} \tag{.1.5}$$

where we make use of the the elements of

$$\begin{aligned}\hat{S}_{ik}^{\gamma,(s+1)} &:= \mathbb{E}(z_i^{\gamma}z_i^{\gamma\intercal}|\ell_{ik}=1,\mathbf{Y},\hat{\boldsymbol{\Theta}}^{(s)}) = \\ &= \int_{\mathbb{R}} z_i^{\gamma}z_i^{\gamma\intercal}\cdot\frac{\prod_t^T\prod_g^G\mathbb{P}(y_{igt}^{\gamma}|\exp(z_{igt}^{\gamma}))\cdot\mathcal{MN}_{GT}(z_i^{\gamma}|\text{vec}(M_k^{(s),\gamma|\alpha,\beta}),\Sigma_k^{(s),\gamma|\alpha,\beta}\otimes\Phi_k^{(s)})}{\int_{\mathbb{R}}\prod_t^T\prod_g^G\mathbb{P}(y_{igt}^{\gamma}|\exp(z_{igt}^{\gamma}))\cdot\mathcal{MN}_{GT}(z_i^{\gamma}|\text{vec}(M_k^{(s),\gamma|\alpha,\beta}),\Sigma_k^{(s),\gamma|\alpha,\beta}\otimes\Phi_k^{(s)})dz_i^{\gamma}}dz_i^{\gamma}. \end{aligned} \tag{.1.6}$$

As before, the samples generated to calculate the first moment $\hat{m}_{ik}^{\gamma,(s+1)}$ can be reused to compute the matrix $\hat{S}_{ik}^{\gamma,(s+1)}$ by calculating the mean of the inner product between them.

Finally, for Equation 5.3.5, let us define by $\hat{\sigma}_{k,gd}^{(s),\beta\beta}$ the $(g,d)^{th}$ element of the block $\hat{\Sigma}_k^{-1(s),\beta\beta}$. Then, the $(h,t)^{th}$ element of $Z_i^{\beta\intercal}\Sigma_k^{\beta\beta}Z_i^{\beta}$ is $\sum_{d=1}^{O}\sum_{g=1}^{O} z_{i,gh}\hat{\sigma}_{k,gd}^{(s),\beta\beta}z_{i,dt}$, and we get

$$\begin{aligned}\hat{C}_{ik}^{(s)} &:= \mathbb{E}(Z_i^{\beta\intercal}\Sigma_k^{\beta\beta}Z_i^{\beta}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) = \\ &= \left(\sum_{d=1}^{O}\sum_{g=1}^{O}\hat{S}_{ik,[(h-1)J+g,(t-1)J+d]}^{\beta,(s+1)}\hat{\sigma}_{k,gd}^{(s),\beta\beta}\right)_{h,t}. \end{aligned} \tag{.1.7}$$

For $\hat{A}_{ik}^{(s)}$, let indicate by $\hat{\sigma}_{k,gd}^{(s),\gamma\gamma}$ the $(g,d)^{th}$ element of block $\hat{\Sigma}_k^{-1(s),\gamma\gamma}$. Then, the $(h,t)^{th}$ element of $Z_i^{\gamma\intercal}\Sigma_k^{\gamma\gamma}Z_i^{\gamma}$ is $\sum_{d=1}^{O}\sum_{g=1}^{O} z_{i,gh}\hat{\sigma}_{k,gd}^{(s),\gamma\gamma}z_{i,dt}$, and we get

$$\begin{aligned}\hat{A}_{ik}^{(s)} &:= \mathbb{E}(Z_i^{\gamma\intercal}\Sigma_k^{\gamma\gamma}Z_i^{\gamma}|\ell_{ik}=1,\hat{\boldsymbol{\Theta}}^{(s)},\mathbf{Y}) = \\ &= \left(\sum_{d=1}^{O}\sum_{g=1}^{O}\hat{S}_{ik,[(h-1)J+g,(t-1)J+d]}^{\gamma,(s+1)}\hat{\sigma}_{k,gd}^{(s),\gamma\gamma}\right)_{h,t}. \end{aligned} \tag{.1.8}$$

## .2 Cluster interpretation

In this section interpretations for other clusters referred in Section 5.5.3 are given.

- **Cluster 1**: 94 units.
  - **Means**: the cluster means show the highest values for Dividend and Grade, and the second highest for Volume. The results for LogReturn are more shaded: the cluster has the lowest mean for 2019, 2020 (the only one negative for that year) and 2023. At the same time, it is the only cluster to have non-negative LogReturns for 2022.

  - **Correlation in time**: the cluster is characterized by a fading and weaker correlations among times than other clusters, especially regarding 2022 to the previous years.
  - **Correlation among variables**: the cluster is characterized by feeble correlations among Returns, Grade and Dividend, yet these correlations are stronger than in other clusters. Some soft negative correlations are estimated between Volume, Grade and Return.

  We can describe Cluster 1 as the cluster of more "traditional" stocks. Stocks belonging to this cluster have good grades, usually grant dividends and are among the most exchanged, ensuring good liquidity.
  By looking at Figure D1a, we can notice that the cluster is the ones with more variety of composing sectors. This might explain why it is the only cluster that experienced a fall in LogReturns in 2020, at the height of the COVID-19 pandemic and of the consequent lock-downs, which had a major impact on more traditional sectors. Figure 5.7 suggests that indeed the stocks gave right to dividends even for the entirety of them in 2019 and 2020. We can also point out to the fact that during 2020 and 2021 the percentage of stocks marked as "Buy" for this cluster increased, probably in view of the end of toughest pandemic period and in light of the lower prices of the stocks. The grades distribution changes during 2022 and 2023 mostly in favor of "Neutral". The correlations among times suggest that the behaviour is less constant in time with respect of the other clusters. Moreover, the negative correlations between Volume, LogReturns and Grade may indicate that the increase in volume exchange is generally related to selling, as the volume increase when grades and returns decreases.

- **Cluster 2**: 50 units.

  - **Means**: the cluster has the highest means regarding LogReturns for the first two years and the second most important negative value for 2022. It is the only cluster with relatively strong negative values for Dividend. It has also the lowest values for Volume.
  - **Correlation in time**: it is the cluster with the strongest positive correlation in time.
  - **Correlation variables**: the cluster is characterized by the presence of weak correlation between LogReturn, Grade ans Dividend, and of a weak negative correlation between Volume and LogReturn.

  Cluster 2 has the main characteristics to be the only cluster with negative values for Dividend. A look at Figure 5.7 shows us that indeed that almost none of the stocks allocated to the cluster gave right to a dividend, a situation that slightly improves in 2023. The low values for Volume compared to the other clusters indicate that the stocks in this cluster are among the less exchanged. The grades distribution show that there is a high percentage of stocks marked as "Buy" until 2021, but it decreases and in 2023 the cluster has the highest percentage of stocks marked as "Underperform". 2022 appears to be a bad year for the stocks belonging to the cluster, but with the expect of this year the cluster has the most stable values for LogReturn. The sector composition of the cluster shows a dominance of the sectors "Healthcare" and "Technology", which might explain the good performance during the pandemic, as these sectors were among the ones to actually profit during the pandemic. The same reason might explain the 2022 performance, where staff lay-offs and decrease in investments due to over-investments during the pandemics hit particularly the IT sector.

- **Cluster 3**: 154 units.
  - **Means**: the cluster has the second highest means for LogReturns for 2019 and 2021, and the lowest negative value for 2022. It has the second highest values for Dividend and the second smallest values for Volume.
  - **Correlation in time**: the cluster has the overall strong positive correlations in time.
  - **Correlation among variables**: the cluster is mainly characterized by the weak negative correlation between Volume and LogReturn, and the absence of other meaningful correlations.

  Cluster 3 can be seen as cluster between Cluster 1 and Cluster 2: both Volume and Grade have values in between the two, and the same can be almost be said for LogReturn. The main exception to this description is Dividend, since for Cluster 3 the values are high, and if we look at Figure 5.7 almost 100% of the stocks gave right to a dividend. Besides, the percentage of stocks releasing dividends is surprisingly stable over time.
  Moreover, concerning the variables Grade, the cluster is the one with the smallest percentage of stocks classified as "Buy", while it has the highest percentage of stocks marked as "Neutral" among all the clusters.
  Its main sector is "Industrials", but we can see from Figure D1 that its composition is diversified, more like Cluster 1 than Cluster 2.

## .3 Simulations

Table C1: Means matrices for simulation

| **Cluster 1** | T1 | T2 | T3 |
|---|---|---|---|
| V1 | 1.75 | 1.75 | 1.75 |
| V2 | 1.75 | 1.75 | 1.75 |
| V3 | -0.25 | -0.25 | -0.25 |
| V4 | 1 | 1 | 1 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.75 | 2.75 | 2.75 |
| V2 | 2.75 | 2.75 | 2.75 |
| V3 | 0.25 | 0.25 | 0.25 |
| V4 | 2.5 | 2.5 | 2.5 |

## .4 Real data

Table D1: Clusters' means over time. The estimated parameter $\hat{\pi} = (0.287, 0.156, 0.460, 0.096)$

| **Cluster 1** | 2019 | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|---|
| Return | 19.77 | -3.34 | 28.72 | 0.03 | 5.07 |
| Grade | 3.93 | 4.16 | 4.07 | 4.07 | 3.71 |

| | | | | | |
|---|---|---|---|---|---|
| Dividend | 4.07 | 4.04 | 3.44 | 3.51 | 3.58 |
| Volume | 7.35 | 7.59 | 7.38 | 7.44 | 7.33 |
| **Cluster 2** | 2019 | 2020 | 2021 | 2022 | 2023 |
| Return | 34.69 | 31.77 | 26.73 | -27.7 | 22.21 |
| Grade | 3.00 | 3.24 | 3.20 | 3.04 | 2.69 |
| Dividend | -1.57 | -1.92 | -2.05 | -2.00 | -1.68 |
| Volume | 5.72 | 5.82 | 5.52 | 5.70 | 5.66 |
| **Cluster 3** | 2019 | 2020 | 2021 | 2022 | 2023 |
| Return | 27.92 | 9.54 | 28.06 | -10.87 | 12.34 |
| Grade | 3.09 | 3.19 | 3.44 | 3.23 | 3.08 |
| Dividend | 3.34 | 3.65 | 3.58 | 3.81 | 4.01 |
| Volume | 6.02 | 6.15 | 5.91 | 6.00 | 5.98 |
| **Cluster 4** | 2019 | 2020 | 2021 | 2022 | 2023 |
| Return | 21.29 | 22.72 | 24.87 | -43.95 | 36.45 |
| Grade | 4.52 | 3.71 | 3.71 | 3.52 | 3.46 |
| Dividend | 0.50 | 0.38 | -0.88 | -0.54 | -0.13 |
| Volume | 8.04 | 8.84 | 8.44 | 8.52 | 8.33 |

Table D2: Clusters' time covariances

| **Cluster 1** | 2019 | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|---|
| 2019 | 1.36 | 0.94 | 0.75 | 0.61 | 0.66 |
| 2020 | 0.94 | 1.64 | 0.81 | 0.61 | 0.67 |
| 2021 | 0.75 | 0.81 | 1.51 | 0.99 | 0.77 |
| 2022 | 0.61 | 0.61 | 0.99 | 1.68 | 0.83 |
| 2023 | 0.66 | 0.67 | 0.77 | 0.83 | 1.3 |
| **Cluster 2** | 2019 | 2020 | 2021 | 2022 | 2023 |
| 2019 | 2.25 | 1.97 | 1.81 | 1.69 | 1.79 |
| 2020 | 1.97 | 2.42 | 1.94 | 1.78 | 1.92 |
| 2021 | 1.81 | 1.94 | 2.54 | 2.02 | 2.07 |
| 2022 | 1.69 | 1.78 | 2.02 | 2.69 | 2.2 |
| 2023 | 1.79 | 1.92 | 2.07 | 2.2 | 2.73 |
| **Cluster 3** | 2019 | 2020 | 2021 | 2022 | 2023 |
| 2019 | 1.63 | 1.28 | 1.2 | 1.06 | 1.06 |
| 2020 | 1.28 | 1.81 | 1.3 | 1.09 | 1.09 |
| 2021 | 1.2 | 1.3 | 1.93 | 1.29 | 1.27 |
| 2022 | 1.06 | 1.09 | 1.29 | 1.95 | 1.29 |
| 2023 | 1.06 | 1.09 | 1.27 | 1.29 | 1.9 |
| **Cluster 4** | 2019 | 2020 | 2021 | 2022 | 2023 |
| 22019 | 2.61 | 1.57 | 1.59 | 1.5 | 1.61 |
| 2020 | 1.57 | 2.06 | 1.42 | 1.16 | 1.27 |
| 2021 | 1.59 | 1.42 | 1.95 | 1.48 | 1.48 |
| 2022 | 1.5 | 1.16 | 1.48 | 1.92 | 1.52 |
| 2023 | 1.61 | 1.27 | 1.48 | 1.52 | 1.97 |

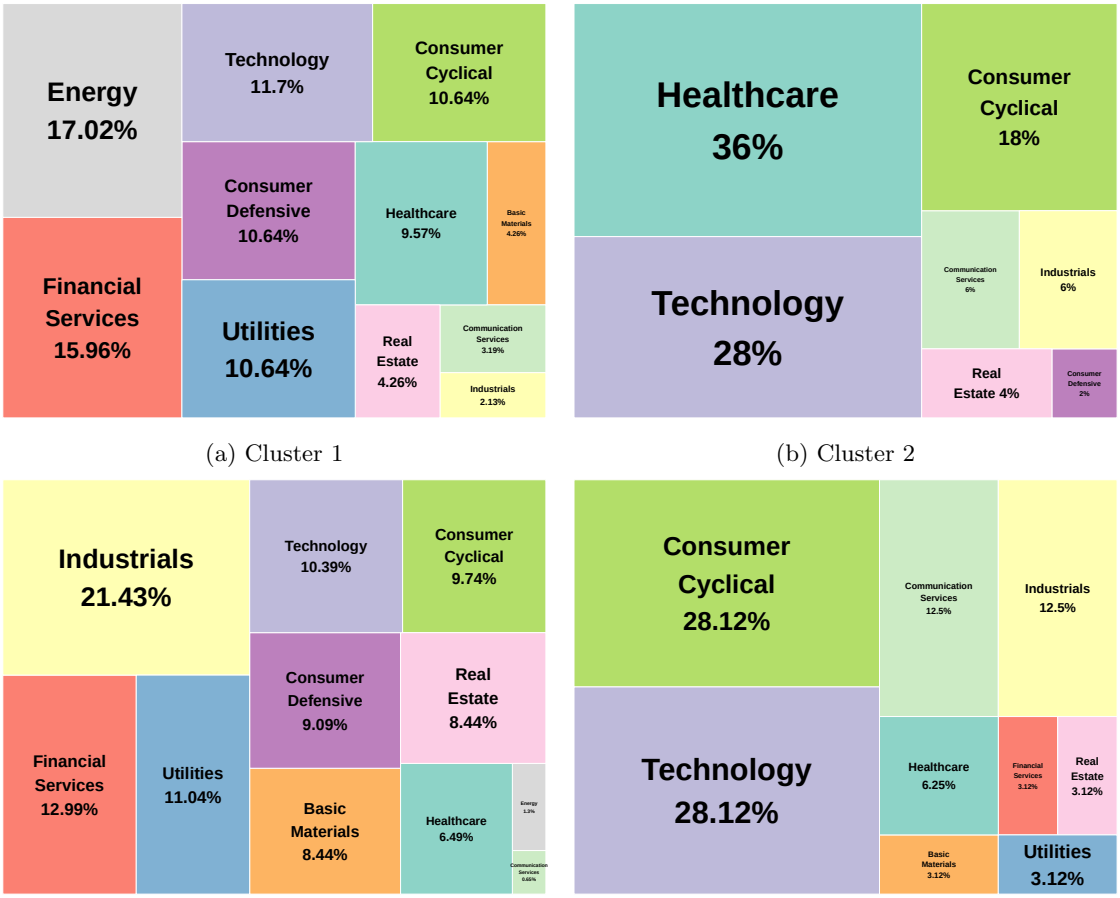
Table D3: Clusters' variables covariances

| **Cluster 1** | Return | Grade | Dividend | Volume |
|---|---|---|---|---|
| Return | 589.69 | 6.4 | 6.22 | -0.87 |
| Grade | 6.4 | 3.36 | 0.92 | -0.06 |
| Dividend | 6.22 | 0.92 | 3.74 | -0.03 |
| Volume | -0.87 | -0.06 | -0.03 | 0.1 |
| **Cluster 2** | Return | Grade | Dividend | Volume |
| Return | 976.02 | 5.09 | 3.85 | -1.82 |
| Grade | 5.09 | 2.02 | 0.54 | 0 |
| Dividend | 3.85 | 0.54 | 1.98 | -0.01 |
| Volume | -1.82 | 0 | -0.01 | 0.15 |
| **Cluster 3** | Return | Grade | Dividend | Volume |
| Return | 521.79 | 1.15 | 1.89 | -0.91 |
| Grade | 1.15 | 3.6 | 0.97 | -0.01 |
| Dividend | 1.89 | 0.97 | 3.89 | -0.01 |
| Volume | -0.91 | -0.01 | -0.01 | 0.07 |
| **Cluster 4** | Return | Grade | Dividend | Volume |
| Return | 2378.8 | 4.74 | -8.1 | 1.2 |
| Grade | 4.74 | 2.64 | 0.61 | -0.05 |
| Dividend | -8.1 | 0.61 | 2.31 | 0.04 |
| Volume | 1.2 | -0.05 | 0.04 | 0.32 |

Table D4: Stocks' tickers in each cluster

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|
| ABBV, AES, AIG | ADBE, ADSK, ANET | A, ACGL, ACN | AAL, AAPL, AMD |
| AMAT, BK, BKR | APTV, AZO, BKNG | ADI, ADP, AEE | AMZN, AVGO, BA |
| BMY, BX, CFG | BSX, CBRE, CDNS | ALB, ALL, AME | CMG, CRWD, CZR |
| CL, CMCSA, CNP | CNC, CRL, CSGP | APD, AVB, AVY | DAL, DIS, EXPE |
| COP, CSCO, CSX | CTLT, DECK, DLTR | AWK, AXP, BALL | F, FCX, GM |
| CVS, CVX, D | DVA, DXCM, EPAM | BAX, BBY, BEN | GOOGL, INTC, MRNA |
| DD, DOW, DVN | EW, FFIV, FSLR | BWA, BXP, CAT | MSFT, NCLH, NFLX |
| EBAY, EOG, EXC | FTNT, GNRC, HOLX | CBOE, CDW, CE | NVDA, PCG, PFE |
| FANG, FE, FIS | HSIC, IDXX, IQV | CF, CHD, CHRW | PYPL, RCL, SPG |
| FITB, FOXA, GILD | ISRG, IT, KMX | CLX, CME, CMI | T, TSLA, UAL |
| GLW, HAL, HBAN | LH, LULU, MHK | CMS, COST, CPB | UBER, WDC |
| HD, HPE, HPQ | MOH, MTCH, MTD | CPT, CTAS, CTSH | |
| IBM, IVZ, JPM | NOW, NVR, ORLY | DE, DFS, DGX | |
| KDP, KHC, KIM | PANW, PAYC, PTC | DHI, DOV, DPZ | |
| KMI, KR, LLY | QRVO, TDG, TMUS | DRI, DTE, DUK | |
| LOW, LUV, LVS | TTWO, URI, VRTX | EA, ED, EFX | |
| MCHP, MDLZ, MDT | WAT, WST | EIX, EL, ELV | |
| MET, MGM, MO | | EMN, EMR, EQR | |
| MOS, MPC, MRK | | ES, ESS, ETN | |

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|
| MRO, NEE, NEM | | ETR, EVRG, EXR | |
| NI, NKE, O | | FDS, FDX, FMC | |
| OKE, ORCL, OXY | | FTV, GD, GPC | |
| PEP, PG, PM | | GS, HCA, HES | |
| PPL, QCOM, RF | | HII, HON, HRL | |
| SBUX, SCHW, SLB | | HSY, HUM, ICE | |
| SO, SYF, TGT | | INTU, IP, IRM | |
| TJX, TPR, TXN | | ITW, JBHT, JBL | |
| UNH, USB, V | | JNPR, K, KKR | |
| VICI, VLO, VST | | KLAC, KMB, LEN | |
| VZ, WBA, WFC | | LMT, LNT, LRCX | |
| WMB, WY, WYNN | | LW, LYB, MA | |
| XOM | | MAS, MCD, MCK | |
| | | MLM, MMC, MMM | |
| | | NDAQ, NRG, NSC | |
| | | NTAP, NTRS, NUE | |
| | | NXPI, ODFL, PAYX | |
| | | PCAR, PEG, PH | |
| | | PHM, PKG, PLD | |
| | | PNC, PNW, PPG | |
| | | PSA, RL, ROK | |
| | | RSG, SBAC, SHW | |
| | | SNA, SRE, STLD | |
| | | STT, STZ, SWK | |
| | | SWKS, SYY, TAP | |
| | | TER, TMO, TRGP | |
| | | TROW, TRV, TSCO | |
| | | TSN, TXT, UDR | |
| | | UHS, UNP, UPS | |
| | | VMC, VRSK, VTR | |
| | | WAB, WEC, WELL | |
| | | WM, WRB, XEL | |
| | | ZTS | |




(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 2

Figure D1: Clusters' sectors composition

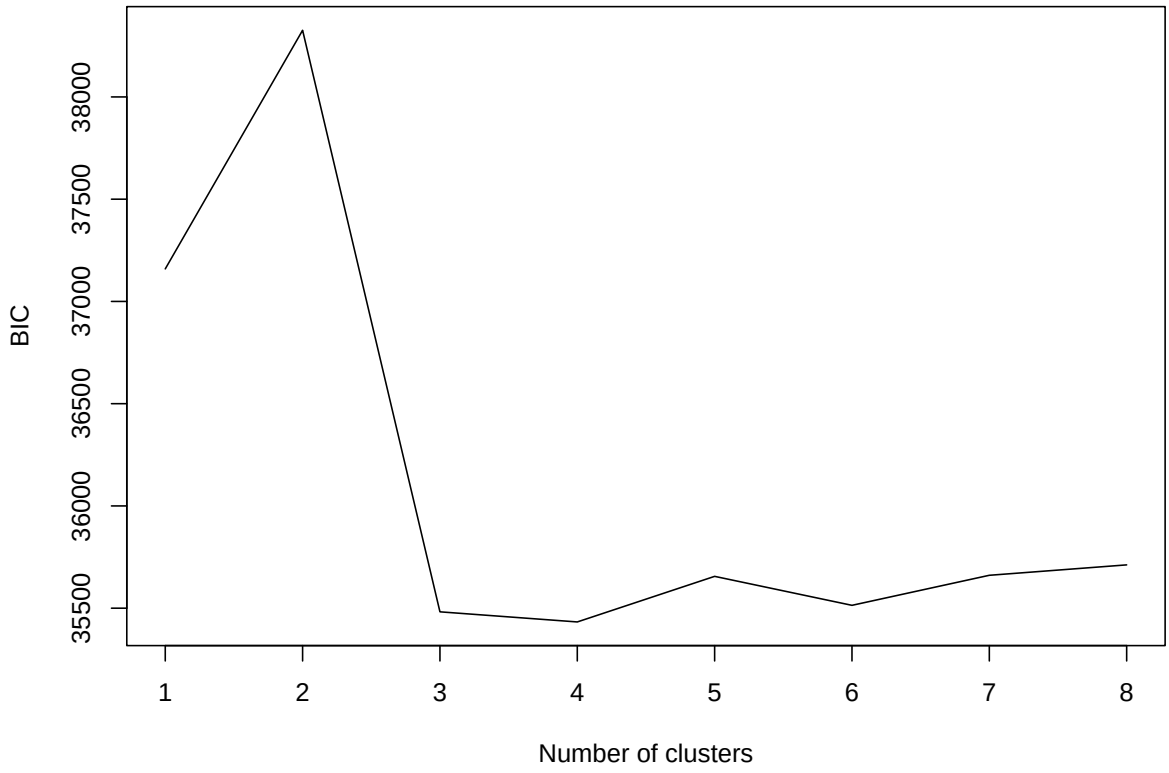



Figure D2: Visualization of BIC for K as results of application on real data. Kmeans++ initialization.

# Chapter 6

# Conclusions

This thesis has explored the challenging task of clustering longitudinal mixed-type data, a critical area in statistics and machine learning with broad applications in social sciences, economics, and medicine. The primary objective was to develop an easily interpretable model-based clustering algorithm capable of handling the complexities inherent in such data. This concluding chapter summarizes the key findings and contributions from the preceding chapters, to finally talk about the future directions this thesis, and the research it inspired, can take.

## 6.1 Model-Based Clustering Framework

In Chapters 1 and 2, we introduced the scientific question that inspired our work and consequently dug into the foundational concepts of clustering and model-based clustering, emphasizing the advantages of probabilistic approaches in cluster analysis. We introduced Finite Mixture Models (FMMs), which represent the probability density function of random variables as a weighted sum of component densities. This framework allows for a principled approach to model "learning" (i.e. inference), leveraging the statistical properties of probability distributions to address practical questions such as determining the number of clusters, detecting outliers, assessing uncertainty in clustering but also enhance clusters' interpretability. The Expectation-Maximisation (EM) algorithm was presented as the main method for estimating the parameters of FMMs. This iterative procedure alternates between the Expectation (E) step and the Maximisation (M) step until convergence, making it a useful tool for maximum likelihood estimation in the presence of latent variables or incomplete data. The chapter also detailed the Gaussian Mixture Model (GMM), a common type of FMM where the component densities are multivariate Gaussian distributions. The GMM's flexibility and ease of use make it suitable for various applications, and its parameters, including mean vectors and covariance matrices, can be estimated using the EM algorithm.

## 6.2 Literature Survey on Clustering Longitudinal Mixed-Type Data

Chapter 3 provided a survey of existing methods for model-based clustering of longitudinal mixed-type data. The chapter highlighted the challenges and solutions proposed in the literature, empha-

sizing the need for models that can jointly address temporal evolution and the heterogeneity of data types. Traditional clustering methods often fail to capture the dependencies between observations over time, leading to suboptimal results. Parametric approaches, such as mixed-effects models and latent Markov models, offer a framework to incorporate temporal dependencies but often suffer from over-parametrization, computational complexity and also . The survey covered various models, including Latent Markov Models (LMMs) and their extensions, Mixed Hidden Markov Models (MHMMs), Growth Mixture Models (GMMs), and Mixtures of Matrix-Normals (MMN), which we used as base for our new model. As we saw, each model has its strengths and limitations. The chapter also discussed the challenges of handling mixed-type data, with the majority of models assuming conditional independence at least between variables of different types given the cluster membership to simplify computations, as it happens for Latent Class Model (LCM) and clustMD. Some more complex models, such as Mixture of Multivariate Generalized Linear Mixed Models (MMGLMM) and Latent Class Linear Mixed Models (LCLMMs) solve this problem by means of random effects. As we saw, this technique is effective, but increases the model complexity and often need to be paired with more advanced computational methods. Moreover, it may be complex for non-statisticians to fully grasp the interpretative nuances that random effects entail.

## 6.3 The MOM model

Chapter 4 we moved our first step towards our final model, by focusing on developing a clustering algorithm for longitudinal ordinal data, probably the most common type of data in questionnaires in social sciences. The proposed Mixture of Ordinal Matrices (MOM) model assumes that an ordinal variable is the discretization of an underlying latent continuous variable, which follows a Gaussian distribution. This model relies on an underlying Mixture of Matrix-Normals (MMN), which is able to account simultaneously for within- and between-time dependence structures. The chapter detailed the EM algorithm for parameter estimation and approaches to deal with computational challenges. The MOM model was evaluated through synthetic data, demonstrating its estimation abilities and advantages over competitors. Lastly, a real-world application concerning changes in eating behaviours during the Covid-19 pandemic in France showcased the model's practical utility. The application on real marketing data and the collaboration with social sciences researchers confirmed the goodness of our approach not only regarding the ability to manage the intended data, but also concerning the simplicity of the interpretation of its parameters and clusters.

## 6.4 The MMM model

Finally, Chapter 5 extended the proposed framework to handle longitudinal mixed-type data, which includes continuous, ordinal, binary, categorical, and count data. The proposed model, an extension of the MOM model, accounts for the correlation structure among different types of data without requiring the conditional independence assumption and the matrix-variate structure allows for a more parsimonious modelling of multivariate longitudinal data than other models in the literature. The chapter detailed the MCMC-EM algorithm for parameter estimation and the model's ability to handle the complexities of mixed-type data. Again, the model was evaluated through synthetic data, demonstrating its estimation abilities and robustness to noise. A real-world application on financial data proved the applicability of the MMM model to a more vast ensemble of purposes, corroborating that the model retains a high degree of interpretability even in the context of mixed-

type data and is indeed able to detect correlation structures among variables of different types and to efficiently model the time behaviours of the data.

## 6.5 Future Directions

However, the proposed model has some limitations. In the paper presenting it, 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, since, as the number of clusters increases, the number of parameters to estimate can still become 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 and Viroli, 2015 and in Sarkar, Zhu, Melnykov, and Ingrassia, 2020a. Another solution to this problem can be the one proposed by Cappozzo, Casa, and Fop, 2024.
Similarly, the matrix-variate structure is not just inherent to multivariate longitudinal data, but can actually be found in many other applications. The MMM model can be employed in such cases as well, with minimal adjustments required. Moreover, EM algorithm can be leveraged to extend the model to deal with incomplete data under the missing at random (MAR).

Moreover, one could as well think of employing, with proper adjustments, different underlying continuous distributions, such as heavy-tailed (Tomarchio, Punzo, and Bagnato, 2020), skewed (Gallaugher and McNicholas, 2018, Melnykov and Zhu, 2018) or t-student (Doğru, Bulut, and Arslan, 2016) distributions to endow the clustering model with different desired properties.

Finally, as mentioned in Chapter 3, we deem models to be useful for researchers in other domains only if they are provided with the right instruments to implement the fitting of those models, mostly through open-access softwares such as R libraries. As we aim for our work to be broadly accessible, we are working on a R package to let practitioners easily implement the MMM model, providing an user-friendly interface and a variety of functions to also allow for the analysis of subsets of all the possible data-types.

# Bibliography


[1] Karl Pearson. "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material". In: *Philosophical Transactions of the Royal Society of London. Series A* 185 (1894), pp. 343–414.

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

[3] 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.

[4] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[5] ES Pearson. "Some aspects of the geometry of statistics: the use of visual presentation in understanding the theory and application of mathematical statistics". In: *Journal of the Royal Statistical Society. Series A (General)* 119.2 (1956), pp. 125–146. DOI: 10.2307/2342880.

[6] Edward W Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". In: *biometrics* 21 (1965), pp. 768–769.

[7] A. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2 (Jan. 1967), pp. 260–269. DOI: 10.1109/TIT.1967.1054010.

[8] 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.

[9] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (Jan. 1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[10] 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.

[11] 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.

[12] 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.

[13] Nan M. Laird and James H. Ware. *Random-Effects Models for Longitudinal Data*. Dec. 1982. DOI: 10.2307/2529876.

[14] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.

[15] B.S. Everitt. *Introduction to Latent Variable Models*. Chapman and Hall, 1984.

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

[17] 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.

[18] J. D. Kalbfleisch and J. F. Lawless. "The Analysis of Panel Data under a Markov Assumption". In: *Journal of the American Statistical Association* (Dec. 1985). ISSN: 1047-8195. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478195.

[19] D Michael Titterington, Smith Afm, Adrian FM Smith, UE Makov, et al. *Statistical analysis of finite mixture distributions*. Vol. 198. Chichester: John Wiley & Sons Incorporated, 1985.

[20] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[21] Trevor Hastie and Robert Tibshirani. "Generalized Additive Models". In: *Statistical Science* 1.3 (Aug. 1986), pp. 297–310. ISSN: 0883-4237. DOI: 10.1214/ss/1177013604.

[22] P. McCullagh. *Generalized Linear Models*. Andover, England, UK: Taylor & Francis, Jan. 1989. ISBN: 978-0-20375373-6. DOI: 10.1201/9780203753736.

[23] T. Hammar. *Democracy and the nation state : aliens, denizens and citizens in a world of international migration*. Aldershot, UK: Gower Publishing Company, 1990.

[24] Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1990.

[25] Rolf Langeheine and Frank Van De Pol. "A Unifying Framework for Markov Modeling in Discrete Space and Discrete Time". In: *Sociological Methods & Research* 18.4 (May 1990), pp. 416–441. ISSN: 0049-1241. DOI: 10.1177/0049124190018004002.

[26] Frank Van de Pol and Rolf Langeheine. "Mixed Markov latent class models". In: *Sociological methodology* (1990), pp. 213–247.

[27] B. H. Juang and L. R. Rabiner. "Hidden Markov Models for Speech Recognition". In: *Technometrics* (Aug. 1991). ISSN: 1048-4833. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484833.

[28] 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.

[29] Linda M. Collins and Stuart E. Wugalter. "Latent Class Models for Stage-Sequential Dynamic Latent Variables". In: *Multivariate Behavioral Research* (Jan. 1992). DOI: 10.1207/s15327906mbr2701_8.

[30] Jeffrey D. Banfield and Adrian E. Raftery. "Model-Based Gaussian and Non-Gaussian Clustering". In: *Biometrics* 49.3 (Sept. 1993), pp. 803–821.

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

[32] Gilles Celeux and Gérard Govaert. "Gaussian parsimonious clustering models". In: *Computational Statistics and Data Analysis* 28.5 (1995), pp. 781–793.

[33] Keith Humphreys. "Classification error adjustments for female labour force transitions using a latent Markov chain with random effects". In: *Applications of Latent Class and Latent Trait Models in the Social Sciences* (1997), pp. 370–380.

[34] Harald Waldrauch and Christoph Hofinger. "An index to measure the legal obstacles to the integration of migrants". In: *Journal of Ethnic and Migration Studies* 23.2 (1997), pp. 271–285. DOI: 10.1080/1369183X.1997.9976590.

[35] C. Fraley and A. E. Raftery. "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis". In: *The Computer Journal* 41.8 (1998), pp. 578–588. ISSN: 0010-4620. DOI: 10.1093/comjnl/41.8.578.

[36] Cornelis Arnoldus Groenendijk, Elspeth Guild, Halil Dogan, et al. *Security of residence of long-term migrants: A comparative study of law and practice in European countries*. Strasbourg: Council of Europe, 1998.

[37] Keith Humphreys. "The latent markov chain with multivariate random effects: An evaluation of instruments measuring labor market status in the british household panel study". In: *Sociological methods & research* 26.3 (1998), pp. 269–299.

[38] Douglas S Massey et al. *Worlds in motion: understanding international migration at the end of the millennium*. Oxford, Clarendon Press, 1998.

[39] Bengt Muthén and Kerby Shedden. "Finite mixture modeling with mixture outcomes using the EM algorithm". In: *Biometrics* 55.2 (1999), pp. 463–469.

[40] Cristophe Biernacki, Gilles Celeux, and Gérard Govaert. "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7 (Aug. 2000), pp. 719–725. DOI: 10.1109/34.865189.

[41] Stephen Castles and Alastair Davidson. *Citizenship and migration: Globalization and the politics of belonging*. New York: Routledge, 2000. DOI: 10.4324/9781003061595.

[42] P. D'Urso. "Dissimilarity measures for time trajectories". In: *Stat. Methods Appl.* 9.1-3 (2000), pp. 53–83. DOI: 10.1007/BF03178958.

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

[44] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2000.

[45] D. Peel and G. J. McLachlan. "Robust mixture modelling using the t distribution". In: *Statistics and Computing* 10.4 (Oct. 2000), pp. 339–348. ISSN: 1573-1375. DOI: 10.1023/A:1008981510081.

[46] Carl de Boor. *A Practical Guide to Splines*. Springer, 2001. ISBN: 978-0-387-95366-3.

[47] Bengt Muthén et al. "General growth mixture modeling for randomized preventive interventions". In: *Biostatistics (Oxford, England)* 3.4 (Dec. 2002), pp. 459–475. ISSN: 1465-4644. DOI: 10.1093/biostatistics/3.4.459. eprint: 12933592.

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

[49] Christophe Biernacki and Stéphane Chrétien. "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM". In: *Statistics & Probability Letters* 61.4 (2003), pp. 373–382. DOI: 10.1016/S0167-7152(02)00396-6.

[50] M. Freudenber. *Composite Indicators of Country Performance*. Paris: OECD Publishing, 2003. DOI: 10.1787/405566708255.

[51] Gérard Govaert and Mohamed Nadif. "Clustering with block mixture models". In: *Pattern Recognition* 36.2 (Feb. 2003), pp. 463–473. ISSN: 0031-3203. DOI: 10.1016/S0031-3203(02)00074-2.

[52] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". In: *Machine Learning* 52.1 (July 2003), pp. 91–118. ISSN: 1573-0565. DOI: 10.1023/A:1023949509487.

[53] Roger E. Millsap and Jenn Yun-Tein. "Assessing Factorial Invariance in Ordered-Categorical Measures". In: *Multivariate Behavioral Research* 39.3 (2004), pp. 479–515. DOI: 10.1207/S15327906MBR3903_4.

[54] Bengt Muthén and David Kaplan. "Handbook of quantitative methodology for the social sciences". In: *Latent variable analysis: growth mixture modeling and related tech niques for longitudinal data. Newbury Park: Sage* (2004), pp. 345–68.

[55] Rinus Penninx and Marco Martiniello. "Integration processes and policies: State of the art and lessons". In: *Citizenship in European cities: Immigrants, local politics and integration policies*. 1 st. Aldershot, UK: Ashgate, 2004, pp. 139–164.

[56] 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.

[57] G. Govaert and M. Nadif. "An EM algorithm for the block mixture model". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.4 (Mar. 2005), pp. 643–647. DOI: 10.1109/TPAMI.2005.69.

[58] 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.

[59] M. Nardo, M. Saisana, A. Saltelli, and S. Tarantola. "Tools for composite indicators building". In: *European Commission, Ispra* 15.1 (2005), pp. 19–20.

[60] Cécile Proust and Hélène Jacqmin-Gadda. "Estimation of linear mixed models with a mixture of distribution for the random effects". In: *Computer Methods and Programs in Biomedicine* 78.2 (May 2005), pp. 165–173. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2004.12.004.

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

[62] Keith G Banting, Will Kymlicka, et al. *Multiculturalism and the welfare state: Recognition and redistribution in contemporary democracies*. Oxford University Press on Demand, 2006. DOI: 10.1093/acprof:oso/9780199289172.001.0001.

[63] Rachel MacKay Altman. "Mixed Hidden Markov Models". In: *Journal of the American Statistical Association* (Mar. 2007). ISSN: 1071-0171. URL: https://www.tandfonline.com/doi/epdf/10.1198/016214506000001086?needAccess=true.

[64] 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.

[65] Christian Dustmann and Ian P Preston. "Racial and economic factors in attitudes to immigration". In: *The BE Journal of Economic Analysis & Policy* 7.1 (2007).

[66] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.

[67] Tsung I Lin, Jack C Lee, and Shu Y Yen. "Finite mixture modelling using the skew normal distribution". In: *Statistica Sinica* 17.3 (July 2007), pp. 909–927.

[68] 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.

[69] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Chichester, England, UK: John Wiley & Sons, Ltd., Apr. 2007. ISBN: 978-0-47120170-0. DOI: 10.1002/9780470191613.

[70] J Niessen et al. *Migrant integration policy index*. Tech. rep. Brussels: British Council and Migration Policy Group, 2007.

[71] Tihomir Asparouhov and Bengt Muthen. "Multilevel mixture models". In: *Advances in Latent Variable Mixture Models*. Information Age Publishing, 2008, pp. 27–51. DOI: 10.1201/9781420011579-15.

[72] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge, England, UK: Cambridge University Press, July 2008. ISBN: 978-0-52185225-8. DOI: 10.1017/CBO9780511790485.

[73] Rolando De la Cruz-Mesia, Fernando A Quintana, and Guillermo Marshall. "Model-based clustering for longitudinal data". In: *Computational Statistics & Data Analysis* 52.3 (2008), pp. 1441–1457.

[74] Bettina Grün and Friedrich Leisch. "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters". In: *Journal of Statistical Software* 28.4 (2008), pp. 1–35. DOI: 10.18637/jss.v028.i04.

[75] Tony Jung and K. A. S. Wickrama. "An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling". In: *Social and Personality Psychology Compass* 2.1 (Jan. 2008), pp. 302–317. ISSN: 1751-9004. DOI: 10.1111/j.1751-9004.2007.00054.x.

[76] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2008.

[77] Bengt Muthen and Tihomir Asparouhov. "Growth mixture modeling: Analysis with non-Gaussian random effects". In: *Longitudinal Data Analysis*. Chapman and Hall/CRC, Aug. 2008, pp. 157–180. DOI: 10.1201/9781420011579-15.

[78] OECD. *Handbook on Constructing Composite Indicators. Methodology and User Guide*. 2008.

[79] Peder J Pedersen, Mariola Pytlikova, and Nina Smith. "Selection and network effects—Migration flows into OECD countries 1990–2000". In: *European Economic Review* 52.7 (2008), pp. 1160–1186.

[80] Heather Andruff et al. "Latent class growth modelling: a tutorial". In: *Tutorials in quantitative methods for psychology* 5.1 (2009), pp. 11–24.

[81] Jan Niessen and Thomas Huddleston. *Legal frameworks for the integration of third-country nationals*. Leiden, The Netherlands: Brill| Nijhoff, 2009. DOI: 10.1163/ej.9789004170698.i-246.

[82] Nilam Ram and Kevin J Grimm. "Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups". In: *International journal of behavioral development* 33.6 (2009), pp. 565–576.

[83] Nilam Ram and Kevin J. Grimm. "Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups". In: *International journal of behavioral development* 33.6 (2009), p. 565. DOI: 10.1177/0165025409343765.

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

[85] Sara Wallace Goodman. "Integration requirements for integration's sake? Identifying, categorising and comparing civic integration policies". In: *Journal of Ethnic and Migration Studies* 36.5 (2010), pp. 753–772. DOI: 10.1080/13691831003764300.

[86] 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.

[87] Hanno Petras and Katherine Masyn. "General growth mixture analysis with antecedents and consequences of change". In: *Handbook of quantitative criminology* (2010), pp. 69–100.

[88] Svend-Erik Skaaning. "Measuring the rule of law". In: *Political Research Quarterly* 63.2 (2010), pp. 449–460. DOI: 10.1177/1065912909346745.

[89] Graeme Boushey and Adam Luedtke. "Immigrants across the US federal laboratory: Explaining state-level innovation in immigration policy". In: *State Politics & Policy Quarterly* 11.4 (2011), pp. 390–414. DOI: 10.2307/41575833.

[90] 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.

[91] Lynette Hunt and Murray Jorgensen. "Clustering mixed data". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.4 (2011), pp. 352–361. DOI: 10.1002/widm.33.

[92] Antonello Maruotti. "Mixed Hidden Markov Models for Longitudinal Data: An Overview". In: *International Statistical Review* 79.3 (Dec. 2011), pp. 427–454. ISSN: 0306-7734. DOI: 10.1111/j.1751-5823.2011.00160.x.

[93] Bauböck Rainer and Helbling Marc. *Which Indicators are Most Useful for Comparing Citizenship Policies?* EUI-RSCAS Working Papers 54. European University Institute (EUI), Robert Schuman Centre of Advanced Studies (RSCAS), 2011.

[94] 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.

[95] 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.

[96] G. Zincone, R. Pennix, and M. Borkert. "Migration policymaking in Europe: The dynamics of actors and contexts in past and present". In: *Migration Policymaking in Europe* (2011). DOI: 10.2307/j.ctt46n178.

[97] Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data*. New York: Chapman and Hall/CRC, 2012. DOI: 10.1201/b13246.

[98] Ahmed M. Gad and Rasha B. El Kholy. "Generalized Linear Mixed Models for Longitudinal Data". In: *International Journal of Probability and Statistics* 1.3 (2012), pp. 41–47. DOI: 10.5923/j.ijps.20120103.03.

[99] Ruud Koopmans, Ines Michalowski, and Stine Waibel. "Citizenship rights for immigrants: National political processes and cross-national convergence in Western Europe, 1980–2008". In: *American journal of sociology* 117.4 (2012), pp. 1202–1245. DOI: 10.1086/662707.

[100] 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.

[101] Ulrike Von Luxburg, Robert C Williamson, and Isabelle Guyon. "Clustering: Science or art?" In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 65–79.

[102] Tjalling J. Ypma. "Historical Development of the Newton–Raphson Method". In: *SIAM Review* (Feb. 2012). DOI: 10.1137/1037125.

[103] Mathias Czaika and Hein De Haas. "The effectiveness of immigration policies". In: *Population and Development Review* 39.3 (2013), pp. 487–508. DOI: 10.1111/j.1728-4457.2013.00613.x.

[104] Andreas Hadjar and Susanne Backes. "Migration background and subjective well-being a multilevel analysis based on the European social survey". In: *Comparative Sociology* 12.5 (2013), pp. 645–676.

[105] Marc Helbling. "Validating integration and citizenship policy indices". In: *Comparative European Politics* 11.5 (2013), pp. 555–576. DOI: 10.1057/cep.2013.11.

[106] Arnost Komarek and Lenka Komárková. "Clustering for multivariate continuous and discrete longitudinal data". In: *The Annals of Applied Statistics* (2013), pp. 177–200.

[107] 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.

[108] Cécile Proust-Lima, Hélène Amieva, and Hélène Jacqmin-Gadda. "Analysis of multivariate mixed longitudinal data: a flexible latent process approach". In: *British Journal of Mathematical and Statistical Psychology* 66.3 (2013), pp. 470–487.

[109] Karthick S Ramakrishnan. "Incorporation versus Assimilation". In: *Outsiders No More?: Models of Immigrant Political Incorporation* (2013), p. 27. DOI: 10.1093/acprof:oso/9780199311316.003.0002.

[110] Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. "Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates". In: *Test* 23 (2014), pp. 433–465.

[111] Justin Gest et al. "Measuring and comparing immigration, asylum and naturalization policies across countries: challenges and solutions". In: *Global Policy* 5.3 (2014), pp. 261–274. DOI: 10.1111/1758-5899.12132.

[112] Shima Ghassempour, Federico Girosi, and Anthony Maeder. "Clustering Multivariate Time Series Using Hidden Markov Models". In: *International Journal of Environmental Research and Public Health* 11.3 (Mar. 2014), p. 2741. DOI: 10.3390/ijerph110302741.

[113] M. S. Gilthorpe et al. "Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures". In: *Journal of Developmental Origins of Health and Disease* 5.3 (Mar. 2014), p. 197. DOI: 10.1017/S2040174414000130.

[114] Arnošt Komárek and Lenka Komárková. "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.

[115] Daniel Manrique-Vallier. "Longitudinal mixed membership trajectory models for disability survey data". In: *The Annals of Applied Statistics* 8.4 (2014), p. 2268.

[116] Cécile Proust-Lima, Mbéry Séne, Jeremy MG Taylor, and Hélène Jacqmin-Gadda. "Joint latent class models for longitudinal and time-to-event data: a review". In: *Statistical methods in medical research* 23.1 (2014), pp. 74–90. DOI: 10.1177/0962280212445839.

[117] 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.

[118] Liv Bjerre, Marc Helbling, Friederike Römer, and Malisa Zobel. "Conceptualizing and measuring immigration policies: A comparative perspective". In: *International Migration Review* 49.3 (2015), pp. 555–600. DOI: 10.1111/imre.12100.

[119] Sara Wallace Goodman. "Conceptualizing and measuring citizenship and integration policy: Past lessons and new approaches". In: *Comparative Political Studies* 48.14 (2015), pp. 1905–1941. DOI: 10.1177/0010414015592648.

[120] Christian Hennig. "What are the true clusters?" In: *Pattern Recognition Letters* 64 (2015), pp. 53–62.

[121] Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. *Handbook of cluster analysis*. CRC press, 2015. ISBN: 978-0-42918547-2. DOI: 10.1201/b19706.

[122] Adalbert Kerber and Rainer Brüggemann. "Problem Driven Evaluation of Chemical Compounds and Its Exploration". In: *MATCH Commun Math Comput Chem* 73 (2015), pp. 577–618.

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

[124] Didier Ruedin. "Increasing validity by recombining existing indices: MIPEX as a measure of citizenship models". In: *Social Science Quarterly* 96.2 (2015), pp. 629–638.

[125] Michel Beine et al. "Comparing immigration policies: An overview from the IMPALA database". In: *International Migration Review* 50.4 (2016), pp. 827–863. DOI: 10.1111/imre.12169.

[126] 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.

[127] Liesbeth Bruckers, Geert Molenberghs, Pim Drinkenburg, and Helena Geys. "A clustering algorithm for multivariate longitudinal data". In: *Journal of Biopharmaceutical Statistics* (July 2016). URL: https://www.tandfonline.com/doi/full/10.1080/10543406.2015.1052476?src=recsys.

[128] 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.

[129] 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.

[130] Blanca Garcés-Mascareñas and Rinus Penninx. *Integration processes and policies in Europe: Contexts, levels and actors*. Cham: Springer Nature, 2016. DOI: 10.1007/978-3-319-21674-4.

[131] 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.

[132] 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.

[133] 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.

[134] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. "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.

[135] Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni. "LMest: An R package for latent Markov models for longitudinal categorical data". In: *Journal of Statistical Software* 81.1 (2017), pp. 1–38.

[136] Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. "High-Dimensional Mixed Graphical Models". In: *Journal of Computational and Graphical Statistics* (Apr. 2017). DOI: 10.1080/10618600.2016.1237362.

[137] Marco Fattore. "Synthesis of Indicators: The Non-aggregative Approach". In: *Complexity in Society: From Indicators Construction to their Synthesis*. Ed. by F. Maggino. Cham: Springer, 2017, pp. 193–212.

[138] Marc Helbling, Liv Bjerre, Friederike Römer, and Malisa Zobel. "Measuring immigration policies: The IMPIC database". In: *European Political Science* 16 (2017), pp. 79–98. DOI: 10.1057/eps.2016.4.

[139] Adalbert Kerber. "Evaluation, Considered as Problem Orientable Mathematics over Lattices". In: *Partial Order Concepts in Applied Sciences*. Ed. by M. Fattore and R. Brüggemann. Dordrecht: Springer, 2017, pp. 87–103.

[140] Filomena Maggino. "Developing Indicators and Managing the Complexity". In: *Complexity in Society: From Indicators Construction to their Synthesis*. Ed. by F. Maggino. Cham: Springer, 2017, pp. 87–114.

[141] Matthieu Marbac, Christophe Biernacki, and Vincent Vandewalle. "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics-Theory and Methods* 46.23 (2017).

[142] Bengt Muthen and Linda Muthen. "Mplus User's Guide. Eighth Edition". In: (2017).

[143] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm". In: *Journal of Statistical Software* 78 (June 2017), pp. 1–56. ISSN: 1548-7660. DOI: 10.18637/jss.v078.i02.

[144] Glenn Rayp, Ilse Ruyssen, and Samuel Standaert. "Measuring and explaining cross-country immigration policies". In: *World Development* 95 (2017), pp. 141–163. DOI: 10.1016/j.worlddev.2017.02.026.

[145] Nicole B Simpson. "Demographic and economic determinants of migration". In: *IZA World of Labor* (2017).

[146] Xinyuan Song, Yemao Xia, and Hongtu Zhu. "Hidden Markov latent variable models with multivariate longitudinal data". In: *Biometrics* 73.1 (Mar. 2017), pp. 313–323. ISSN: 1541-0420. DOI: 10.1111/biom.12536. eprint: 27148857.

[147] Simon N. Wood. *Generalized Additive Models:An Introduction with R, Second Edition*. Andover, England, UK: Taylor & Francis, May 2017. ISBN: 978-1-31537027-9. DOI: 10.1201/9781315370279.

[148] Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press, 2017.

[149] 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.

[150] 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.

[151] Niklas Harder et al. "Multidimensional measure of immigrant integration". In: *Proceedings of the National Academy of Sciences* 115.45 (2018), pp. 11483–11488. DOI: 10.1073/pnas.1808793115.

[152] 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.

[153] 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.

[154] 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.

[155] Amir Ahmad and Shehroz S. Khan. "Survey of State-of-the-Art Mixed Data Clustering Algorithms". In: *IEEE Access* 7 (Mar. 2019), pp. 31883–31902. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2903568.

[156] Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data*. Chapman and Hall/CRC, 2019.

[157] Charles Bouveyron, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108644181.

[158] Charles Bouveyron, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery. *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.

[159] Zvi Gilula, Robert E. McCulloch, Yaacov Ritov, and Oleg Urminsky. "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.

[160] Sara Wallace Goodman. "Indexing immigration and integration policy: Lessons from Europe". In: *Policy Studies Journal* 47.3 (2019), pp. 572–604. DOI: 10.1111/psj.12283.

[161] Peter J Green. "Introduction to finite mixtures". In: *Handbook of Mixture Analysis*. Chapman and Hall/CRC, 2019, pp. 3–20. DOI: 10.1201/9780429055911.

[162] Marc Helbling and David Leblang. "Controlling immigration? How regulations affect migration flows". In: *European Journal of Political Research* 58.1 (2019), pp. 248–269. DOI: 10.1111/1475-6765.12279.

[163] David Ingleby, Roumyana Petrova-Benedict, Thomas Huddleston, and Elena Sanchez. "The MIPEX health strand: a longitudinal, mixed-methods survey of policies on migrant health in 38 countries". In: *European journal of public health* 29.3 (2019), pp. 458–462.

[164] 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.

[165] Margot Selosse, Julien Jacques, Christophe Biernacki, and Florence Cousson-Gélie. "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.

[166] Leonardo Salvatore Alaimo. "Complexity of social phenomena: Measurements, analysis, representations and synthesis". In: *Unpublished doctoral dissertation, University of Rome "La Sapienza", Rome, Italy* (2020).

[167] Leonardo Salvatore Alaimo and Filomena Maggino. "Sustainable Development Goals Indicators at Territorial Level: Conceptual and Methodological Issues — The Italian Perspective". In: *Social Indicators Research* 147.2 (2020), pp. 383–419. DOI: 10.1007/s11205-019-02162-4.

[168] 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.

[169] Agnès François-Lecompte, Morgane Innocent, Dominique Kréziak, and Isabelle Prim-Allaz. "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.

[170] Shuchismita Sarkar, Xuwen Zhu, Volodymyr Melnykov, and Salvatore Ingrassia. "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.

[171] Shuchismita Sarkar, Xuwen Zhu, Volodymyr Melnykov, and Salvatore Ingrassia. "On parsimonious models for modeling matrix data". In: *Computational Statistics & Data Analysis* 142 (2020), p. 106822. DOI: 10.1016/j.csda.2019.106822.

[172] Margot Selosse. "Introducing parsimony to analyse complex data with model-based clustering". Theses. Université de Lyon, Nov. 2020. URL: https://theses.hal.science/tel-04592164.

[173] Margot Selosse, Julien Jacques, and Christophe Biernacki. "Model-based co-clustering for mixed type data". In: *Computational Statistics & Data Analysis* 144 (Apr. 2020). ISSN: 0167-9473. DOI: 10.1016/j.csda.2019.106866.

[174] Giacomo Solano and Thomas Huddleston. "Migrant Integration Policy Index 2020". In: *Barcelona Center for International Affairs (CIDOB)* (2020).

[175] 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.

[176] 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.

[177] Klaas Wardenaar. "Latent Class Growth Analysis and Growth Mixture Modeling using R: A tutorial for two R-packages and a comparison with Mplus." In: *OSF* (Apr. 2020). DOI: 10.31234/osf.io/m58wx.

[178] Leonardo Salvatore Alaimo. "Complex Systems and Complex Adaptive Systems". In: *Encyclopedia of Quality of Life and Well-being Research*. Ed. by F. Maggino. Cham: Springer, 2021, pp. 1–3. DOI: 10.1007/978-3-319-69909-7_104659-1.

[179] Leonardo Salvatore Alaimo. "Complexity and knowledge". In: *Encyclopedia of Quality of Life and Well-being Research*. Ed. by F. Maggino. Cham: Springer, 2021, pp. 1–2. DOI: 10.1007/978-3-319-69909-7_104658-1.

[180] Leonardo Salvatore Alaimo, Alberto Arcagni, Marco Fattore, and Filomena Maggino. "Synthesis of multi-indicator system over time: A poset-based approach". In: *Social Indicators Research* 157.1 (2021), pp. 77–99.

[181] Leonardo Salvatore Alaimo and Emiliano Seri. "Monitoring the main aspects of social and economic life using composite indicators: A literature review". In: *Working papers Research group Economics, Policy Analysis, and Language; Ulster University* W.P. 21-7 (2021), pp. 1–58.

[182] Filomena Maggino and Leonardo Salvatore Alaimo. "Complexity and wellbeing: measurement and analysis". In: *A Modern Guide to the Economics of Happiness*. Ed. by L. Bruni, A. Smerilli, and D. De Rosa. Cheltenham, UK: Edward Elgar Publishing, 2021, pp. 113–128.

[183] Filomena Maggino, Rainer Bruggemann, and Leonardo Salvatore Alaimo. "Indicators in the framework of partial order". In: *Measuring and Understanding Complex Phenomena*. Ed. by Rainer Bruggemann et al. Cham: Springer, 2021, pp. 17–29.

[184] 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.

[185] Giacomo Solano and Thomas Huddleston. "Beyond immigration: Moving from Western to global indexes of migration policy". In: *Global Policy* 12.3 (2021), pp. 327–337. DOI: 10.1111/1758-5899.12930.

[186] Jan Vávra and Arnošt Komárek. "Clustering based on multivariate mixed type longitudinal data with an application to the EU-SILC database". In: *22nd European Young Statisticians Meeting*. 2021, p. 148.

[187] Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. "MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling". In: *Journal of Classification* (2021), pp. 1–24.

[188] Leonardo S. Alaimo et al. "Measuring Equitable and Sustainable Well-being in Italian Regions. The Non-aggregative Approach". In: *Social Indicators Research* 161 (2022). https://doi.org/10.1007/s11205-020-02388-7, pp. 711–733.

[189] Alberto Alesina and Marco Tabellini. *The Political Effects of Immigration: Culture or Economics?* Tech. rep. National Bureau of Economic Research, 2022. DOI: 10.3386/w30079.

[190] Zihang Lu and Wendy Lou. "Bayesian consensus clustering for multivariate longitudinal data". In: *Statistics in Medicine* 41.1 (Jan. 2022), pp. 108–127. ISSN: 1097-0258. DOI: 10.1002/sim.9225.

[191] Filomena Maggino and Leonardo Salvatore Alaimo. "Measuring Complex Socio-economic Phenomena. Conceptual and Methodological Issues". In: *Interdisciplinary Approaches to Climate Change for Sustainable Growth*. Ed. by S. Valaguzza and M. A. Hughes. Cham: Springer, 2022, pp. 43–59.

[192] Zhiwen Tan, Chang Shen, and Zihang Lu. *BCClong package: Bayesian consensus clustering model for mixed-type longitudinal data*. R package version 1.0. 2022.

[193] Salvatore D Tomarchio, Salvatore Ingrassia, and Volodymyr Melnykov. "Modelling students' career indicators via mixtures of parsimonious matrix-normal distributions". In: *Australian & New Zealand Journal of Statistics* (2022). DOI: 10.1111/anzs.12351.

[194] 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.

[195] Leonardo 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.

[196] 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.

[197] Zihang Lu, Mojtaba Ahmadiankalati, and Zhiwen Tan. "Joint clustering multiple longitudinal features: A comparison of methods and software packages with practical guidance". In: *Statistics in Medicine* 42.29 (Dec. 2023), pp. 5513–5540. ISSN: 0277-6715. DOI: 10.1002/sim.9917.

[198] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[199] Rapty Sarker et al. "The WHO declares COVID-19 is no longer a public health emergency of international concern: benefits, challenges, and necessary precautions to come back to normal life". In: *International Journal of Surgery* 109.9 (May 2023), p. 2851. DOI: 10.1097/JS9.0000000000000513.

[200] Anjali Silva et al. "Finite Mixtures of Matrix Variate Poisson-Log Normal Distributions for Three-Way Count Data". In: *Bioinformatics* (Apr. 2023), btad167. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad167.

[201] Jan Vávra and Arnošt Komárek. "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.

[202] Junyi Zhou, Ying Zhang, and Wanzhu Tu. "clusterMLD: An Efficient Hierarchical Clustering Method for Multivariate Longitudinal Data". In: *Journal of Computational and Graphical Statistics* (July 2023). ISSN: 1061-8600. DOI: 10.1080/10618600.2022.2149540.

[203] Francesco Amato and Julien Jacques. "MMM: Clustering Multivariate Longitudinal Mixed-type Data". working paper or preprint. Nov. 2024. URL: https://hal.science/hal-04807626.

[204] 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.

[205] Andrea Cappozzo, Alessandro Casa, and Michael Fop. "Sparse Model-Based Clustering of Three-Way Data via Lasso-Type Penalties". In: *Journal of Computational and Graphical Statistics* (Dec. 2024). ISSN: 1061-8600. DOI: 10.1080/10618600.2024.2429705.

[206] Sjoerd Hermes, Joost van Heerwaarden, and Pariya Behrouzi. "Copula Graphical Models for Heterogeneous Mixed Data". In: *Journal of Computational and Graphical Statistics* (Jan. 2024). ISSN: 1061-8600. DOI: 10.1080/10618600.2023.2289545.

[207] Francis K. C. Hui, Khue-Dung Dang, and Luca Maestrini. "Simultaneous Coefficient Clustering and Sparsity for Multivariate Mixed Models". In: *Journal of Computational and Graphical Statistics* (Oct. 2024). ISSN: 1061-8600. DOI: 10.1080/10618600.2024.2402904.

[208] William Ruth. *A review of Monte Carlo-based versions of the EM algorithm*. 2024. arXiv: 2401.00945 [stat.CO]. URL: https://arxiv.org/abs/2401.00945.

[209] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.32.6. 2024. URL: https://mc-stan.org/.

[210] Jan Vávra, Arnošt Komárek, Bettina Grün, and Gertraud Malsiner-Walli. "Clusterwise multivariate regression of mixed-type panel data". In: *Statistics and Computing* 34.1 (Feb. 2024), pp. 1–20. ISSN: 1573-1375. DOI: 10.1007/s11222-023-10304-5.

[211] Geert Molenberghs and Geert Verbeke. *Linear Mixed Models for Longitudinal Data*. New York, NY, USA: Springer. ISBN: 978-1-4419-0300-6.

# Appendix A

# Supplement: Mixed clustering

## A.1 Weight of each data type

This experiment aims at studying the model behaviour when all but one of the data types are hold equal among clusters. The results are presented in Figure A.1 and the parameters used are in Table A.1. Not being particularly interested in the influence of size or noise, we performed the experiment just for $N = 500$ and $\tau = 0.1$.

As we can see, when the parameters reported in Table A.1 are kept fixed, then count data appear to have a major weight, while continuous and categorical data have a fair proportionate weight. Binary data-type appears to be the one with less weight, almost none.

However, this result may be influenced by the fact that the parameters in Table A.1 do not have equal distances, meaning that the mean parameters for continuous and ordinal data are closer between the clusters than the ones for binary and count data-type. In order to account for that, we performed again the simulations for these two data-type, but this time readjusting the parameters in order for them to have the same distance as the continuous and ordinal ones. The new parameters are reported in Table A.2.
As we can see in Figure A.1, the weight of count data-type is drastically reduced in this way, even if it still out-weights continuous and categorical data-types, due to the way parameters of the latent continuous variable are linked to the observed count data through the exponential function. On the other hand, binary data-type still seems to have a limited weight. This may be due to the fact that they still have a smaller margin of difference with respect to the threshold, which leads to difficulties into performing proper clustering when the other variables are held still, especially with relatively high positive values as in this case.
Overall, we can say that for similar parameters continuous, ordinal and count data-types have comparable weights in determining the partitioning. The same does not hold for binary data-type, for which a larger distance among clusters is needed for them to be properly detected.

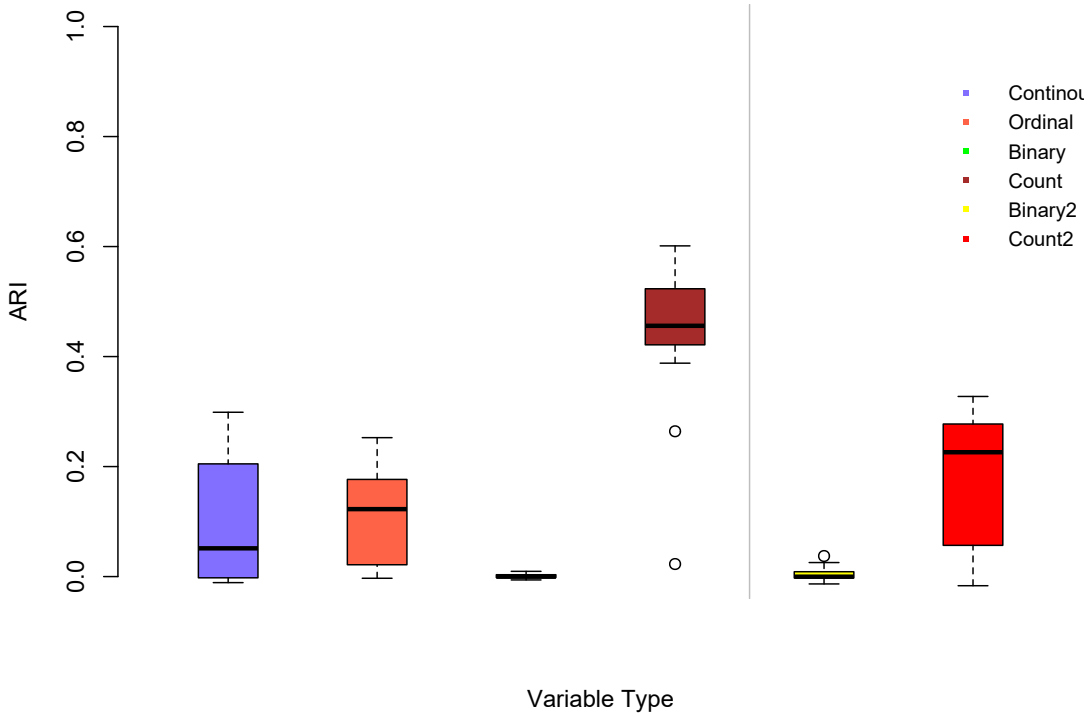



Figure A.1: ARI for different kind of variable which is kept different from the others.

## A.2 Tables

Table A.1: Means matrices for fixed data-type simulation

| **Continous** | | | |
|---|---|---|---|
| **Cluster 1** | T1 | T2 | T3 |
| V1 | 1.75 | 1.75 | 1.75 |
| V2 | 2.255 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 1.75 | 1.75 | 1.75 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.75 | 2.75 | 2.75 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 1.75 | 1.75 | 1.75 |

| **Categorical** | | | |
|---|---|---|---|
| **Cluster 1** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 1.75 | 1.75 | 1.75 |
| V3 | 0 | 0 | 0 |
| V4 | 1.75 | 1.75 | 1.75 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.75 | 2.75 | 2.75 |
| V3 | 0 | 0 | 0 |
| V4 | 1.75 | 1.75 | 1.75 |

| **Binary** | | | |
|---|---|---|---|
| **Cluster 1** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | -0.25 | -0.25 | -0.25 |
| V4 | 1.75 | 1.75 | 1.75 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0.25 | 0.25 | 0.25 |
| V4 | 1.75 | 1.75 | 1.75 |

| **Count** | | | |
|---|---|---|---|
| **Cluster 1** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 1 | 1 | 1 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 2.5 | 2.5 | 2.5 |

Table A.2: New means matrices for fixed data-type simulation

| **Binary2** | | | |
|---|---|---|---|
| **Cluster 1** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | -0.5 | -0.5 | -0.5 |

| | | | |
|---|---|---|---|
| V4 | 1.75 | 1.75 | 1.75 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0.5 | 0.5 | 0.5 |
| V4 | 1.75 | 1.75 | 1.75 |

**Count2**

| **Cluster 1** | T1 | T2 | T3 |
|---|---|---|---|
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 1.75 | 1.75 | 1.75 |
| **Cluster 2** | T1 | T2 | T3 |
| V1 | 2.25 | 2.25 | 2.25 |
| V2 | 2.25 | 2.25 | 2.25 |
| V3 | 0 | 0 | 0 |
| V4 | 2.75 | 2.75 | 2.75 |

# Appendix B

# A comparison of migrant integration policies via Mixture of Matrix-Normals

This chapter has been published in October 2022 in the *Social Indicators Research* journal, volume 165 (Alaimo et al., 2023). We have reproduced the entire article as released. For this reason, some concepts may be repeated, particularly concerning Section B.4 with regard to Chapters 2 and 3. The work in question proves the goodness of the Gaussian matrix-variate mixture model in an application on longitudinal continuous data in a social science context, specifically, a clustering application on the Migrant Integration Policy Index. The model allowed for the identification of clusters of countries with similar patterns of migrant integration policies over time.

**Abstract.**In recent decades, there has been a growing interest in comparative studies about migrant integration, assimilation and the evaluation of policies implemented for these purposes. Over the years, the Migrant Integration Policy Index (MIPEX) has become a reference on these topics. This index measures and evaluates the policies of migrants' integration in 52 countries over time. However, the comparison of very different countries can be difficult and, if not well conducted, can lead to misleading interpretations and evaluations of the results. The aim of this paper is to improve this comparison and facilitate the reading of the considered phenomenon, by applying a Mixture of Matrix-Normals classification model for longitudinal data. Focusing on data for 7 MIPEX dimensions from 2014 to 2019, our analysis identify 5 clusters of countries, facilitating the evaluation and the comparison of the countries within each cluster and between different clusters.

**Keywords.**Mixture of matrix-normals, MIPEX, Model-based classification, Migration policies

## B.1 Introduction

Immigration regulation and immigrants assimilation have been salient political issues in all industrialised countries for many decades, mainly because of their cultural and economic effects (Alesina and Tabellini, 2022). The growing interest in the study of immigration, starting from citizenship and moving more recently to integration, has led to a variety of attempts to quantify immigration policies. Policy indices have become mandatory in the study of immigrant-related policies implemented by different countries. However, the study of these phenomena from a quantitative point of view is rather recent, due to the previous lack or difficulties to access of data (Bjerre, Helbling, Römer, and Zobel, 2015). Moreover, quantifying migrant integration is a difficult challenge, linked to its complex nature and lack of uniformity in migration policies of many countries, which are based on multiple criteria.

In this work, we focus on the Migrant Integration Policy Index (MIPEX) (Niessen et al., 2007, Solano and Huddleston, 2020), a complex system of 167 policy indicators across 8 domains of citizenship and integration, combined into a single composite indicator in order to evaluate the migrant integration policies of each considered country over the years. MIPEX has quickly become a solid and useful tool for evaluating and comparing what governments are doing to promote the migrants' integration in a cross-country setting. Indeed, it informs and engages key policy actors about how to use indicators to improve integration governance and policy effectiveness, with the aim to measure policies that promote integration in both socio-economic and civic terms. Although not without its critics, this index has become a reference for comparative studies on migrant integration over the last decade and its data has been widely used in literature (Hadjar and Backes, 2013, Ruedin, 2015, Rayp, Ruyssen, and Standaert, 2017, Ingleby, Petrova-Benedict, Huddleston, and Sanchez, 2019). This paper aims to deeply look at how similar, or dissimilar, countries really are and to add new reading perspectives on the MIPEX data, by discovering structures and patterns in the behaviour of the considered countries. The underlying idea is that, given the complex and multidimensional nature of the phenomenon and the differences in socio-economic and civic terms between the examined countries, it can be misleading to compare all of the units with each others. Therefore, the present work aims at improving the analysis, by grouping countries in order to facilitate the comparison and interpretation of the phenomenon. Thus, the research question to which we try to answer:

- *In order to improve the comparison between the countries regarding their migrant integration policies, is it possible to identify homogeneous groups over time among them, i.e. groups of countries which behave similarly across and within time?*

To answer this research question, a Finite Mixture of Matrix-Normals model has been applied to cluster the units, taking into account the longitudinal dimension along 6 years, on the 52 available countries for 7 of the 8 dimensional indicators of the MIPEX. We relied on an unsupervised parametric clustering approach to minimize the risk of arbitrariness[1] in the choices made and to be able to better evaluate the results.

The paper is structured as follows. Section B.2 describes the immigrants integration framework and some works related to migration indicators. Section B.3 presents the description of the analysed data and the structure of the MIPEX theoretical framework. In Section B.4 we present

[1]Subjectivity is an essential element in any measurement process, but its presence does not make the process arbitrary (Alaimo, 2020).

the methodology implemented. Section B.5 reports data analysis and the results and Section B.6 concludes.

## B.2 Theoretical framework and related works

### B.2.1 Immigrants integration framework

Immigration can be generally defined as the set of policies that determine who can enter or exit a country under what conditions, as well as how immigrants are considered once they are settled in a country. Many factors contribute to the migratory flows and stocks (forced or voluntary) to destination countries, which have been extensively addressed in the literature (Dustmann and Preston, 2007, Pedersen, Pytlikova, and Smith, 2008, Simpson, 2017). We distinguish short-term migrants (seasonal agricultural workers, students, tourists, or temporary residents) and long-term migrants that include permanent residents, the first step on a path towards the creation of members, namely the citizenship (Goodman, 2019, Solano and Huddleston, 2021). Migration and migrant integration dynamics influence the number and characteristics of migrants entering a country, as well as the integration outcomes (Massey et al., 1998, Czaika and De Haas, 2013, Garcés-Mascareñas and Penninx, 2016, Helbling and Leblang, 2019). At the same time, the receiving society defines all the laws and policies that relate to the selection, admission, integration, settlement, and full membership of migrants in a country (Hammar, 1990, Bjerre, Helbling, Römer, and Zobel, 2015, Solano and Huddleston, 2021). Citizenship, migration, and integration policy, albeit in different ways, are distinct policy domains and creates the conditions that support or hinder migrants' inclusion in the destination society. More attention has been paid to integration policies in recent years, so much so that, in modern countries, they have evolved into very complex legal constructs (Zincone, Pennix, and Borkert, 2011), whereas previously the focus was more on immigrant or assimilation policies. Moreover, as reported in Ramakrishnan, 2013, in several countries terms like *assimilation*, *adaptation*, *incorporation* and *integration*, often refer to the same concept and some efforts were needed to provide more conceptual clarity, especially in finding unambiguous definitions of fundamental concepts on the matter. Castles and Davidson, 2000 highlight that countries have three main policy options with respect to managing social diversity. The first option is *exclusion*. Although this model is not considered legitimate by humanitarian standards and formally not accepted, it should be noted that it is still predominant in large areas of the world. The second option is *assimilation*. According to this policy model, immigrants should be granted full citizenship: the immigrants' distinct culture is seen as in transition and it is expected that they fully adopt the national culture and generally accepted social norms. The third option is *integration*, with respect which policy makers are aware that immigrants do not abandon their distinct culture immediately and, therefore, their cultural identity can be considered an opportunity. Legal integration, intended as an immigrant's legal status, residence rights, citizenship, and equal access to rights, goods, services, and resources, receives wide expert acceptance as the first step in promoting societal integration. It is considered a key determinant (Penninx and Martiniello, 2004) and can hardly be overestimated as either "a firm base" for societal integration or a "clear signal" committing public authorities to an inclusive agenda (Groenendijk, Guild, Dogan, et al., 1998). These differences are strictly linked to the complex nature of immigration policies, which involve different political, social and economical spheres that are interconnected with each other. As explained in Niessen and Huddleston, 2009, integration is developed by policymakers in conjunction with their policies on social inclusion/cohesion, employment, demography, competitiveness. It follows that immigrant

integration is only one part of the broader good governance framework. In recent years, various studies have tried to develop this framework and quantitative indices of immigration policies have been proposed. These indices play a central role in the study of immigrant-related policies, starting with citizenship and moving to immigration and integration (Helbling, 2013, Goodman, 2015, 2019). The next sub-section, although not exhaustively, present some of the most used immigrant-related policy indexes, highlighting how over time they assume greater specificity in relation to integration policies.

### B.2.2 Immigration policies indexes: a literature review

The policy indices reflect the tendency in social sciences to reduce the complexity of socio-economic phenomena, allowing comparisons across countries and times (Skaaning, 2010, Rainer and Marc, 2011). A sample of immigrant-related policy indexes will be presented below, providing information on index content, type, scope, and source. All of the indices reported in this paper make important and innovative contributions to the field of comparative immigration policy research. It is not our goal to discuss whether and which indexes are better than others. Each index has different methodological and conceptual assumptions and answers specific research questions. In the migratory field, the first index was proposed by Waldrauch and Hofinger, 1997 in a study on citizenship, examining the Legal Obstacles to Integration (LOI). But indexing did not stop at citizenship. Several studies have documented the expansion of indexing from citizenship to integration, assuming more specificity for immigration policies (Helbling, 2013, Goodman, 2015, 2019). The first immigrant-related policy indexes proposed, do not differentiate between immigration and integration policy domains. An exception is represented by the index proposed by Boushey and Luedtke, 2011, who first consider the distinction between immigration control and immigrant integration measures. This index provides "conceptual clarification to indexing by distinguishing immigration as control policies [that] deal with keeping out "unwanted immigrants" and integration policy as dictat[ing] the transition and settlement of resident immigrants" (Goodman, 2019, p. 579). Recently, an interdisciplinary community of scholars has developed multi-dimensional indices capable of differentiating across types of policies, target groups, and instruments (Goodman, 2010, Koopmans, Michalowski, and Waibel, 2012, Goodman, 2019). We briefly present some of the main ones:

- First released by Banting, Kymlicka, et al., 2006, the *Multiculturalism Policy Index* (MCP) is a scholarly research project that monitors the evolution of multiculturalism policies in 21 Western democracies. The MCP is designed to provide information about multiculturalism policies in a standardized format that aids comparative research and contributes to the understanding of State-minorities relations. The project provides an index at 3 points in time: 1980, 2000, 2010, and for 3 types of minorities: one index relating to immigrant groups; one relating to historic national minorities; one index relating to indigenous peoples.

- The Migrant Integration Policy Index (MIPEX) (Niessen et al., 2007, Solano and Huddleston, 2020) is a complex system of 167 policy indicators across 8 domains of citizenship and integration combined into a single composite indicator, in order to evaluate the migrant integration policies of each considered country (for details, see Section B.3).

- Based on the selection of data for 9 countries, between 1999 and 2008, and with the aim of measuring and comparing immigration, asylum, and naturalization policies across countries, the *International Migration Policy and Law Analysis* (IMPALA) database collects comparable

data on immigration law and policy across 6 major areas of migration legislation: economic migration, family reunification, humanitarian migration, irregular migration, student migration, and the acquisition and loss of citizenship for migrants resident (Gest et al., 2014, Beine et al., 2016).

- Helbling, Bjerre, Römer, and Zobel, 2017 presented the *Immigration Policies in Comparison* (IMPIC) project, which proposes a data set that allows to measure immigration regulations.
- *The Canadian Index for Measuring Integration* (CIMI), is an interactive tool that allows for measuring the outcomes of immigrants in Canadian regions. It is a data-driven index that examines 4 dimensions of immigrants' integration in Canada to assess the gaps between immigrants and the Canadian-born population. The CIMI identifies factors that underline successful immigrants' integration, assesses changes and trends over time (currently from 1991 to 2020), enables detailed examination of 4 dimensions of integration and provides rankings based on empirical evidence for Canadian geographies.
- The *Immigration Policy Lab* (IPL) (Harder et al., 2018) is a survey-based measure of immigrant integration, to provide scholars with a short instrument that can be implemented across survey modes, with the aim to strike a pragmatic compromise to help generate cumulative knowledge on immigrant integration. The IPL captures 6 dimensions of integration: psychological, economical, political, social, linguistical, and navigational.

With the proliferation of such policy indices, scholars have more refined tools than ever for classifying and comparing policy plans and practices. Immigration and integration policies vary across dimensions, and limiting them to a single dimension reduces the ability to observe variations that could be significant. For this reason, we focused our analysis on MIPEX dimensions instead of the final composite indicator.

## B.3 Data

Analyzing a complex phenomenon (Alaimo, 2021b) is often connected to the measuring of some non-directly measurable latent variables (Maggino and Alaimo, 2021, Maggino, Bruggemann, and Alaimo, 2021, Maggino and Alaimo, 2022). The measurement process in social sciences is associated with the construction of system of indicators. The indicators within a system are interconnected and new properties typical of the system emerge from these interconnections. As it can be easily understood, these kinds of systems are complex systems (Alaimo, 2021a). Therefore, a system of indicators allows the measurement of a complex concept that would not otherwise be measurable by taking into account the indicators individually (Alaimo and Maggino, 2020).

The MIPEX is a system of 167 policy indicators[2] and it includes 52 countries and collects data from 2007 to 2019, in order to provide a view of integration policies across a broad range of differing environments. The values of each indicator are chosen by experts from each country, by means of a questionnaire. The MIPEX synthetic indicator is constructed by means of an aggregative-compensative approach (Nardo, Saisana, Saltelli, and Tarantola, 2005, OECD, 2008, Alaimo and Maggino, 2020). The 167 basic indicators are first aggregated in 58 indicators (for more information, please consult Solano and Huddleston, 2020), which cover the 8 policy areas designed to benchmark

[2] A policy indicator is a question relating to a specific policy component of one of the 8 policy areas.

current laws and policies against the highest standards through consultations with top scholars and institutions,[3]. The policy areas of integration covered are the following::

- Labour Market Mobility (X1)
- Family Reunion (X2)
- Education (X3)
- Political Participation (X4)
- Long-term Residence (X5)
- Access to Nationality (X6)
- Anti-discrimination (X7)
- Health[4]

For each area, a synthetic measure (dimensional) is calculated as the arithmetic mean of the elementary indicators[5], i.e. those selected for measuring each policy area. Each dimensional synthetic indicator is bounded between $[0, 100]$: the higher the value, the better the situation in that policy area.

The method and the approach adopted for the construction of the synthetic index have not been without criticism. Even if it is the most widespread among the aggregation methods for composite indicators construction, the arithmetic mean it has been highly criticized. The main advantage of this method is that it is simple, largely known and gives easy-to-understand results. The main drawback is that it is a full compensative method; consequently, low values in some indicators can be compensated by high values in other ones (OECD, 2008). This assumption is very strong and has a great impact on the results obtained, leading in many cases to an extreme flattening of the differences between the units (Alaimo and Seri, 2021). Despite its success, the aggregative-compensative approach has been deeply criticized as inappropriate and often inconsistent, from both conceptual and methodological point of view (Freudenber, 2003, Fattore, 2017, Maggino, 2017). To address and try to overcome the limitations of this approach, in recent years alternative procedures to synthesis have been developed in the literature (for instance, see: Kerber and Brüggemann, 2015, Kerber, 2017, Alaimo, Arcagni, Fattore, and Maggino, 2021, Alaimo et al., 2022). However, the purpose of this paper is to improve the analysis of the dimensions of MIPEX in its present form, albeit we suggest a critical read of it. The analysis carried out in the present work uses the listed above dimensions (excluding health), of which we are going to give a brief description in the following sub-sections[6].

### B.3.1 Labour Market Mobility

Integration of immigrants into the labor market is a process that happens over time and depends on general policies, context, immigrants' skills and the reason for migration. Labour market mobility policies qualify as only halfway favourable for promoting equal quality employment over the long-term. In most countries, family members and permanent residents can access the labour market and job training, as well as social security and assistance. However, according to Solano and Huddleston, 2020, full equality of rights and opportunity in the labour market is still far from being achieved, especially in the public sector.

[3]The highest standards are drawn from Council of Europe Conventions, European Union Directives and international conventions (for more information see: http://mipex.eu/methodology).
[4]This dimensions was excluded from the analysis, because it presents data only available for years 2014 and 2019.
[5]The elementary indicators are described in (Solano and Huddleston, 2020).
[6]A more extensive explanation is given in (Solano and Huddleston, 2020).

### B.3.2 Family Reunion

Family reunification policies determine if and when separated families can reunite and settle in their new home. According to Solano and Huddleston, 2020, policies are more favourable in traditional destination countries, Northern European countries and new countries of labour migration (e.g. Italy, Portugal and Spain). On the other hand, for family reunification some countries require a high fee to pay and little support (e.g. Austria, Denmark, France, Germany, the Netherlands, Switzerland, UK). Increasingly, countries make exceptions for the highly-skilled and the wealthy, but rarely for the most vulnerable (minors and beneficiaries of international protection).

### B.3.3 Education

Despite being an increasing priority for integration, education is the greatest weakness in the integration policies of many countries. Most immigrant pupils receive little support in finding the right school or class, or in 'catching up' with their peers. As described in Solano and Huddleston, 2020, Australia, Canada and New Zealand have developed strong targeted education policies through multiculturalism, while the US focuses additional support on vulnerable racial and social groups. In contrast, the education systems of Austria, France, Germany and Luxembourg are less responsive to the needs of their relatively large number of immigrant pupils. New destination countries with small immigrant communities offer inconsistent targeted support (e.g. Japan and Central Europe).

### B.3.4 Political Participation

In most countries, foreign citizens are not enfranchised or regularly informed, consulted or involved in local civil society and public life. Political participation is one of the weakest areas of integration (Solano and Huddleston, 2020). Foreign citizens' political opportunities differ enormously from one country to another. For instance, in Australia, New Zealand and Western Europe, they enjoy greater voting rights, stronger consultative bodies, more funding for immigrant organisations and greater support from mainstream organisations. With the exception of Korea, immigrants in Asian countries enjoy almost none of these rights unless they (can) naturalise. Despite European norms and promising regional practices, political participation is still almost absent from integration strategies in Bulgaria, Lithuania, Romania and Slovakia.

### B.3.5 Long-term Residence

The security of permanent residence may be a fundamental step on the path to full citizenship and better integration outcomes. Permanent residence is a normal part of the integration process in top-scoring countries in the MIPEX composite indicator, such as Canada, most Latin American countries (Brazil, Chile and Mexico), Nordic countries (Finland and Sweden), and few other European countries (Hungary, Iceland, Slovenia, Ukraine). In contrast, many newcomers are ineligible for permanent residence in China, Denmark, Ireland, Israel, Japan, Switzerland and Turkey. Countries rarely reform their legal routes to permanent residence. The limited major reforms of recent years have been driven by the politicisation of immigration. Brazil, Estonia, Macedonia, Russia, and Turkey have removed previous restrictions, while Austria, Denmark, Korea, Norway, Poland, Ukraine and the US have imposed new ones.

### B.3.6 Access to Nationality

Facilitating access to nationality can significantly increase naturalisation rates and boost integration outcomes. Nationality policies are a major area of weakness in most European and non-European countries (Solano and Huddleston, 2020), especially Austria, Bulgaria, the Baltics, Eastern Europe, and India. By contrast, immigrants have favourable opportunities to become citizens in many countries, e.g., Sweden and the traditional destination countries (Canada, New Zealand and US). Since 2014, nationality policies have become more restrictive in Argentina, Denmark, Greece and Italy, while immigrants' access to nationality has improved significantly in Brazil and Luxembourg and, to lesser extent, in China, Greece, Latvia, Moldova, Portugal, Spain, Switzerland and Turkey.

### B.3.7 Anti-discrimination

Anti-discrimination laws are becoming increasingly widespread. Victims of discrimination are often too poorly informed or supported to take the first step in the long path to justice, so most do not report their experience to the authorities. Victims are best informed and supported to seek justice in traditional destination countries (Canada, New Zealand and the US) and some EU Member States (Finland, Portugal and Sweden). Since the adoption of EU law in 2000, anti- discrimination has been the greatest and most consistent area of improvement in integration policy across Europe. Over the past 5 years, 7 countries have made positive reforms to discrimination policy (Croatia, Finland, Iceland, Ireland, Luxemburg, Slovenia and Turkey) and more than half of the MIPEX countries now protect against ethnic, racial, religious and nationality discrimination in all areas of public life (Solano and Huddleston, 2020). China, India, Japan, Russia and Switzerland are critically behind schedule on these international trends.

## B.4 Methodology

The basic finite mixture model assumes that data are drawn from a density modelled as a convex combination of components each of specified parametric form (Green, 2019). The usage of finite mixture models as clustering procedures comes clear when supposing that the population from which we are sampling is heterogeneous and so there are multiple groups. Model-based clustering refers to the use of statistical models to cluster data, where the (multivariate) observations are assumed to have been generated from a finite mixture of component distributions, each regarded as a cluster, whose specific probability distribution has generated the units belonging to it (Titterington, Afm, Smith, Makov, et al., 1985, Hennig, Meila, Murtagh, and Rocci, 2015). Model-based clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows the analysis 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 (Bouveyron, Celeux, Murphy, and Raftery, 2019a). In our case, we deal with longitudinal data; model-based clustering of such data is far from simple. Indeed, longitudinal data, sometimes referred to as panel data, track the same sample taking measurements at different time occasions. They are very different from time series: in the longitudinal case we observe short sequences of data in correspondence to a large number of individuals or statistical units, whereas in the time series case we observe long sequences of data referred to one or few statistical units (Bartolucci, Farcomeni, and Pennoni, 2019). The ideal way to model these data would be to take into account the temporal evolution and models all the responses at the same

time. Thus, the analysis will exhibit typical temporal evolution behaviours, which are the objects that researchers in human and social sciences wish to study.

In this paper, we adopt a clustering approach to longitudinal data that consists of 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 (Viroli, 2011a) and within a Bayesian one (Viroli, 2011b). 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 MMN is provided in Wang and Melnykov, 2020). Very recently, Tomarchio, Ingrassia, and Melnykov, 2022 applied MMN to cluster longitudinal students' career indicators for Italian universities.

### B.4.1 Mixture of Matrix-Normals

MMN, as introduced in Viroli, 2011a, can be a useful tool to cluster time-dependent data. Suppose we observe $N$ independent and identically distributed random matrices $Y_1, \ldots, Y_N$ of dimension $J \times T$, with $J$-variate vector observations measured repeatedly over $T$ time points (i.e. $Y \in \mathbb{R}^{J \times T}$), as in a longitudinal study case. Assume that $Y$ follows a matrix-normal distribution, $Y \sim \mathcal{MN}_{(J \times T)}(M, \Phi, \Omega)$, 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 $\Omega \in \mathbb{R}^{J \times J}$ is the covariance matrix containing the variance and covariances of the $J$ variables. The matrix-normal probability density function (pdf) is:

$$f(Y \mid M, \Phi, \Omega) = \\ = (2\pi)^{-\frac{TJ}{2}} \mid \Phi \mid^{-\frac{J}{2}} \mid \Omega \mid^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}[\Omega^{-1}(Y-M)\Phi^{-1}(Y-M)^{\intercal}]\right\} \quad \text{(B.4.1)}$$

Being a particular specification 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 and Nagar, 2000). The pdf of the MMN model is:

$$f(Y \mid \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi^{(J \times T)}(Y \mid M_k, \Phi_k, \Omega_k) \quad \text{(B.4.2)}$$

where $K$ 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 $\Theta_k = \{M_k, \Phi_k, \Omega_k\}$.

Matrix-variate models suffer from over-parametrization that leads to estimation issues. This issue is addressed in Sarkar, Zhu, Melnykov, and Ingrassia, 2020b and Zhu, Sarkar, and Melnykov, 2021, with the aim to explain the data with as few parameters as possible. To do so, the spectral decomposition of the covariance matrix (Banfield and Raftery, 1993, Celeux and Govaert, 1995) is used. The spectral decomposition of the general covariance matrix $\Omega_k$ is given by $\Omega_k = \lambda_k \Gamma_k \Delta_k \Gamma_k^{\intercal}$, where $\lambda_k = \mid \Omega_k \mid^{1/J}$, $\Gamma_k$ is the matrix consisting of the eigenvectors of $\Omega_k$ and $\Delta_k$ is the diagonal

matrix composed by the eigenvalues. From a geometrical interpretation point of view, $\lambda_k$ mirrors the volume of the $k$-th mixture component, $\Gamma_k$ the orientation and $\Delta_k$ the shape. In MMN, there are two covariance matrices, one measuring covariance in time and one among variables. For identifiability issues of the model, the determinant of the time-covariance matrix must be restricted to be $\mid \Phi_k \mid = 1$, hence imposing $K$ restrictions and making $\lambda_k = 1$ for the matrix $\Phi_k$. Moreover, two kinds of mean matrices $M$ are considered: a general (no constraints) and an additive one. An additive matrix $M_k$ has the structure $M_k = \alpha_k \mathbf{1}_T^\intercal + \mathbf{1}_J \beta_k^\intercal$, where $\mathbf{1}_T$ represents a $T$-dimensional vector of 1s, $\alpha_k$ is the $J$-dimensional mean vector for the variables (row-wise) and $\beta_k$ is the $T$-dimensional mean vector across time (column-wise). This structure gives rise to identifiability issues, which are resolved by imposing $K$ constraints $\beta_{k,T} = 0$. Last, as introduced in McNicholas and Murphy, 2010, the time-covariance matrix can be further decomposed through the modified Cholesky decomposition to parameters interpretable in an Auto-Regressive (AR) fashion. Any or all among volume, shape or orientation can be constrained across mixture components. Following the conventional notation in Bouveyron, Celeux, Murphy, and Raftery, 2019a, for the covariance matrices parameterizations E stands for equal, V denotes variable, I represents identity, configuring different types of constraints that can be imposed. Since $\Omega_k$ can be decomposed in 3 submatrices, and $\Phi_k$ in 2, we have 14 different possible combination for the former and 8 (including AR) for the latter, giving rise to $14 \times 8 = 112$ different parametrizations. Since the mean matrix $M_k$ can be in turn parametrized with a general or an additive structure, in total we can fit $2 \times 112 = 224$ differently parametrized models.

## B.5 Analysis and results

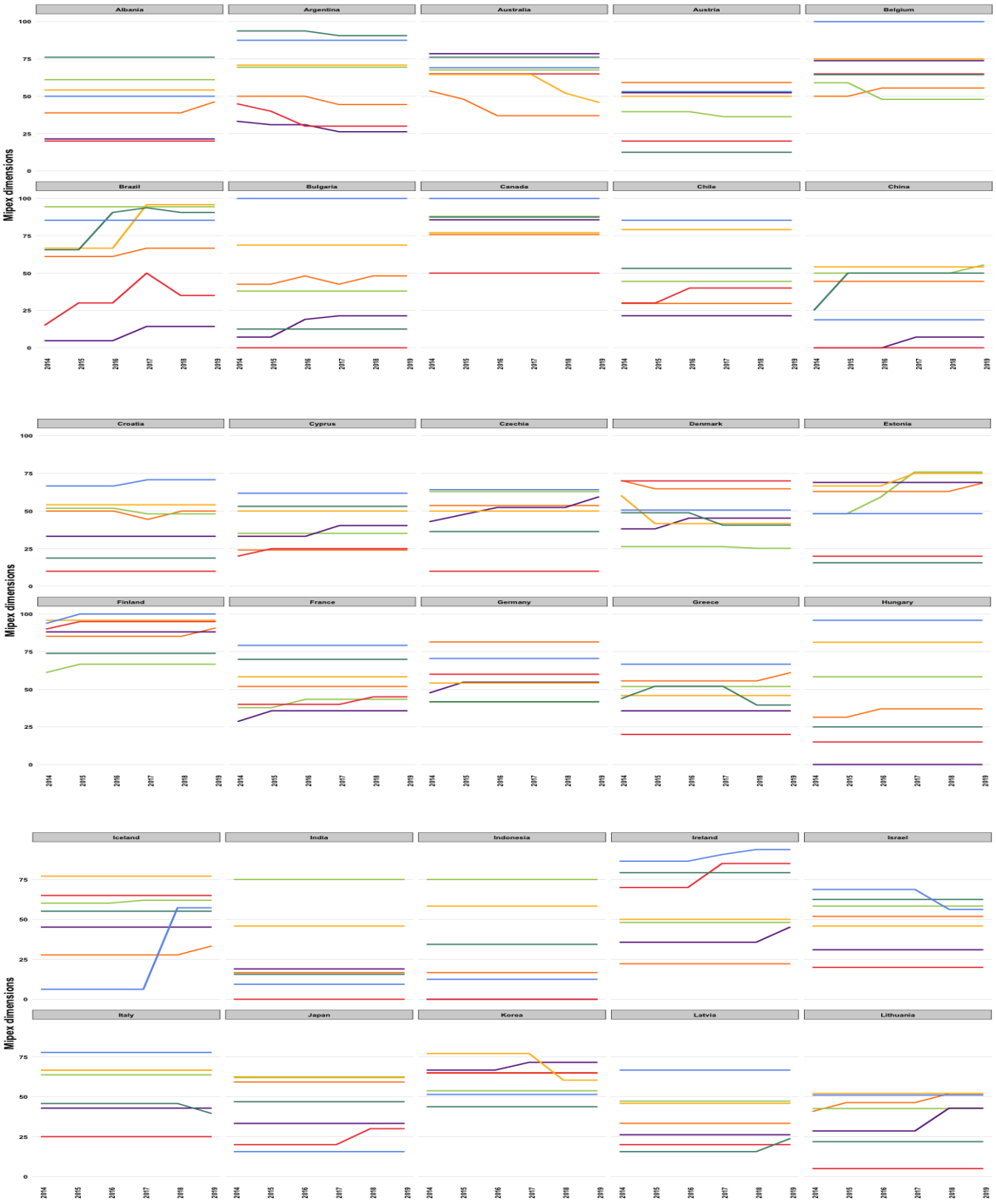
Data used are freely downloadable from the Migrant Integration Policy Index website[7]. For sake of brevity, during the analysis and in all the Tables and Figures, we name the indicators using one-word labels or the codes reported in Section B.3.

The analysis has been carried out by considering 7 MIPEX dimensions explained in Section B.3. In this paper, we deal with a three-way "time data array" of the type "units × variables × times" (D'Urso, 2000) that can be algebraically formalised as follows:

$$\mathbf{Y} \equiv \left\{ y_{ijt} : i = 1, \ldots, N; j = 1, \ldots, J; t = 1, \ldots, T \right\} \tag{B.5.1}$$

where the indices $i$, $j$ and $t$ stand, respectively, for the units, the quantitative variables and the times. In this paper, $i = 1, 2, \ldots, 52$ indicates the generic country, $j = 1, 2, \ldots, 7$ the generic MIPEX dimensional indicator and $t = 204, 2015, \ldots, 2019$ the generic year; consequently, $y_{ijt}$ represents the determination of the $j$-th indicator in the $i$-th country at the $t$-th year. The first step is to give a geometrical representation of the initial data array $\mathbf{Y}$ to obtain information on the form of the data and the relationships between the basic indicators (Pearson, 1956). Figure B.1 outlines that the trajectories of most of the indicators appears quite flat, which means that most of the countries does not change much the values of their indicators (and so the related policies) over time. For instance, Canada, India, Indonesia, Mexico and Romania have no improvement or worsening in any indicator during the considered period; while other countries (for instance, Albania, Austria, Hungary, Italy and Latvia) have just a small change in only one of the considered years. We can also observe that in most of the countries (for instance, Belgium, Bulgaria, Canada, and so on) the labour dimension is the one that rank higher; at the same time, the residence dimension rank lower.

[7] https://www.mipex.eu/download-pdf.


Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Korea
Latvia
Lithuania
Mipex dimensions
0
25
50
75
100
2014
2015
2016
2017
2018
2019

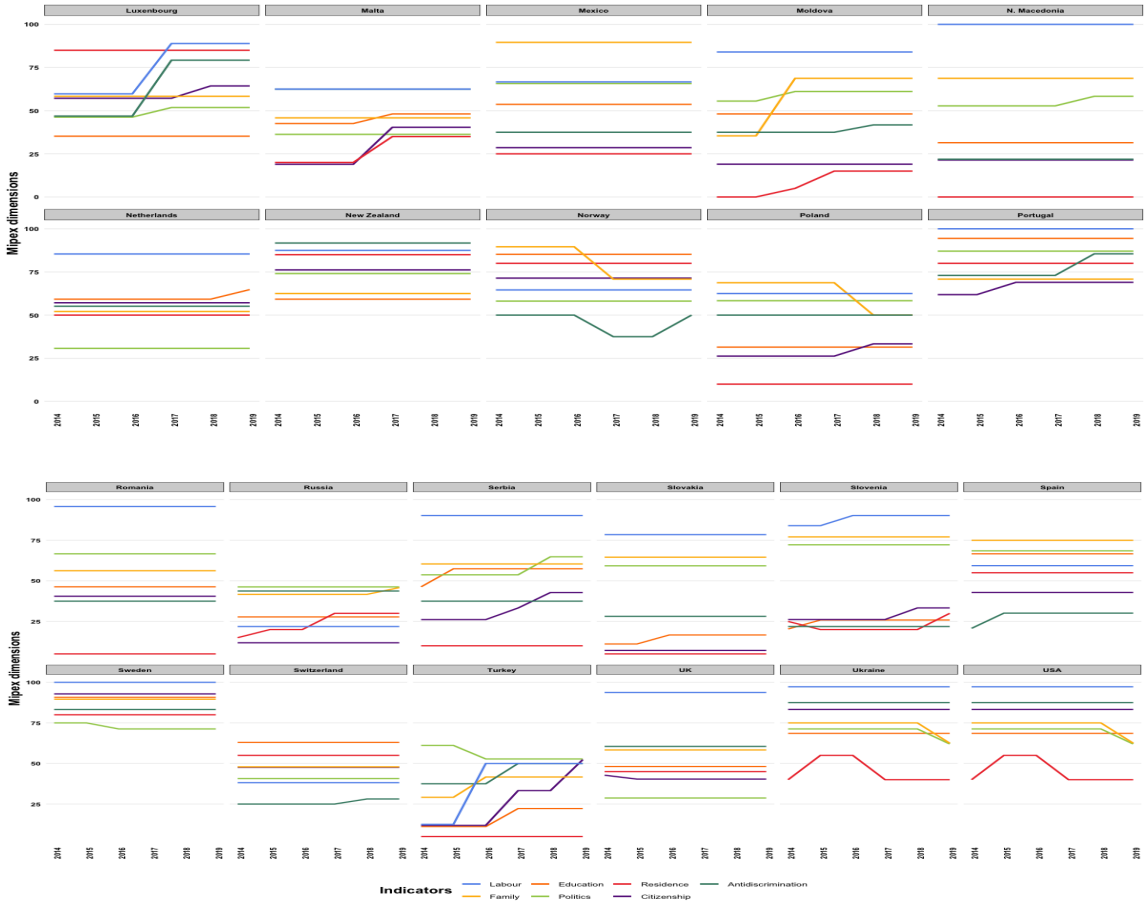



Figure B.1: Country trajectories of the 7 MIPEX dimensions. 52 countries; years 2014 – 2019.

However, this is not true for most of the Asian countries, where the family and politics dimensions tend to rank higher and the labour dimension lower. The MMN will be used to model together the changes between and within time, grouping together the units which behave similarly across and within time.

The cluster analysis have been performed with the package MatTransMix (Zhu, Sarkar, and Melnykov, 2021) of the statistical software R. As usual when performing clustering, the main parameter to set is represented by the number of clusters $K$. Moreover, it is important that the clusters are interpretable (Forgy, 1965, Fraley and Raftery, 1998). Since our dataset is composed by 52 units, we carried out the MMN model for $K$ ranging from 1 to 8 and we run the model several times in order to choose the best number of clusters by means of the Bayesian Information Criterion (BIC): the lowest the BIC, the better the model. The selected number of $K$ is 5. The best parametrization of the model, as expressed in Section B.4.1, is A-VEV-VV[8], which means that the means $M_k$ are better parsimoniously parametrized in additive way, $\Omega_k$ with varying volume, equal shape and varying orientation (in a two components case, it would be ellipsoidal with equal shape) and $\Phi_k$ has both varying shape and orientation.

Because of the matrices $\Phi_k$ and $\Omega_k$, each MMN component models not only the conditional means, but also covariances of the response variables and the covariances among times. This, of course, is visible in the clustering as well, since MMN tends to cluster together not only the units with similar response conditional means, but with conditional covariances among times and variables as well. In this way, each cluster provides a broad profile of units belonging to it. It should be notice that a low correlation in time within cluster means that there have been changes in migration polices in the countries belonging to the cluster; on the other hand, a high correlation in time would signal that little changed. Equally, purified from temporal effect, positive variables correlations mean that the policies' dimensional scores move homogeneously country-wise within cluster. The values of the correlation in time are reported in Figure B.2, the values of the correlations among variables in Figure B.3 and the countries that belongs to each cluster in Figure B.4. The values of the clusters' means over time are reported in Table 1.

A description and interpretation of the clustering results is as follow:

- **Cluster 1**: Estonia and Slovenia.
  - **Correlation in time**: with respect to the other clusters, Cluster 1 is the one with the lowest correlations within time.
  - **Means**: this is the cluster with the lowest mean values in the Citizenship strand. With respect to the other clusters, it has low values in the Politics indicator but high values for Family, Residence and Anti-discrimination.
  - **Correlation among indicators**: the Labour indicator presents negative correlations with almost all the other indicators except for Family. The correlation is particularly high between the indicators Labour and Anti-discrimination.

  In Cluster 1, we observe relatively low levels of temporal correlation, and this is due to the fact that Estonia has important changes in Family indicator in 2016 and 2017 and Residence in 2017, while Slovenia has important changes in the Anti-discrimination in 2016, in Education

[8] The total number of estimated parameters is given by $K+(J-1)+KJ(J-1)/2+KT(T-1)/2-K+K(J+T-1) = 251$, to be estimated from a total of $J\times T\times N = 7\times 6\times 52 = 2184$ observations. For a non parsimoniously parametrized matrix-variate normal mixture the number of parameters would be $K[JT + J(J+1)/2 + T(T+1)/2] - 1 = 454$.

in 2018 and Politics in 2019. Cluster 1 is characterized by lower correlations in time between the first 3 years (2014-2016) and the second ones (2017-2019). Moreover, it has negative correlation between Labour Market Mobility and the other dimensions, with the exception of Family Reunion. Countries in this cluster have the lowest score for the Access To Nationality and rank low for Political Participation as well, while ranking high for Family Reunion, Long-term Residence and Anti-discrimination legislation.

- **Cluster 2**: Belgium, Canada, Chile, Hungary, India, Indonesia, Israel, Japan, Mexico, New Zealand, North Macedonia, Poland, Portugal, Romania, Slovakia, Sweden, Switzerland.
  - **Correlation in time**: Cluster 2 presents high correlation values in time.
  - **Means**: with respect to the other clusters, the values of the means of this group are quite low in Politics and Education and high in Family, Residence and Anti-discrimination.
  - **Correlation among indicators**: almost all the indicators of this cluster are positively correlated, with particularly high values between Education and Labour, Politics and Labour, Politics and Education, Citizenship and Education and Citizenship and Politics.

  During the analysed period, countries belonging to this cluster did not change much their policies, and they usually rank high in all the areas. The countries of this group tend to have good policies for Residence, Family and Anti-discrimination, but rank low for Education and Politics.

- **Cluster 3**: Albania, Austria, China, Croatia, Cyprus, Finland, Germany, Greece, Iceland, Ireland, Italy, Korea, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Russia, Serbia, Spain, Ukraine, UK, USA.
  - **Correlation in time**: Cluster 3 presents the highest correlations in time with respect to the other clusters.
  - **Means**: with respect to the other clusters, this group does not present low mean values for any indicator. It presents medium values in Politics, Labour, Family, Education and Citizenship indicators and quite high values in Residence and Anti-discrimination.
  - **Correlation among indicators**: almost all the correlations values among indicators are low, with exception for Residence and Family.

  The characteristic of Cluster 3 is its high stability in time, that is the tendency to not make huge changes in the legislation, with some remarkable exceptions such as Iceland in Anti-discrimination in 2018 and Citizenship and Anti-discrimination in Luxembourg in 2017. To this cluster, belongs the countries that reformed less their immigration legislation during the study period. They tend to rank average in most of the policies areas, with the exception of Residence and Anti-discrimination laws, where they tend to rank higher. This group could be seen as the "average" cluster, grouping countries which could be located at the middle of the MIPEX overall rank. This does not mean that any country of this cluster do not present high or low values in any indicator, but that overall, among the indicators the tendency is towards the center. However, low correlation among variables signals that countries do not move homogeneously among the policies areas.

- **Cluster 4**: Bulgaria, Czech Republic, France, Turkey.

  - **Correlation in time**: it presents high values but they shades with time.
  - **Means**: with respect to the other clusters, Cluster 4 have the lowest mean values for Politics and quite low values in Education, Citizenship and Labour. It has high mean values in Anti-discrimination.
  - **Correlation among indicators**: it generally presents low correlations with the exception for an high positive value between Anti-discrimination and Residence.

  Cluster 4 is mainly characterised by its relatively low values of Politics in every country, including France. Important positive improvements in Education across time for all the countries mostly explaining the time-correlation behaviour. Despite ranking generally high for Anti-discrimination policies, countries within this cluster tend to rank low for policies in Education, Citizenship and Labour, while scoring average for Residence legislation. Yet, low correlation among variables indicates that the countries do not move homogeneously among the dimensions, with the exception of policies regarding Residence and Anti-discrimination, that have high positive correlation. Countries belonging to this cluster have seen their score moderately changing in time, indicating that some changes in the legislation have happened.

- **Cluster 5**: Argentina, Australia, Brazil, Denmark, Moldova.

  - **Correlation in time**: it presents high values but they shade faster.
  - **Means**: with respect to the other clusters, the values of the means of Cluster 5 are quite low in Education and Politics, medium in Labour and high for the other indicators.
  - **Correlation among indicators**: the values of the correlations are generally low.

  Cluster 5 collects countries with smooth evolution, in both positive and negative directions and it generally presents low values in Education (with the exception of Australia). Changes are to be noted in Residence, where all the countries (with the exception of Argentina) see their values change in time (in both directions). Countries belonging to this cluster have high correlation values in time, but they tend to decrease faster with time, meaning that some changes in the policies have been made especially in the last years. Countries of this cluster, are characterized for generally ranking low in policies related to Educational support for foreign pupils and Politics, but high in Family, Residence, Citizenship and Anti-discrimination. However, the low correlation among the dimensions, means that the countries tend not to move homogeneously among them.

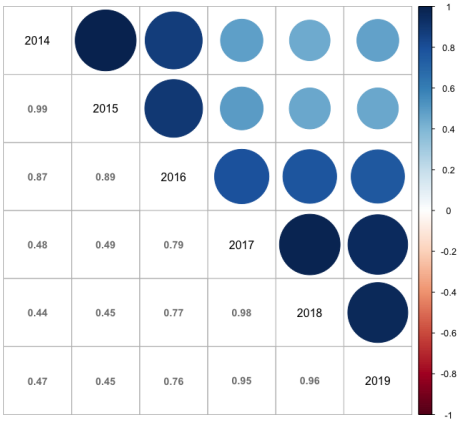
Looking at the details of the countries assigned to each cluster, it could be noticed that in the clustering process the algorithm gave more importance to the temporal and variables' dynamics (captured by $\Phi$ and $\Omega$) than to their overall scores (captured in $M$). The clustering privileged the similarity in trajectory rather than in magnitude. This gives us an idea on how the clustering should be read and explains why countries that one could think are quite different in their policies are in the same cluster.

## B.6 Conclusions

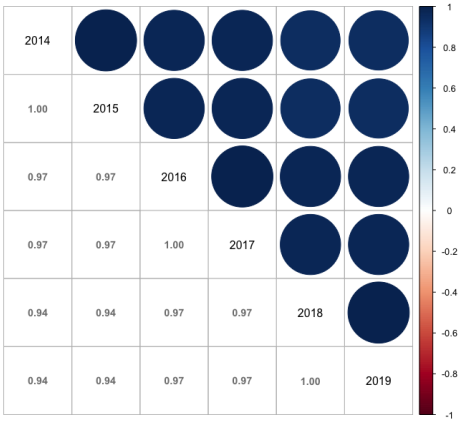
This paper has explored immigrant regulation and immigrant assimilation policies, analyzing 7 dimensions of the Migrant Integration Policy Index from the year 2014 to 2019. The need for the

analysis carried out came from the statement that when comparing very different countries from each other on social and civil issues, the identification of homogeneous groups of units substantially improves the ease of reading and the interpretation of the results. In this paper, we addressed this issue trough the application of an unsupervised clustering approach for longitudinal data namely MMN. The exploration and visualization of the data show that for the 7 MIPEX dimensions analyzed, the considered countries tend to change little over time. This behaviour led us to rely on an approach as MMN, that accounts simultaneously for the within and between time dependency structures. The identification of groups of countries with similar behaviour over time allows the comparison of clusters with each other and the comparison of the countries within each cluster. Moreover, the correlations in time shows the general trend of each indicator over time in each cluster, and the correlations between variables purified from the time effect underline the behaviour of each indicator in relation to the others within each cluster. This analysis allowed the addition of new levels of interpretation of the migration policies and of several new information about the phenomena. Specifically, the information added helps to better understand which countries have similar legislative attitudes regarding migration policies and which are following similar trends, whether they are virtuous toward integration, static, or toward the marginalization of migrants. For instance, the evidence that Bulgaria and France are both in Cluster 4 highlights that they both have relatively low values for the Politics dimension and they both improved the Citizenship dimension over the considered years.

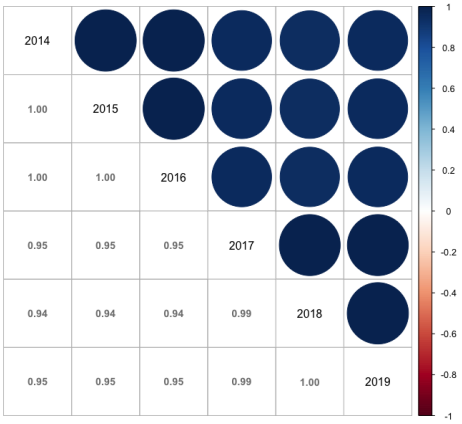
As future developments of this work, we expect, as the data will be available, to add to the analysis the Health dimension. This would be of particular interest especially during the last years of COVID-19 pandemic. Moreover, if as we expect, there will be changes in the migration policies of many of the countries considered, and, consequently, there will be changes over time in the trajectories of the considered indicators. Moreover, it will be of particular interest to estimate the probabilities to move trough the clusters along the time, through the application of Latent Markov models.



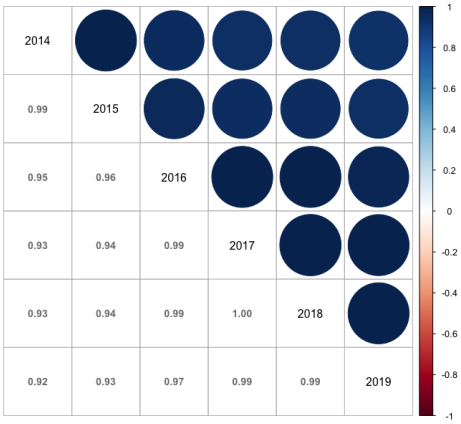

(a) Cluster 1



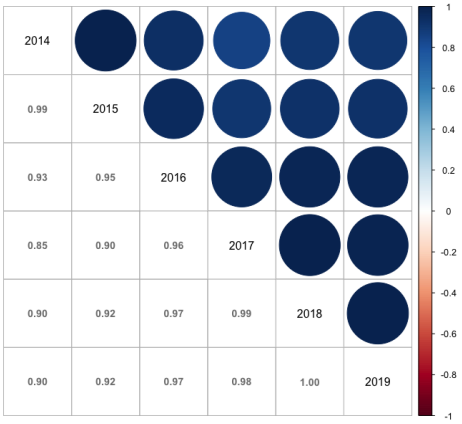

(b) Cluster 2




(c) Cluster 3

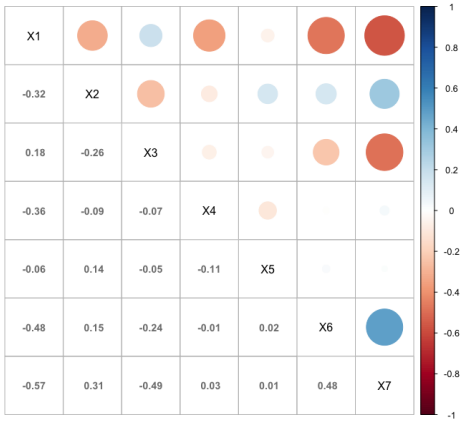



(d) Cluster 4

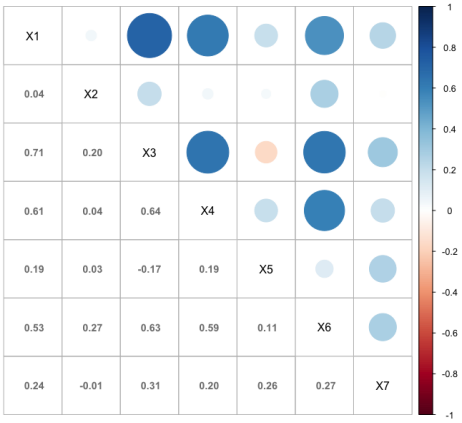



(e) Cluster 5

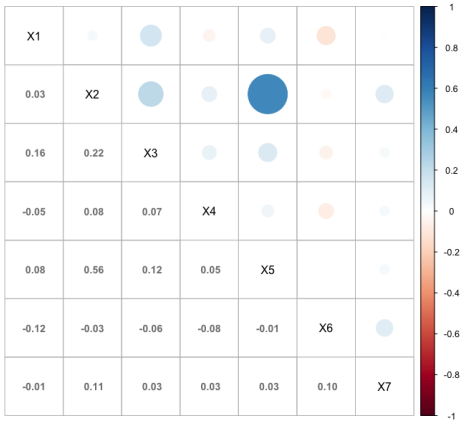
Figure B.2: MMN clusters' corr-plots in time.



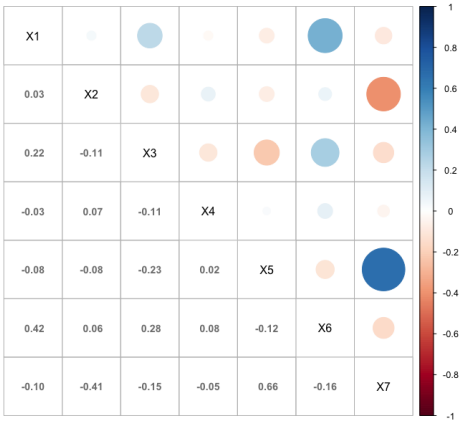

(a) Cluster 1



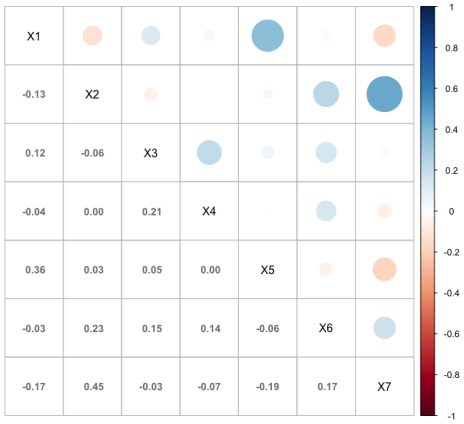

(b) Cluster 2




(c) Cluster 3




(d) Cluster 4




(e) Cluster 5

Figure B.3: MMN clusters' corr-plots among indicators. X1 Labour, X2 Family, X3 Education, X4 Politics, X5 Residence, X6 Citizenship, X7 Anti-discrimination

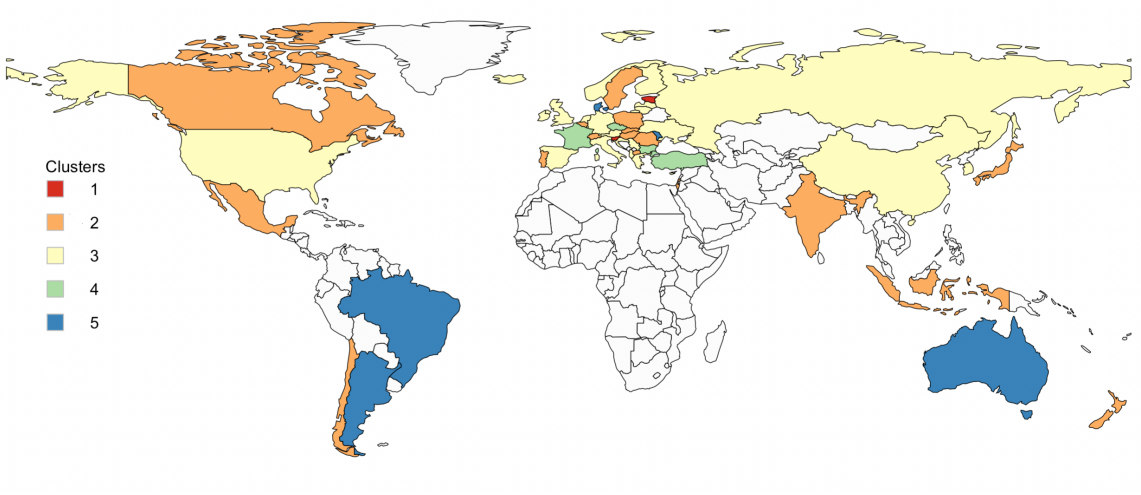



Figure B.4: MIPEX dimensional indices: MMN clusters' composition. 52 countries; years 2014 – 2019.

# Appendices

# .1 Tables

Table 1: MMN clusters' means over time.

| **Cluster 1** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
|---|---|---|---|---|---|---|
| Labour | 42.34 | 42.38 | 43.62 | 45.40 | 45.91 | 47.02 |
| Family | 65.96 | 66.00 | 67.24 | 69.02 | 69.53 | 70.64 |
| Education | 46.68 | 46.72 | 47.96 | 49.74 | 50.25 | 51.36 |
| Politics | 19.14 | 19.18 | 20.42 | 22.21 | 22.72 | 23.83 |
| Residence | 71.85 | 71.89 | 73.13 | 74.92 | 75.43 | 76.54 |
| Citizenship | 16.64 | 16.68 | 17.92 | 19.71 | 20.22 | 21.33 |
| Anti-discrimination | 66.12 | 66.16 | 67.40 | 69.19 | 69.70 | 70.81 |
| **Cluster 2** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
| Labour | 48.34 | 48.34 | 48.50 | 48.50 | 48.56 | 48.56 |
| Family | 64.09 | 64.09 | 64.25 | 64.25 | 64.31 | 64.31 |
| Education | 39.51 | 39.51 | 39.67 | 39.67 | 39.73 | 39.73 |
| Politics | 32.52 | 32.52 | 32.68 | 32.68 | 32.74 | 32.74 |
| Residence | 66.79 | 66.79 | 66.95 | 66.95 | 67.00 | 67.00 |
| Citizenship | 49.45 | 49.45 | 49.60 | 49.60 | 49.66 | 49.66 |
| Anti-discrimination | 71.20 | 71.20 | 71.36 | 71.36 | 71.42 | 71.42 |
| **Cluster 3** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
| Labour | 51.65 | 52.37 | 52.37 | 52.95 | 53.41 | 53.53 |
| Family | 49.99 | 50.71 | 50.71 | 51.29 | 51.75 | 51.87 |
| Education | 44.42 | 45.14 | 45.14 | 45.71 | 46.18 | 46.30 |
| Politics | 39.46 | 40.18 | 40.18 | 40.75 | 41.22 | 41.34 |
| Residence | 58.94 | 59.65 | 59.65 | 60.23 | 60.70 | 60.81 |
| Citizenship | 49.02 | 49.74 | 49.74 | 50.31 | 50.78 | 50.90 |
| Anti-discrimination | 65.39 | 66.11 | 66.11 | 66.68 | 67.15 | 67.27 |
| **Cluster 4** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
| Labour | 38.85 | 39.28 | 41.76 | 43.25 | 43.63 | 44.56 |
| Family | 46.45 | 46.87 | 49.35 | 50.84 | 51.22 | 52.16 |
| Education | 28.89 | 29.32 | 31.80 | 33.29 | 33.67 | 34.60 |
| Politics | 11.13 | 11.56 | 14.03 | 15.53 | 15.91 | 16.84 |
| Residence | 50.61 | 51.04 | 53.51 | 55.01 | 55.39 | 56.32 |
| Citizenship | 37.59 | 38.02 | 40.49 | 41.99 | 42.36 | 43.30 |
| Anti-discrimination | 67.18 | 67.61 | 70.09 | 71.58 | 71.96 | 72.89 |
| **Cluster 5** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
| Labour | 52.29 | 51.65 | 53.22 | 54.82 | 54.03 | 53.85 |
| Family | 62.32 | 61.69 | 63.26 | 64.85 | 64.06 | 63.89 |
| Education | 34.76 | 34.13 | 35.69 | 37.29 | 36.50 | 36.32 |
| Politics | 40.98 | 40.34 | 41.91 | 43.51 | 42.72 | 42.54 |
| Residence | 61.75 | 61.12 | 62.69 | 64.28 | 63.50 | 63.32 |
| Citizenship | 65.94 | 65.30 | 66.87 | 68.46 | 67.68 | 67.50 |
| Anti-discrimination | 74.32 | 73.68 | 75.25 | 76.85 | 76.06 | 75.88 |

# List of Figures

# List of Tables

# Résumé Long en Français

## Chapitre 1 : Introduction et contexte scientifique

Les enquêtes longitudinales, qui suivent les individus ou les unités au fil du temps, sont essentielles pour étudier les trajectoires de développement, les changements comportementaux et l'impact des interventions. Cependant, l'hétérogénéité et les dépendances temporelles dans ces données rendent les techniques de clustering traditionnelles inadéquates. Ces enquêtes collectent souvent des données de types mixtes, incluant des réponses nominales, ordinales, quantitatives et textuelles, ce qui pose des défis significatifs pour l'analyse statistique. Cette nature mixte des réponses aux enquêtes—où différents types de données coexistent—ajoute une couche supplémentaire de complexité.
Le clustering est une technique d'apprentissage non supervisé, ce qui signifie qu'il n'y a pas de "vérité terrain" ou de labels pré-définis pour guider l'analyse. L'objectif est de maximiser la similarité intra-cluster tout en minimisant la similarité inter-cluster. Les applications du clustering sont vastes et incluent des domaines tels que la biologie, l'analyse textuelle, l'économie et la sociologie. Par exemple, la classification biologique de Linnaeus en 1735 est un exemple précoce de clustering, où les plantes et les animaux étaient catégorisés en fonction de critères subjectifs (Bouveyron, Celeux, Murphy, and Raftery, 2019b). Plus formellement, le clustering est défini comme le processus de partitionnement d'un ensemble de données en $K$ clusters, où $K$ est un nombre prédéfini de groupes. Différentes approches de clustering existent, telles que le clustering basé sur les centroïdes, le clustering hiérarchique, le clustering basé sur la densité et le clustering probabiliste. Chaque méthode définit les clusters différemment et implique des façons distinctes de traiter les questions et les défis courants en clustering, comme le choix du nombre de clusters et l'évaluation de l'incertitude du regroupement (Hennig, Meila, Murtagh, and Rocci, 2015).

Les méthodes actuelles pour traiter ces données transforment souvent les variables ordinales et de comptage en variables continues, car elles sont plus faciles à manipuler. Cependant, cette approche introduit des biais ou entraîne une perte d'information significative (Liddell and Kruschke, 2018). Par exemple, la transformation des données ordinales en échelle de Likert, bien que courante, peut introduire des distances artificielles entre les catégories, biaisant ainsi les résultats (Lewis et al., 2005). De plus, ignorer l'ordre des catégories ordinales en les traitant comme des données nominales entraîne une perte d'information cruciale (Vermunt and Magidson, 2005).
En outre, les enquêtes longitudinales nécessitent des modèles capables de capturer l'évolution temporelle des comportements. Les approches actuelles analysent souvent chaque phase temporelle indépendamment, puis tentent de trouver des liens entre ces analyses a posteriori. Cette méthode ne permet pas de modéliser correctement l'évolution temporelle, rendant l'analyse des comportements longitudinaux complexe et imprécise (Selosse, Jacques, Biernacki, and Cousson-Gélie, 2019).

Lorsque ces deux défis scientifiques sont réunis, la littérature scientifique devient encore plus mince, car les chercheurs auraient besoin d'un outil capable de gérer des données longitudinales de type mixte, comme celles recueillies dans (François-Lecompte, Innocent, Kréziak, and Prim-Allaz, 2020). Cette thèse se concentre sur le développement d'algorithmes de clustering probabiliste pour l'analyse de données mixtes longitudinales. Pour répondre à ces défis, cette thèse propose un nouveau modèle de clustering pour les données longitudinales de types mixtes. Ce modèle utilise des melanges de distributions normales matricielles (Viroli, 2011a) pour modéliser simultanément l'hétérogénéité, l'association entre les réponses et la structure de dépendance temporelle. L'objectif est de fournir une méthode capable de regrouper les unités ayant des comportements similaires au fil du temps, tout en étant accessible aux chercheurs et praticiens sans formation statistique avancée.

## Chapitre 2 : Clustering probabiliste

une méthode clé du clustering probabiliste sont les modèles de mélanges finis (FMMs), où la densité de probabilité d'une observation est exprimée comme une somme pondérée de densités de composantes. Cette approche a été introduite par Pearson, 1894 et a depuis été largement utilisée dans divers domaines.

**Modèles de Mélanges Finis (FMMs)** Les modèles de mélanges finis expriment la densité de probabilité d'une observation $y_i$ comme une somme pondérée de densités de composantes :

$$f(y_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k) \tag{C.1}$$

où $\pi_k \geq 0$ pour tout $k$, et $\sum_{k=1}^{K} \pi_k = 1$. Les paramètres $\pi_k$ sont les proportions de mélange, et $f_k(y_i;\theta_k)$ sont les densités de composantes avec paramètres $\theta_k$, avec $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. Le modèle peut être caractérisé hiérarchiquement en introduisant un vecteur latent $\ell_i \in \{0,1\}^K$, où $\ell_{ik} = 1$ si $y_i$ appartient à la $k$-ème composante, et $\ell_{ik} = 0$ sinon. La densité conjointe de $(y_i, \ell_i)$ est :

$$f(y_i, \ell_i;\boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k f_k(y_i;\theta_k)]^{\ell_{ik}} \tag{C.2}$$

La densité marginale de $y_i$ est alors :

$$f(y_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k) \tag{C.3}$$

La log-vraisemblance des données observées est :

$$l_o(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k) \tag{C.4}$$

Lorsque toutes les composantes $f_k$ appartiennent à la même famille de distributions et leur forme paramétrique est connue, on parle de FMMs homogènes paramétriques, comme les modèles de mélanges gaussiens (GMMs) (McLachlan and Peel, 2000).

**Algorithme Expectation-Maximisation (EM)** L'algorithme EM est la méthode principale pour estimer les paramètres des FMMs (Titterington, Smith, and Makov, 1985). Il a été développé par Dempster, Laird, and Rubin, 1977 pour résoudre les problèmes d'estimation du maximum de vraisemblance (MLE) impliquant des variables latentes ou des données incomplètes. L'algorithme alterne entre deux étapes : l'étape d'expectation (E-step) et l'étape de maximisation (M-step).
L'E-step consiste à calculer l'expectation de la log-vraisemblance complète conditionnellement aux données observées et aux paramètres actuels :

$$\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)}) = \mathbb{E}\left(l_c(\boldsymbol{\theta}; y, \boldsymbol{\ell}) \mid \hat{\boldsymbol{\theta}}^{(s)}, y\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\tau}_{ik}^{(s+1)} \left\{\log \hat{\pi}_k^{(s)} + \log f_k(y_i; \hat{\theta}_k^{(s)})\right\}, \tag{C.5}$$

où $\hat{\tau}_{ik}^{(s+1)}$ est la probabilité postérieure que $y_i$ appartient à la $k$-ème composante à l'itération $s+1$ :

$$\hat{\tau}_{ik}^{(s+1)} = \frac{\hat{\pi}_k^{(s)} f_k(y_i; \hat{\theta}_k^{(s)})}{\sum_{k'=1}^{K} \hat{\pi}_{k'}^{(s)} f_{k'}(y_i; \hat{\theta}_{k'}^{(s)})} \tag{C.6}$$

L'M-step consiste à maximiser $\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)})$ par rapport à $\boldsymbol{\theta}$ :

$$\hat{\boldsymbol{\theta}}^{(s+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(s)}) \tag{C.7}$$

Les proportions de mélange $\hat{\pi}_k^{(s+1)}$ sont mises à jour comme suit :

$$\hat{\pi}_k^{(s+1)} = \frac{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)}}{N}, \forall k \in \{1, \ldots, K\} \tag{C.8}$$

Les paramètres $\hat{\boldsymbol{\theta}}_k^{(s+1)}$ sont mis à jour en résolvant :

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\tau}_{ik}^{(s+1)} \frac{\partial \log f_k(y_i; \theta_k)}{\partial \boldsymbol{\theta}} = 0 \tag{C.9}$$

L'algorithme EM alterne entre les étapes E et M jusqu'à convergence. Bien qu'il ne garantisse pas la convergence vers le maximum global de la log-vraisemblance observée, il est assuré de converger vers un maximum local ou un point selle sous certaines conditions (Wu, 1983).

**Modèles de Mélanges Gaussiens (GMMs)** Les modèles de mélanges gaussiens (GMMs) sont une classe commune de FMMs où les densités de composantes sont des distributions gaussiennes multivariées. Les paramètres $\theta_k$ incluent le vecteur de moyenne $\mu_k$ et la matrice de covariance $\Sigma_k$. Le GMM peut être écrit comme :

$$f(y_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_J(y_i \mid \mu_k, \Sigma_k) \tag{C.10}$$

où $\mathcal{N}_J(y_i \mid \mu_k, \Sigma_k)$ est la densité d'une distribution gaussienne multivariée :

$$\mathcal{N}_J(y_i \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^J |\Sigma_k|}} \exp\left\{-\frac{1}{2}(y_i - \mu_k)^\top \Sigma_k^{-1}(y_i - \mu_k)\right\} \tag{C.11}$$

Les paramètres du modèle sont mis à jour dans l'M-step en maximisant la log-vraisemblance complète attendue. Les mises à jour des paramètres sont données par :

$$\hat{\pi}_k^{(s+1)} = \frac{\hat{n}_k^{(s)}}{N}, \quad \hat{\mu}_k^{(s+1)} = \frac{\sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} y_i}{\hat{n}_k^{(s+1)}} \tag{C.12}$$

$$\hat{\Sigma}_k^{(s+1)} = \frac{1}{\hat{n}_k^{(s+1)}} \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)} (y_i - \hat{\mu}_k^{(s+1)})(y_i - \hat{\mu}_k^{(s+1)})^\top \tag{C.13}$$

où $\hat{n}_k = \sum_{i=1}^N \hat{\tau}_{ik}^{(s+1)}$.

**Paramétrisation des GMMs** Pour réduire le nombre de paramètres à estimer, la matrice de covariance $\Sigma_k$ peut être décomposée en utilisant les valeurs propres et les vecteurs propres :

$$\Sigma_k = \lambda_k D_k A_k D_k^\top \tag{C.14}$$

où $\lambda_k \in \mathbb{R}$ détermine le volume et est égal à $|\Sigma_k|^{1/J}$, $D_k \in \mathbb{R}^{J \times J}$ est la matrice des vecteurs propres de $\Sigma_k$ et détermine l'orientation, et $A_k \in \mathbb{R}^{J \times J}$ est une matrice diagonale telle que $|A_k| = 1$ avec les valeurs propres normalisées de $\Sigma_k$ et détermine la forme de la $k$-ème composante. Différentes restrictions sur $\lambda_k, D_k$ et $A_k$ conduisent à quatorze modèles différents, chacun avec des propriétés géométriques distinctes (Banfield and Raftery, 1993, Celeux and Govaert, 1995).

## Chapitre 3 : Revue de la Littérature sur le Clustering probabiliste pour les Données Mixtes Longitudinales

La littérature sur le clustering des données longitudinales de types mixtes est limitée. Néanmoins, compte tenu de la littérature limitée disponible sur ces méthodes, nous avons jugé utile pour le lecteur intéressé d'inclure également une enquête sur les principales approches de clustering (probabiliste) des données de type mixte et longitudinales.

**Données longitudinales** Les approches traditionnelles de clustering échouent souvent à capturer les dépendances entre les observations au fil du temps, conduisant à des résultats sous-optimaux ou souffrent de sur-paramétrisation et de complexité computationnelle. Par exemple, les modèles de Markov latents (LMMs) et leurs extensions, comme les modèles de Markov latents mixtes (MHMMs), offrent une manière de modéliser l'évolution temporelle par des états latents. Cependant, les LMMs supposent une homogénéité des trajectoires entre les individus, ce qui peut ne pas être vrai dans des populations hétérogènes. Les MHMMs adressent cette limitation en introduisant une variable aléatoire supplémentaire pour capturer l'hétérogénéité entre les individus, mais cela augmente la complexité computationnelle et l'interprétation des résultats (Bartolucci, Farcomeni, and Pennoni, 2012, 2014).

**Données mixtes** Le traitement des données de types mixtes pose également des défis significatifs. Les modèles comme le modèle de classes latentes (LCM) (Everitt, 1984) et clustMD (McParland and Gormley, 2016) supposent une indépendance conditionnelle entre les variables de différents types étant donnée l'appartenance à un cluster, ce qui simplifie la modélisation mais peut ignorer des dépendances importantes entre les variables. Le modèle de blocs latents multiples (MLBM) (Selosse, Jacques, and Biernacki, 2020), bien qu'il offre une co-clustering flexible, ne permet pas aux clusters de variables d'être de types différents, limitant ainsi sa capacité à capturer les associations entre variables de types différents.

**Données longitudinales mixtes** L'intégration des données longitudinales et de types mixtes pose des défis supplémentaires. Les modèles comme les mélanges de modèles linéaires généralisés (MMGLMM) (Komarek and Komárková, 2013) et les modèles de classes latentes linéaires mixtes (LCLMMs) (Proust-Lima, Philipps, and Liquet, 2017) offrent des solutions prometteuses en combinant des effets aléatoires et des variables latentes pour capturer à la fois les dépendances temporelles et les dépendances entre variables de types différents. Cependant, ces modèles sont souvent complexes à estimer et à interpréter. De plus, l'interprétation des résultats de ces modèles complexes reste un défi pour les praticiens sans formation statistique avancée, car les relations entre les variables et les clusters peuvent ne pas être directement évidentes.

**Conclusion** La disponibilité des logiciels est également un aspect crucial. Bien que plusieurs packages R fournissent des implémentations de nombreux modèles, ils se concentrent souvent sur des types de données spécifiques ou nécessitent une expertise statistique avancée pour une utilisation efficace. Le développement d'outils logiciels plus conviviaux et complets pourrait considérablement améliorer l'accessibilité et l'applicabilité de ces modèles dans des contextes pratiques. En conclusion, bien que des progrès significatifs aient été réalisés dans le développement de modèles pour les données longitudinales de types mixtes au cours de la dernière décennie, il peut encore y avoir un besoin de solutions plus parcimonieuses, interprétables et computationnellement efficaces. Les recherches futures devraient se concentrer sur la résolution des problèmes de sur-paramétrisation, de complexité computationnelle et de l'interprétation des résultats. De plus, le développement d'outils logiciels plus conviviaux pourrait grandement faciliter l'adoption de ces méthodes dans divers domaines, permettant aux praticiens d'inclure efficacement le clustering des données longitudinales de types mixtes dans leur flux de travail.

## Chapitre 4 : Clustering des Données Ordinales Longitudinales via Mélanges Finis de Distributions Matricielles.

On présente un modèle de clustering pour les données ordinales longitudinales, appelé Mixture of Ordinal Matrices (MOM) (Amato, Jacques, and Prim-Allaz, 2024). Ce modèle suppose qu'une variable ordinale est la discrétisation d'une variable continue sous-jacente et utilise un mélange de distributions matricielles normales pour modéliser les données. Le modèle est capable de capturer simultanément l'hétérogénéité, l'association entre les réponses et la structure de dépendance temporelle.

**Mélange de normales matricielles et modèle MOM** Soit $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, une distribution matricielle normale où $M \in \mathbb{R}^{J\times T}$ est la matrice des moyennes, $\Phi \in \mathbb{R}^{T\times T}$ est la

matrice de covariance contenant les variances et covariances entre les $T$ occasions ou temps, et $\Sigma \in \mathbb{R}^{J\times J}$ est la matrice de covariance contenant les variances et covariances des $J$ variables. La fonction de densité de probabilité (pdf) de la distribution matricielle normale est donnée par :

$$f(Z \mid M, \Phi, \Sigma) = (2\pi)^{-\frac{JT}{2}} |\Phi|^{-\frac{J}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^{\top}\right]\right\} \tag{C.15}$$

La distribution matricielle normale est une extension naturelle de la distribution normale multivariée, car si $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, alors $\text{vec}(Z) \sim \mathcal{N}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$, où vec(.) est l'opérateur de vectorisation et $\otimes$ désigne le produit de Kronecker (Gupta and Nagar, 2000).

Le modèle MOM suppose que chaque variable ordinale $y_{ijt}$ est la manifestation d'une variable continue sous-jacente $z_{ijt}$ qui suit une distribution normale. Le modèle MOM utilise un mélange de distributions matricielles normales (MMN) (Viroli, 2011a) pour modéliser les données. La pdf du modèle MMN est donnée par :

$$f(Y \mid \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{MN}_{(J\times T)}(Y \mid M_k, \Phi_k, \Sigma_k) \tag{C.16}$$

où $K$ est le nombre de composantes du mélange, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$ est le vecteur des proportions de mélange, et $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^{K}$ est l'ensemble des paramètres spécifiques aux composantes avec $\Theta_k = \{M_k, \Phi_k, \Sigma_k\}$.

**Hypothèses du modèle MOM** Pour mapper de $Y_i$ à $Z_i$, soit $\gamma_j$ le vecteur de seuils de dimension $C_j + 1$ pour la $j$-ème variable ordinale qui a $C_j$ niveaux et soit les seuils contraints tels que $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \ldots \leq \gamma_{j,C_j} = \infty$. Si la latente $z_{ijt}$ est telle que $\gamma_{j,c-1} < z_{ijt} < \gamma_{j,c}$, alors la réponse ordinale observée $y_{ijt} = c$.

Soit $\mathcal{O}^{J\times T}$ l'ensemble de toutes les matrices ordinales possibles de taille $J \times T$ dont la ligne générale $j$ prend des valeurs dans $\{1, \ldots, C_j\}$. Chaque élément de $\mathcal{O}^{J\times T}$ est appelé un patron de réponse, c'est-à-dire chaque élément de l'ensemble représente l'une des configurations possibles (patron) de la matrice ordinale $J \times T$, étant donné les niveaux $C_j$. Soit $R$ la cardinalité de $\mathcal{O}^{J\times T}$. Chaque patron de réponse $Y_r \in \mathcal{O}^{J\times T}$ est généré par une portion $\Omega_r$ de l'espace latent $\mathbb{R}^{J\times T}$ selon les seuils $\gamma := \{\gamma_j\}_{j=1}^{J}$. Soit le vecteur binaire $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iR})$ une représentation one-hot de $Y_i$ telle que si le $r$-ème patron est observé alors $\tilde{Y}_{ir} = 1$ et toute autre entrée du vecteur est égale à zéro. Nous pouvons dériver la densité conjointe de $Z_i, \tilde{Y}_i, \ell_i$ comme :

$$f(\tilde{Y}_i, Z_i, \ell_i) = f(\tilde{Y}_i \mid Z_i, \ell_i) f(Z_i \mid \ell_i) f(\ell_i) \tag{C.17}$$

En supposant que :

$$\begin{aligned} &\ell_i \sim \mathcal{M}(1, \boldsymbol{\pi}), \boldsymbol{\pi} := (\pi_1, \ldots, \pi_K), \\ &Z_i \mid \ell_{ik} = 1 \sim \mathcal{MN}_{(J\times T)}(Z_i \mid \Theta_k), \Theta_k := \{M_k, \Phi_k, \Sigma_k\}, \\ &\tilde{Y}_i \mid Z_i, \ell_{ik} = 1 \sim \mathcal{M}(1, \xi_i), \xi_i := (\mathbf{1}_{\Omega_1}(Z_i), \ldots, \mathbf{1}_{\Omega_R}(Z_i)), \end{aligned} \tag{C.18}$$

Où $\mathcal{M}$ indique la distribution multinomiale et $\mathbf{1}_{\Omega_r}(Z_i)$ est la fonction indicatrice qui est égale à 1 lorsque les éléments de $Z_i$ ont des valeurs qui déterminent le $r$-ème patron. Donc, lorsque $\tilde{Y}_{ir} = 1$,

le vecteur $\xi_i$ est un vecteur dont le $r$-ème élément est égal à 1 et tous les autres sont égaux à 0. Dans la suite, $\boldsymbol{Z} := \{Z_i\}_{i=1}^N, \boldsymbol{\ell} := \{\ell_i\}_{i=1}^N$ et $\boldsymbol{\Theta} := \{\Theta_k, \pi_k\}_{k=1}^K$ indiquent les ensembles de $Z_i, \ell_i$ et des paramètres, respectivement. Enfin, soit $\tilde{\mathbf{Y}} := \{\tilde{Y}_i\}_{i=1}^N$ la collection des vecteurs de patrons de réponse observés $Y_i$.

**Inférence** Ainsi, la log-vraisemblance complète peut être écrite comme :

$$\log \mathcal{L}_C(\boldsymbol{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \boldsymbol{\ell}) = \sum_{i=1}^{N} \left\{ \sum_{r=1}^{R} \widehat{Y}_{ir} \mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K} \ell_{ik} \left[ \log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \right.\right.$$
$$\left.\left. \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} \operatorname{tr}\left[ \Sigma_k^{-1} (Z_i - M_k) \Phi_k^{-1} (Z_i - M_k)^\top \right] \right] \right\} \tag{C.19}$$

L'inférence est faite grace à un algorithm EM-MCMC (McLachlan and Krishnan, 2008).

**Evaluation sur données synthétiques** Les résultats sur les données synthétiques montrent que le modèle MOM est capable de récupérer les partitions simulées avec une bonne précision, même en présence de bruit. L'évaluation montre que le modèle est robuste et capable de capturer les tendances temporelles et de regrouper les individus en fonction de leurs comportements.

**Application** L'application du modèle MOM à une étude réelle sur les changements de comportements alimentaires pendant la pandémie de Covid-19 en France (François-Lecompte, Innocent, Kréziak, and Prim-Allaz, 2020) montre que le modèle est capable de capturer les tendances temporelles et de regrouper les individus en fonction de leurs comportements alimentaires. Les résultats identifient plusieurs clusters de comportements alimentaires, ce qui facilite l'évaluation et la comparaison des individus au sein de chaque cluster et entre les différents clusters au fil du temps. Cette application démontre l'utilité pratique du modèle pour analyser des données complexes dans des contextes réels et fournit des insights précieux pour la compréhension des comportements longitudinaux. Par exemple, un cluster regroupe les individus qui pense que la pandemie a été une occasion pour mieux reflechir sur leur style de vie et habitudes alimentaires, tandis qu'un autre cluster regroupe ceux qui n'ont pas changé leur habitudes alimentaires. Ces clusters permettent de mieux comprendre les dynamiques de changement de comportement en réponse à des événements externes, comme la pandémie.

## Chapitre 5 : MMM : Clustering de Données Multivarié longitudinales Mixtes

Le modèle MMM (Mixture of Mixed-Matrices) (Amato and Jacques, 2024) est une extension du modèle MOM (Mixture of Ordinal Matrices) pour les données de types mixtes. Il repose sur un mélange de distributions matricielles normales pour effectuer le clustering dans la dimension latente. Les variables continues, ordinales, binaires, nominales et de comptage sont modélisées comme suit :

- **Variables continues** Les variables continues sont modélisées comme suit :

$$z_{ijt} \sim \mathcal{N}(\mu_j, \sigma_j^2) \tag{C.20}$$

où $z_{ijt}$ est la valeur de la variable continue $j$ pour l'unité $i$ au temps $t$, $\mu_j$ est la moyenne et $\sigma_j^2$ est la variance.

- **Variables ordinales** Les variables ordinales sont modélisées comme des manifestations de variables latentes continues. Soit $y_{ijt}$ la valeur observée de la variable ordinale $j$ pour l'unité $i$ au temps $t$, et $z_{ijt}$ la variable latente continue sous-jacente. Les seuils $\gamma_j$ sont fixés de manière à partitionner la droite réelle en $C_j$ intervalles, où $C_j$ est le nombre de niveaux de la variable ordinale $j$. La probabilité d'observer le niveau $c$ est donnée par :

$$\mathbb{P}(y_{ijt} = c) = \Phi\left(\frac{\gamma_{j,c} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,c-1} - \mu_j}{\sigma_j}\right) \tag{C.21}$$

où $\Phi$ est la fonction de répartition cumulative de la distribution normale standard.

- **Variables nominales** Les variables nominales sont modélisées comme des manifestations de variables latentes continues multidimensionnelles. Soit $y_{ijt}$ la valeur observée de la variable nominale $j$ pour l'unité $i$ au temps $t$, et $z_{ijt}$ la variable latente continue sous-jacente. La variable latente $z_{ijt}$ est un vecteur de dimension $C_j - 1$, où $C_j$ est le nombre de niveaux de la variable nominale $j$. La probabilité d'observer le niveau $c$ est donnée par :

$$y_{ijt} = \begin{cases} 1 & \text{si } \max_s\{z_{ijt}^s\} < 0 \\ c & \text{si } z_{ijt}^{c-1} = \max_s\{z_{ijt}^s\} \text{ et } z_{ijt}^{c-1} > 0 \text{ pour } s = 2, \ldots, C_j \end{cases} \tag{C.22}$$

- **Variables de comptage** Les variables de comptage sont modélisées comme des manifestations de variables latentes continues suivant une distribution normale, discrétisées selon des seuils fixes. Soit $y_{ijt}$ la valeur observée de la variable de comptage $j$ pour l'unité $i$ au temps $t$, et $z_{ijt}$ la variable latente continue sous-jacente. Les seuils $\gamma_j$ sont fixés de manière à partitionner la droite réelle en $C_j$ intervalles, où $C_j$ est le nombre de niveaux de la variable de comptage $j$. La probabilité d'observer le niveau $c$ est donnée par :

$$\mathbb{P}(y_{ijt} = c) = \Phi\left(\frac{\gamma_{j,c} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,c-1} - \mu_j}{\sigma_j}\right) \tag{C.23}$$

où $\Phi$ est la fonction de répartition cumulative de la distribution normale standard.

**Modèle conjoint** Nous pouvons alors considérer $Y_i$ comme une matrice par blocs et la diviser commodément entre les $C$ premières lignes, représentant les variables continues observées, suivies des $O$ lignes représentant les variables catégoriques et des $J - C - O = G$ lignes restantes, représentant les variables de comptage. Notez que la division s'effectue uniquement sur les lignes et non sur les colonnes. Nous pouvons alors écrire $Y_i = [Y_i^\alpha, Y_i^\beta, Y_i^\gamma]^\intercal$, où $Y_i^\alpha \in \mathbb{R}^{C \times T}$ contient les variables continues, $Y_i^\beta \in \mathbb{N}^{O \times T}$ regroupe les variables catégoriques (codées sous forme d'entiers) ainsi que les variables binaires, et $Y_i^\gamma \in \mathbb{N}_0^{G \times T}$ contient les variables de comptage.

À ce stade, nous supposons que chaque bloc observé de la matrice $Y_i$ est en réalité la manifestation du bloc correspondant de la matrice aléatoire latente $Z_i = [Z_i^\alpha, Z_i^\beta, Z_i^\gamma]^\intercal$, et que cette matrice latente est liée à $Y_i$ par différentes relations en fonction du type de chaque variable $y_{ijt}$, comme décrit précédemment.
Nous supposons alors une mixture de distributions normales matricielles dans l'espace latent. Par conséquent, nous pouvons écrire

$$\begin{pmatrix} Z_i^\alpha \\ Z_i^\beta \\ Z_i^\gamma \end{pmatrix} \sim \sum_{k=1}^{K} \pi_k \, \mathcal{MN}_{(J\times T)} \left( \begin{pmatrix} M_k^\alpha \\ M_k^\beta \\ M_k^\gamma \end{pmatrix}, \Phi_k, \begin{pmatrix} \Sigma_k^{\alpha\alpha} & \Sigma_k^{\alpha\beta} & \Sigma_k^{\alpha\gamma} \\ \Sigma_k^{\beta\alpha} & \Sigma_k^{\beta\beta} & \Sigma_k^{\beta\gamma} \\ \Sigma_k^{\gamma\alpha} & \Sigma_k^{\gamma\beta} & \Sigma_k^{\gamma\gamma} \end{pmatrix} \right). \tag{C.24}$$

À partir de là, nous pouvons dériver le modèle conjoint. Pour assurer la cohérence de la notation, définissons $\tilde{Y}_i^\beta$ comme l'encodage one-hot de la partie catégorielle de $Y_i$. En plus de $Z_i$, nous introduisons un vecteur binaire d'allocation latente de dimension $K$, indiquant si l'unité $i$ appartient au $k$-ième cluster, $\ell_i = (\ell_{i1}, \ldots, \ell_{iK})$, où $\ell_{ik} = 1$ si l'unité $i$ appartient au cluster $k$.
En rappelant les liens entre les variables observées et latentes, nous pouvons exprimer notre modèle à travers les hypothèses distributionnelles suivantes :

$$\begin{aligned}
&\ell_i \sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} := (\pi_1, \ldots, \pi_K) \\
&Z_i^\alpha | \ell_{ik} = 1 \sim \mathcal{MN}_{(C\times T)}(Z_i^\alpha | \Theta_k^\alpha),\ \Theta_k^\alpha := \{M_k^\alpha, \Phi_k, \Sigma_k^\alpha\}, \\
&Z_i^\beta | Z_i^\alpha, \ell_{ik} = 1 \sim \mathcal{MN}_{(O\times T)}(Z_i^\beta | \Theta_k^{\beta|\alpha}),\ \Theta_k^\beta := \{M_k^{\beta|\alpha}, \Phi_k, \Sigma_k^{\beta|\alpha}\}, \\
&Z_i^\gamma | Z_i^\alpha, Z_i^\beta, \ell_{ik} = 1 \sim \mathcal{MN}_{(G\times T)}(Z_i^\gamma | \Theta_k^{\gamma|\alpha,\beta}),\ \Theta_k^\gamma := \{M_k^{\gamma|\alpha,\beta}, \Phi_k, \Sigma_k^{\gamma|\alpha,\beta}\}; \\
&\tilde{Y}_i^\beta | Z_i^\beta, \ell_{ik} = 1 \sim \mathcal{M}(1, \xi_i),\ \xi_i := (\mathbf{1}_{\Omega_1}(Z_i^\beta), ..., \mathbf{1}_{\Omega_R}(Z_i^\beta)), \\
&Y_{igt}^\gamma | Z_{igt}^\gamma \sim \mathcal{P}(\exp(Z_{igt}^\gamma)),
\end{aligned}$$

où $\mathcal{M}$ désigne la distribution multinomiale et $\mathbf{1}_{\Omega_r}(Z_i^\beta)$ est la fonction indicatrice qui vaut 1 lorsque les éléments de $Z_i^\beta$ prennent des valeurs déterminant le $r$-ième motif. Ainsi, lorsque $\tilde{Y}_{ir}^\beta = 1$, le vecteur $\xi_i$ est un vecteur dont le $r$-ième élément vaut 1 et tous les autres 0.
De plus, afin d'éviter de supposer l'indépendance entre les différents blocs, pour relier les distributions latentes matricielles, nous conditionnons un bloc à un autre en exploitant les propriétés des distributions normales matricielles (Gupta and Nagar, 2000). Ainsi, $\Theta_k^{\gamma|\alpha,\beta} := \{M_k^{\gamma|\alpha,\beta}, \Phi_k, \Sigma_k^{\gamma|\alpha,\beta}\}$, plus précisément $M_k^{\gamma|\alpha,\beta} = M_k^\gamma + \Sigma_k^{\gamma\cdot}\Sigma_k^{-1,\cdot\cdot}(Z_i^{\alpha,\beta} - M_k^{\alpha,\beta})$ et $\Sigma_k^{\gamma|\alpha,\beta} = \Sigma_k^{\gamma\gamma} - \Sigma_k^{\gamma\cdot}\Sigma_k^{-1,\cdot\cdot}\Sigma_k^{\cdot\gamma}$, et où $\Theta_k^{\beta|\alpha} := \{M_k^{\beta|\alpha}, \Phi_k, \Sigma_k^{\beta|\alpha}\}$, plus précisément $M_k^{\beta|\alpha} = M_k^\beta + \Sigma_k^{\beta\alpha}\Sigma_k^{-1,\alpha\alpha}(Y_i^\alpha - M_k^\alpha)$ et $\Sigma_k^{\beta|\alpha} = \Sigma_k^{\beta\beta} - \Sigma_k^{\beta\alpha}\Sigma_k^{-1,\alpha\alpha}\Sigma_k^{\alpha\beta}$.

Enfin, en supposant que le motif observé de $\tilde{Y}_i^\beta$ est $r$ pour simplifier la notation, nous pouvons composer la distribution de chaque matrice mixte observée comme suit :

$$\begin{aligned}
Y_i \sim \sum_{k=1}^{K} \pi_k \, &\mathcal{MN}_{(C\times T)}(Z_i^\alpha | \Theta_k^\alpha) \cdot \int_{\Omega_r} \mathcal{MN}_{(O\times T)}(Z_i^\beta | \Theta_k^{\beta|\alpha}) dZ_i^\beta \\
&\cdot \int_{\mathbb{R}} \prod_t^T \prod_g^G \mathcal{P}(y_{igt}^\gamma | \exp(z_{igt}^\gamma)) \cdot \mathcal{MN}_{(G\times T)}(Z_i^\gamma | \Theta_k^{\gamma|\alpha,\beta}) dZ_i^\gamma \,.
\end{aligned} \tag{C.25}$$

**Inférence** Ainsi, on peut écrire la vraisemblance complete comme:

$$\mathcal{L}_C(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{Z}, \boldsymbol{\ell}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\Big[\pi_k \cdot \Big(\prod_{t}^{T}\prod_{g}^{G}\mathcal{P}(y_{igt}^{\gamma}|\exp(z_{igt}^{\gamma}))\Big) \cdot \mathcal{MN}_{(G\times T)}(Z_i^{\gamma}|\Theta_k^{\gamma|\alpha,\beta}) \cdot \mathcal{MN}_{(O\times T)}(Z_i^{\beta}|\Theta_k^{\beta|\alpha}) \cdot \mathcal{MN}_{(C\times T)}(Z_i^{\alpha}|\Theta_k^{\alpha}) \cdot \prod_{r=1}^{R}\mathbf{1}_{\Omega_r}(Z_i^{\beta})^{\tilde{Y}_{ir}^{\beta}}\Big]^{\ell_{ik}}. \tag{C.26}$$

L'inférence est faite grace à un algorithm EM-MCMC (McLachlan and Krishnan, 2008).

**Evaluation sur des données synthétiques** La performance du modèle MMM est évaluée à l'aide de données synthétiques, démontrant sa capacité à regrouper avec précision des données longitudinales multivariées de types mixtes. Le modèle traite efficacement la complexité des types de données mixtes et des dépendances temporelles, montrant des capacités d'inférence robustes. Les simulations révèlent que le modèle MMM peut récupérer les structures de clusters sous-jacentes et estimer les paramètres avec précision, même en présence de différentes distributions de données et de motifs temporels. Les résultats mettent en lumière la flexibilité du modèle et son potentiel pour des applications pratiques dans divers domaines scientifiques.

**Application** Le modèle MMM est appliqué à un ensemble de données du monde réel constitué de données boursières pendant la pandémie de COVID-19. Les données incluent un mélange de variables continues, telles que les prix des actions et les volumes de transactions, de variables ordinales représentant le sentiment du marché ou les niveaux de risque, et de variables de comptage comme le nombre de transactions ou d'événements d'actualité. Cet ensemble de données diversifié capture les dynamiques complexes des marchés financiers pendant une période de volatilité et d'incertitude significatives. Le modèle MMM identifie avec succès des clusters distincts d'actions basés sur leur comportement temporel et leurs caractéristiques de types mixtes, révélant des motifs qui seraient difficiles à discerner à l'aide de méthodes de clustering traditionnelles.

## Chapitre 6 : Conclusion

Cependant, le modèle proposé présente certaines limitations. Dans l'article le présentant, nous nous sommes concentrés uniquement sur la structure la plus simple de la distribution matricielle normale. Bien que considérablement plus parcimonieuse qu'un mélange de distributions normales multivariées, le modèle semble sensible aux petits échantillons, car à mesure que le nombre de clusters augmente, le nombre de paramètres à estimer peut encore devenir problématique. Pour améliorer cet aspect, les matrices de covariance peuvent être davantage décomposées pour obtenir des modèles plus flexibles et parcimonieux, comme cela a été fait, par exemple, dans Anderlucci and Viroli, 2015 et Sarkar, Zhu, Melnykov, and Ingrassia, 2020a. Une autre solution à ce problème peut être celle proposée par Cappozzo, Casa, and Fop, 2024.

De même, la structure matricielle n'est pas seulement inhérente aux données longitudinales multivariées, mais peut en réalité être trouvée dans de nombreuses autres applications. Le modèle MMM peut également être utilisé dans ces cas, avec un minimum d'ajustements requis. De plus,

l'algorithme EM peut être exploité pour étendre le modèle afin de traiter les données incomplètes sous l'hypothèse de données manquantes aléatoires (MAR).
Nous estimons que les modèles ne sont utiles pour les chercheurs dans d'autres domaines que s'ils sont fournis avec les bons outils pour implémenter l'ajustement de ces modèles, principalement via des logiciels en accès libre tels que les bibliothèques R. Étant donné que nous visons à ce que notre travail soit largement accessible, nous travaillons sur un package R pour permettre aux praticiens d'implémenter facilement le MMM, en fournissant une interface conviviale et une variété de fonctions pour permettre également l'analyse de sous-ensembles de tous les types de données possibles.
Enfin, on pourrait également envisager d'utiliser, avec les ajustements appropriés, différentes distributions continues sous-jacentes, telles que des distributions à queue lourde (Tomarchio, Punzo, and Bagnato, 2020), asymétriques (Gallaugher and McNicholas, 2018, Melnykov and Zhu, 2018) ou de Student (Doğru, Bulut, and Arslan, 2016) pour doter le modèle de clustering de différentes propriétés souhaitées.

## Annexe B : Comparaison des politiques d'intégration des migrants via mélange de lois matricielles normales.

Une étude sur le regroupement des politiques d'intégration des migrants est présenté à l'aide du modèle de Mélange de lois Matricielles Normales (MMN) (Viroli, 2011a). Les données utilisées dans cette analyse proviennent de l'Indice des Politiques d'Intégration des Migrants (MIPEX) (Niessen et al., 2007, Solano and Huddleston, 2020), qui mesure et évalue les politiques d'intégration des migrants dans 52 pays sur la période de 2014 à 2019. L'indice MIPEX est composé de 167 indicateurs de politiques répartis en huit domaines, qui sont agrégés en un indicateur composite unique pour évaluer l'efficacité des politiques d'intégration des migrants dans chaque pays. Pour cette étude, sept des huit dimensions de MIPEX ont été considérées, à l'exclusion de la dimension santé en raison des contraintes de disponibilité des données.
Le modèle MMN a été appliqué pour regrouper les pays en fonction de leurs politiques d'intégration des migrants, en tenant compte de la nature longitudinale des données. Cette approche permet de regrouper les pays qui présentent des schémas similaires dans leurs politiques d'intégration au fil du temps. Le regroupement a été effectué à l'aide du package MatTransMix de R, et le nombre optimal de groupes a été déterminé à l'aide du Critère d'Information Bayésien (BIC). Le modèle le mieux ajusté a identifié cinq groupes distincts, chacun représentant un ensemble de pays avec des trajectoires de politiques d'intégration similaires.
Les résultats de l'analyse de regroupement fournissent des informations précieuses sur les similitudes et les différences dans les politiques d'intégration des migrants dans les 52 pays. Le Cluster 1, composé de l'Estonie et de la Slovénie, se caractérise par de faibles corrélations temporelles et des scores relativement bas dans les dimensions Citoyenneté et Politique, mais des scores élevés dans les dimensions Famille, Résidence et Anti-discrimination. Le Cluster 2, le groupe le plus important, comprend des pays comme la Belgique, le Canada et la Suède, et se distingue par des corrélations temporelles élevées et des scores généralement élevés dans les dimensions Famille, Résidence et Anti-discrimination. Le Cluster 3, qui inclut des pays comme l'Albanie, l'Autriche et l'Allemagne, montre la plus grande stabilité temporelle et des scores moyens dans la plupart des dimensions. Le Cluster 4, composé de la Bulgarie, la République tchèque, la France et la Turquie, se distingue par des scores bas dans les dimensions Politique et Éducation, mais des scores élevés en Anti-discrimination. Le modèle MMN capture efficacement les complexités des politiques d'intégration

des migrants, offrant une compréhension nuancée des paysages politiques dans différents pays.