Clustering longitudinal ordinal data

Julien JACQUES & Francesco AMATO
Université Lyon 2, ERIC lab.

Brescia, May 2023

My works on ordinal data

- modle-based clustering of ordinal data
  - *C. Biernacki and J. Jacques (2016), Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm, Statistics and Computing, 26 [5], 929-943.*
- co-clustering of ordinal data
  - *J. Jacques and C. Biernacki (2018), Model-based co-clustering for ordinal data, Computational Statistics and Data Analysis, 123, 101-115.*
  - *M. Selosse, J. Jacques, C. Biernacki and F. Cousson-Gélie (2019). Analyzing health quality survey using constrained co-clustering model for ordinal data and some dynamic implication, Journal of the Royal Statistical Society, Series C, 68 [5], 1327-1349.*
- R package for classification, clustering and co-clustering of ordinal data
  - *M. Selosse, J. Jacques and C. Biernacki (2020). ordinalClust: a package for analyzing ordinal data, R journal, 12[2], 173-188.*
- regression with ordinal response and functional inputs
  - *J. Jacques, S. Samardzic (2022). Analyzing cycling sensors data through ordinal logistic regression with functional covariates. Journal of the Royal Statistical Society, Series C, 71[4], 969-986.*

Motivating data

Motivating study

- study about eating behaviors in France during the Covid-19 pandemic
- what is the impact of lockdown in terms of sustainable food?
- longitudinal study from March, 2020 to March, 2021

François-Lecompte A., Innocent M., Kréziak D., Prim-Allaz I. (2020), Confinement et comportements alimentaires : Quelles évolutions en matière d'alimentation durable ?, Revue Française de Gestion, 293/8, 55-80.

Motivating study

The survey consists of 55-78 questions, among which:

- *In the last month, you would say that you have preferred in your purchases seasonal products*:
  ☐ *much less than before lockdown*
  ☐ *less than before lockdown*
  ☐ *a little less than before lockdown*
  ☐ *as before lockdown*
  ☐ *a little more much less than before lockdown*
  ☐ *more than before lockdown*
  ☐ *much more than before lockdown*
- *About the foods, you have the impression of wasting*
- *You have paid attention to the expiration dates*
- ...

Motivating study

- *This period is ideal to rethink our way of consuming* :
  ☐ *high disagreement*
  ☐ *disagreement*
  ☐ *low disagreement*
  ☐ *neutral*
  ☐ *low agreement*
  ☐ *agreement*
  ☐ *high agreementt*
- *This period is ideal to test more environmentally responsible ways of living*
- . . .

A longitudinal study

The surveys has been conduced at 5 times:

- March 26 - April 5, 2020: beginning of the 1st lockdown
- April 30 - May 11, 2020: end of the 1st lockdown
- June 9 - June 16, 2020: post-lockdown
- October 28 - November 9, 2020: beginning of the 2nd lockdown
- March 5 - March 25, 2021: just before the 3rd lockdown

Number of participants: from 724 (for the 1st survey) to 337 (who answered to the 5 surveys)

Number of questions: from 78 (for the 1st survey) to 55 (for the 5th survey)

Motivating question

Extract typical consumption behavior of French people during pandemic, and in particular how these behaviors **have evolved**

Motivating question

Extract typical consumption behavior of French people during pandemic, and in particular how these behaviors **have evolved**

We are faced to a **clustering** question, with:

- ordinal variables
- repeated measurements along time

We need a clustering method for **ordinal longitudinal data**

Related work

Ordinal data

- Ordinal data occur when the categories are ordered
- Ordinality is a characteristic of the meaning of measurements [Stevens, 1946]
- Distinct levels of an ordinal variable differ in degree of dissimilarity

S. S. Stevens. "On the Theory of Scales of Measurement". In: Science 103.2684 (June 1946), pp. 677–680

Bad practices with ordinal data

- They are often **transformed into quantitative data** (Likert scale)
  $\Rightarrow$ introduces an *artificial* notion of distance between categories
  $\Rightarrow$ could lead to bias in the analysis
- Sometimes, they are considered as **nominal categoridal data**
  $\Rightarrow$ lost of order information

Ordinal data modelling and clustering (1/2)

McParland & Gormley, 2011; Ranalli & Rocci, 2016:

- ordinal variables are viewed as a **discretization of Gaussian** latent variables
- clustering: Gaussian Mixture Model (GMM)

Giordan and Diana, 2011; Jollois and Nadif, 2009:

- ordinal variables are assumed to arise from a **constrained multinomial** distribution,
- constraints are imposed to respect the ordinal properties : unimodal distribution with decrease of the probabilities around the mode
- clustering: Mixture Model with the constrained multinomial

Ordinal data modelling and clustering (2/2)

D'Elia and Piccolo, 2005; Piccolo & Simone 2019; ...:

- **define a distribution for ordinal data**: the CUB model
- CUB model: mixture of Binomial and Uniform, to reflect respondent choice and uncertainty
- clustering: ?

Biernacki and Jacques, 2016

- **define a distribution for ordinal data**: the BOS model
- BOS model: parametric distribution with position and precision parameters
- clustering: mixture of BOS models
- extension to co-clustering (Selosse et al., 2020)

Longitudinal data clustering (1/3)

Mc Nicholas & Murphy, 2010:

- vector of repeated observations modellized by a Gaussians,
- covariance matrix decomposition in term expressing time dependence (modified Choleski decomposition)
- clustering: GMM
- adapted only for univariate data

Cagnone et al., 2018; Komárek et al., 2014

- consider Generalized Linear Model
- need covariates (without covariates such models are equivalent to multinomial ones)

Vávra et al., 2021:

- binary, ordinal and continuous variables are assumed to come from latent Gaussian variables
- clustering: GMM
- may take into account covariate

Longitudinal data clustering (2/3)

Recent approaches consider longitudinal data as **three-way data**, where:

$$y_{i,j,t}$$

is the observation of:

- variable j
- at time t
- for individual i

Modelling can be done using **matrix-variate distributions**

Longitudinal data clustering (3/3)

Matrix-variate distribution approaches:

- Mixture of Matrix Normal distributions (MMN): Virolli, 2011, 2012;
- Mixture of non-normal skewed distribution: Dogru et al. 2016; Gallaugher et al., 2018; Melnykov et al., 2018, 2019

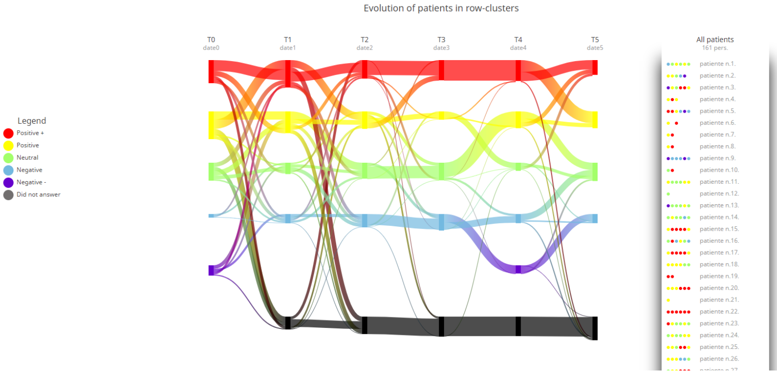
Advantages of matrix-variate approaches:

- parsimony of modelling without conditional independence assumption
- interpretability

Longitudinal ordinal data clustering

In Selosse et al. (2019):

- independent ordinal data clustering using the BOS model are performed at each time
- path of individual among the clusters are a posteriori studied



M. Selosse, J. Jacques, C. Biernacki and F. Cousson-Gélie (2019). Analyzing health quality survey using constrained co-clustering model for ordinal data and some dynamic implication, Journal of the Royal Statistical Society, Series C, 68 [5], 1327-1349.

Our idea

- consider ordinal variables as a **discretization of Gaussian** latent variables
- consider **Mixture of Matrix Normal** distribution for the latent variables

$\Rightarrow$ **Mixture of Ordinal Matrices** model: **MOM**

The MOM model

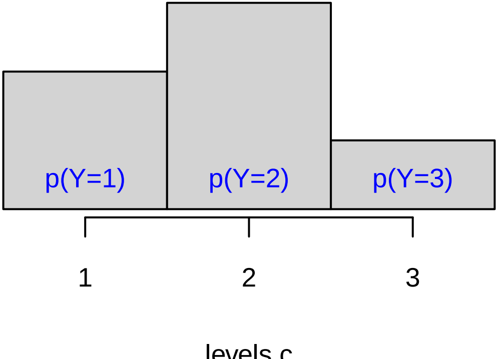
Ordinal distribution

- Y is an ordinal variable taking value in a set of C ordinal levels, coded $\{1, 2, \ldots, C\}$
- The distribution of Y is defined by

$$p(Y = c)$$

for any $c \in \{1, 2, \ldots, C\}$

Representation for $C = 3$ levels:



Latent Gaussian assumption

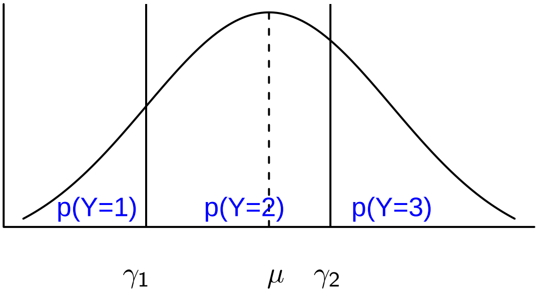
- Hyp. 1: each ordinal variable Y is the manifestation of an underlying latent continuous variable Z:

$$Y = c \quad \text{if} \quad \gamma_{c-1} < Z < \gamma_c$$

where $-\infty = \gamma_0 \leq \gamma_1 \leq ... \leq \gamma_C = \infty$ are some thresholds.

- Hyp. 2: $Z \sim \mathcal{N}(\mu, \sigma^2)$

Representation for $C = 3$ levels:



Latent Gaussian assumption

The ordinal distribution of C is thus defined by:

- the parameters of the Gaussian: μ, σ^2
- the thresholds $\gamma_1, \ldots, \gamma_{C-1}$

Latent Gaussian assumption

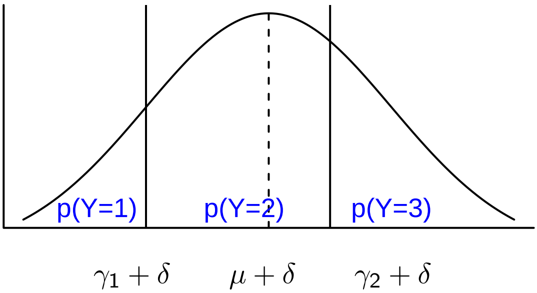
The ordinal distribution of C is thus defined by:

- the parameters of the Gaussian: μ, σ^2
- the thresholds $\gamma_1, \ldots, \gamma_{C-1}$

These parameters are not identifiable:

- adding any constant δ to μ and $\gamma_1, \ldots, \gamma_{C-1}$ does not change the distribution of Y

Representation for $C = 3$ levels:

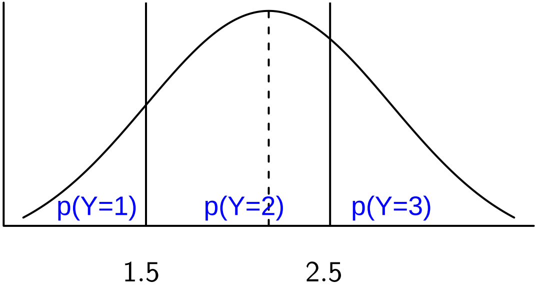


Latent Gaussian assumption

In order to fix identifiability, we choose to fix the γ_c:

$$\{\gamma_1, \gamma_2, \dots, \gamma_{C-1}\} = \{1.5, 2.5, \dots, C - .5\}$$

Representation for $C = 3$ levels:



Three-ways data

Let's go back to our 3-ways data:

- for each individual, we observe a $J \times T$ matrix of ordinal data:

$$y_i = (y_{i,j,t})_{j,t}$$

Notations:

- $\mathcal{O}^{J\times T}$: set of ordinal $J \times T$-matrices, in which row j takes values in $\{1,\ldots,C_j\}$.
- $R = \#\mathcal{O}^{J\times T}$
- each $\tilde{Y}_r \in \mathcal{O}^{J\times T}$ is generated by a portion Ω_r of the latent space $\mathbb{R}^{J\times T}$ according to thresholds $\gamma := \{\gamma_j\}_{j=1}^{J}$
- $\tilde{Y}_i = (\tilde{Y}_{i1},\ldots,\tilde{Y}_{iR})$: one-hot encoding of $\tilde{Y}_i$, s.t. if the r-th pattern is observed then $\tilde{Y}_{ir} = 1$ and any other entry in the vector equals zero.

Latent matrix normal distribution

Let's go back to our 3-ways data:

- for each individual, we observe a $J \times T$ matrix of ordinal data:

$$y_i = (y_{i,j,t})_{j,t}$$

- each y_i is assumed to be the realization of a **ordinal random matrix**:

$$Y$$

- itself coming from an underlying **continuous random matrix** Z distributed according to a **matrix normal distribution**

$$Z \sim \mathcal{MN}_{(J \times T)}(M, \Phi, \Sigma)$$

Matrix Normal distribution

Parameters of $\mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$ are:

- $M \in \mathbb{R}^{J\times T}$: matrix of means,
- $\Phi \in \mathbb{R}^{T\times T}$: covariances between the T times
- $\Sigma \in \mathbb{R}^{J\times J}$: covariances between the J variables

The p.d.f. $f(Z|M, \Phi, \Sigma)$ is:

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

Matrix normal vs multivariate normal distribution

Matrix Normal distribution is a specific multivariate Normal distribution :

$$Z \sim \mathcal{MN}_{(J \times T)}(M, \Phi, \Sigma) \Leftrightarrow \text{vec}(Z) \sim \mathcal{N}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$$

where:

- vec(.) is the vectorization operator
- $\otimes$ is 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**.

Matrix normal vs multivariate normal distribution

Advantage of the Matrix normal distribution:

- **interpretability**:
  - Φ express the time dependence
  - Σ express the dependence between variables
- **parsimony**:
  - $\Phi \otimes \Sigma$ has $J(J+1)/2 + T(T+1)/2$ parameters
  - a full covariance matrix of size JT has $JT(JT+1)/2$ parameters
  - ex: $J = T = 5$: 30 versus 325 parameters

Model-based clustering

In presence of an heterogeneous data set of matrix variate data $(y_i)_i$, we assume that they are realizations of an matrix ordinal variable Y coming from a latent continuous Z issued from a finite **Mixture of Matrix-Normals** (MMN) distribution:

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

Model-based clustering

In presence of an heterogeneous data set of matrix variate data $(y_i)_i$, we assume that they are realizations of an matrix ordinal variable Y coming from a latent continuous Z issued from a finite **Mixture of Matrix-Normals** (MMN) distribution:

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

Let introduce $\ell_i \in \{0,1\}^K$ s.t. $\ell_{ik} = 1$ is y_i belong to cluster k.

Model-based clustering

The generative process is then:

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

Note that the last step is not stochastic since only one elements of $\boldsymbol{\xi}_i$ is equal to 1.

Model-based clustering

The joint density of (Y_i^R, Z_i, ℓ_i) is then:

$$f(Y_i^R, Z_i, \ell_i) = f(Y_i^R | Z_i, \ell_i) f(Z_i | \ell_i) f(\ell_i).$$

with

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

Model-based clustering

The joint density of (Y_i^R, Z_i, ℓ_i) is then:

$$f(Y_i^R, Z_i, \ell_i) = f(Y_i^R | Z_i, \ell_i) f(Z_i|\ell_i) f(\ell_i).$$

with

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

The parameter to estimate are:

$$\boldsymbol{\Theta} := \{\pi_k, M_K, \Phi_k, \Sigma_k\}_{k=1}^{K}$$

Inference

Maximum likelihood estimation

We have to estimate:

$$\boldsymbol{\Theta} := \{\pi_k, M_K, \Phi_k, \Sigma_k\}_{k=1}^{K}$$

from the observed data:

$$\tilde{\mathbf{Y}} := \{\tilde{Y}_i\}_{n=1}^{N}$$

in presence of **latent variables**:

$$\boldsymbol{Z} := \{Z_i\}_{i=1}^{N}, \qquad \text{and} \qquad \boldsymbol{\ell} := \{\ell_i\}_{i=1}^{N}$$

Maximum likelihood estimation

We have to estimate:

$$\boldsymbol{\Theta} := \{\pi_k, M_K, \Phi_k, \Sigma_k\}_{k=1}^{K}$$

from the observed data:

$$\tilde{\mathbf{Y}} := \{\tilde{Y}_i\}_{n=1}^{N}$$

in presence of **latent variables**:

$$\boldsymbol{Z} := \{Z_i\}_{i=1}^{N}, \qquad \text{and} \qquad \boldsymbol{\ell} := \{\ell_i\}_{i=1}^{N}$$

$\Rightarrow$ **EM algorithm**

EM algorithm

Starting from an initialization $\boldsymbol{\Theta}^{(0)}$, the **EM algorithm is an iterative algorithm** which alternates

- the E step (*Expectation*): compute

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

- the M step (*Maximisation*): compute

$$\boldsymbol{\Theta}^{(s+1)} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(s)})$$

until convergence of the log-likelihood.

EM algorithm - E step

The complete log-likelihood $\log \mathcal{L}(\mathbf{\Theta}; \tilde{\mathbf{Y}}, \mathbf{Z}, \ell)$ is:

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

EM algorithm - E step

Computing $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(s)})$ requires to compute:

- $\mathbb{E}(\ell_{ik}|\tilde{Y}_{ir}=1, \hat{\boldsymbol{\Theta}}^{(s)}) = \frac{\pi_k^{(s)} \int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ}{\sum_{k=1}^{K} \pi_k^{(s)} \int_{\Omega_r} f(Z|\Theta_k^{(s)})dZ} =: \tau_{ik}^{(s+1)}$
- $\mathbb{E}(z_i|\ell_{ik}=1, \tilde{Y}_{ir}=1, \hat{\boldsymbol{\Theta}}^{(s)}) =: m_{ik}^{(s+1)}$
- $\mathbb{E}(z_i z_i^{\mathsf{T}}|\ell_{ik}=1, \tilde{Y}_{ir}=1, \hat{\boldsymbol{\Theta}}^{(s)}) =: S_{ik}^{(s+1)}$

The terms involving z_i requires to compute the moments of a truncated matrix-variate Gaussian, what is a complex task. In order to avoid it, a **Gibbs sampler** is considered.

Note that we work the vectorisation version of z_i for practical reasons

EM algorithm - M step

All the updates of the M step are explicit:

$$
\begin{aligned}
\hat{\pi}_k^{(s+1)} &= \frac{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)}}{N} \\
\hat{M}_k^{(s+1)} &= \frac{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)} \hat{M}_{ik}^{(s+1)}}{\sum_{i=1}^{N} \hat{\tau}_{ik}^{(s+1)}} \\
\hat{\Sigma}_k^{(s+1)} &= \frac{\sum_{i=1}^{N} \tau_{ik}^{(s+1)} [D_{ik}^{(s+1)} - \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} M_{ik}^{\mathsf{T}(s+1)} - M_{ik}^{(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\mathsf{T}(s+1)} + \hat{M}_k^{(s+1)} \hat{\Phi}_k^{-1(s)} \hat{M}_k^{\mathsf{T}(s+1)}]}{T \sum_{i=1}^{N} \tau_{ik}^{(s+1)}} \\
\hat{\Phi}_k^{(s+1)} &= \frac{\sum_{i=1}^{N} \tau_{ik}^{(s+1)} [C_{ik}^{(s+1)} - \hat{M}_k^{\mathsf{T}(s+1)} \hat{\Sigma}_k^{-1(s+1)} M_{ik}^{(s+1)} - M_{ik}^{\mathsf{T}(s+1)} \Sigma_k^{-1(s+1)} \hat{M}_k^{(s+1)} + \hat{M}_k^{\mathsf{T}(s+1)} \hat{\Sigma}_k^{-1(s+1)} \hat{M}_k^{(s+}}{J \sum_{i=1}^{N} \tau_{ik}^{(s+1)}}
\end{aligned}
$$

Initialization

The initialization $\mathbf{\Theta}^{(0)}$ can be:

- multiple random initialization
- using kmeans++, applied on the vectorized version of the data

Model selection

The number of cluster K is selected by minimizing the BIC criterion

$$BIC_k = -2 \log \mathcal{L}(\mathbf{\Theta}; \tilde{\mathbf{Y}}) + \nu \log N$$

where ν is the number of model parameters:

$$\nu = K - 1 + K(JT) + KJ(J+1)/2 + KT(T+1)/2$$

Numerical study on simulated data

Numerical study on simulated data

Goals:

- check that parameter estimation is consistent with N
- compare the different initialization strategies
- robustness to noise
- evaluate the efficiency of BIC to choose K
- comparison with competitors (*continuous* model)

Simulation setup

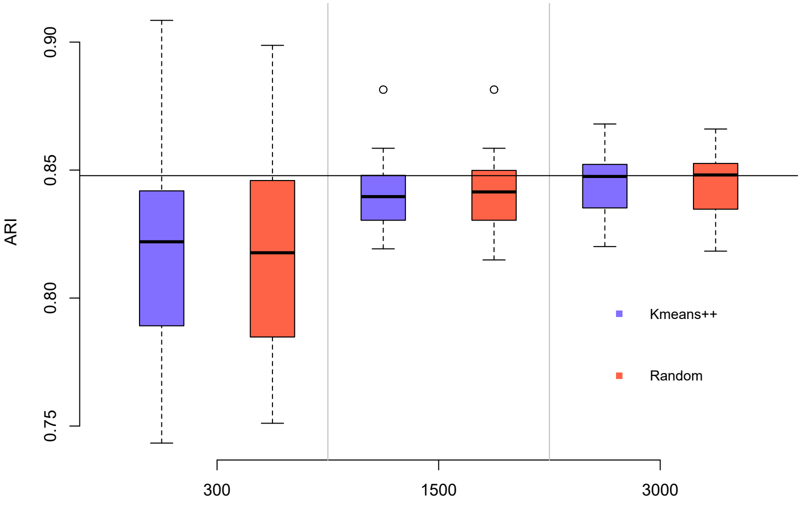
- 20 data sets simulated according to the MOM model:
  - $K = 3$, $J = T = 5$, $C_j = 5$, $\pi = (.3, .4, .3)$
  - each sample is drawn from a matrix-variate Gaussian (with M_1, M_2 and M_3 constant matrix of resp. 1.75, 2.5 and 3.25, and identity covariance matrices) and then discretized using $\gamma = (1.5, 2.5, 3.5, 4.5)$
- $N \in \{300, 1500, 3000\}$
- in each data set, a proportion of noise is added using a uniform distribution over the levels: 0% (scenario 1),10% (scenario 2), 20% (scenario 3)

Indicators

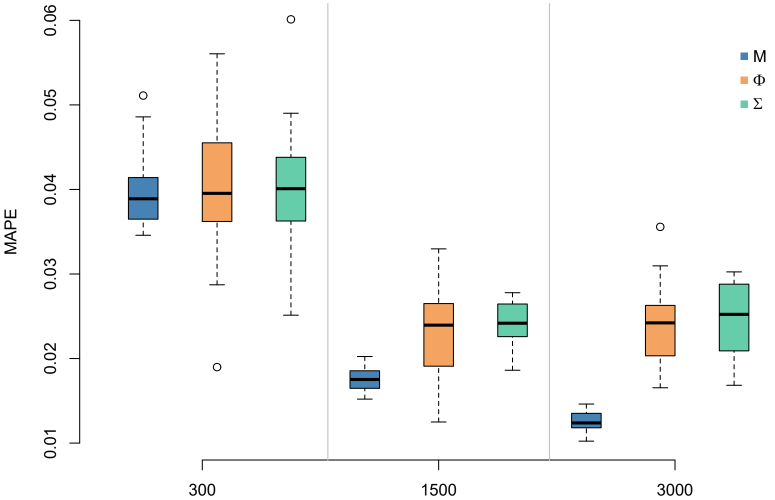
Efficiency is evaluated thanks to:

- Adjusted Rand Index (ARI) between the estimated partition and the actual one
- MAPE between the estimated parameters and the actual ones

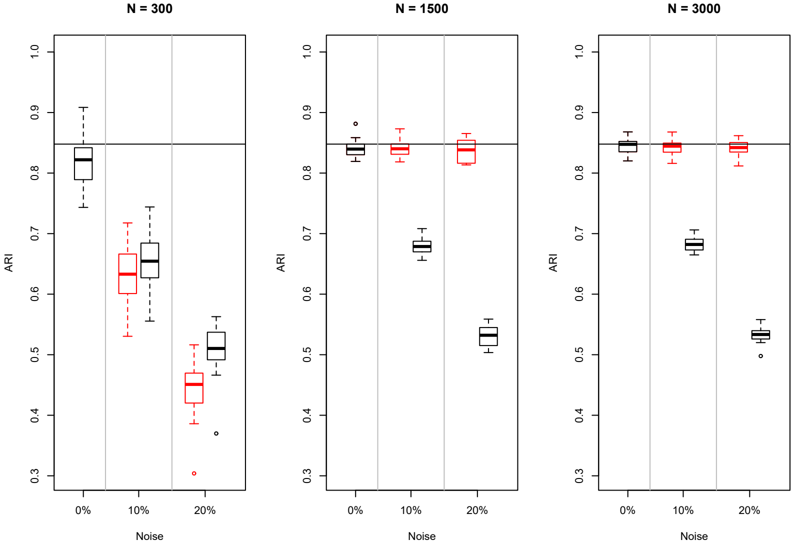
Influence of initialization and sample size



Influence of initialization and sample size



Robustness to noise



Model selection

Number of choice of K among $\{1, \ldots, 6\}$

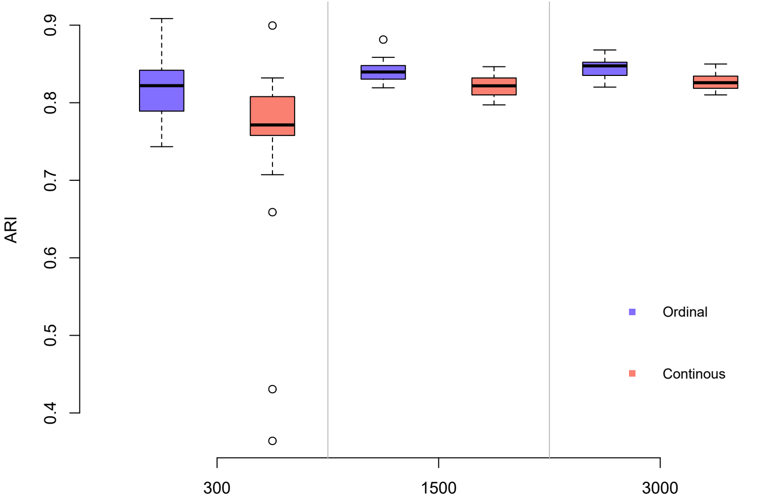
Scenario $\tau = 0$						
N/K	1	2	3	4	5	6
300	0	14	6	0	0	0
1500	0	0	20	0	0	0
3000	0	0	20	0	0	0

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

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

Comparison with competitors

Our MOM model is compare to the Mixture of Matrix Normal distribution applied on levels $\{1, \ldots, 5\}$ as they were continuous numbers in $\mathbb{R}$



Real data application

The data

- study about eating behaviors in France during the Covid-19 pandemic
- what is the impact of lockdown in terms of sustainable food?
- longitudinal study from March, 2020 to March, 2021

For our analysis: - a subset of $J = 11$ questions is considered - $N = 337$ individuals have answered to each of the $T = 5$ surveys - each questions has ordinal answer on $C_j = 7$ levels

François-Lecompte A., Innocent M., Kréziak D., Prim-Allaz I. (2020), Confinement et comportements alimentaires : Quelles évolutions en matière d'alimentation durable ?, Revue Française de Gestion, 293/8, 55-80.

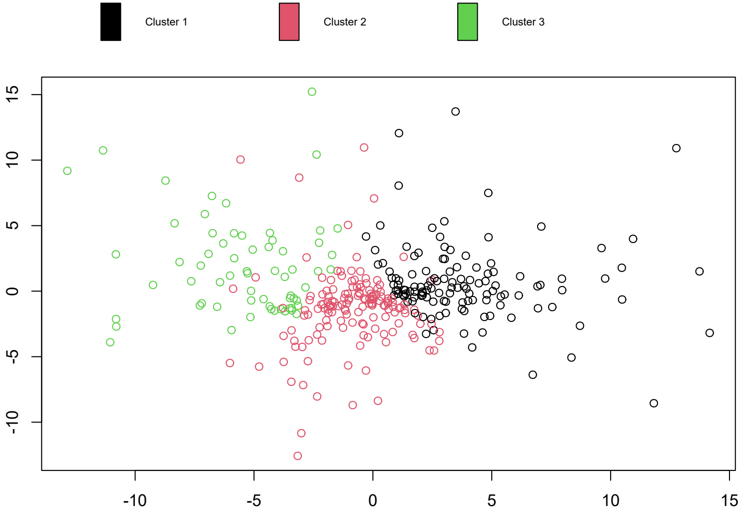
The 11 questions

- 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

Results

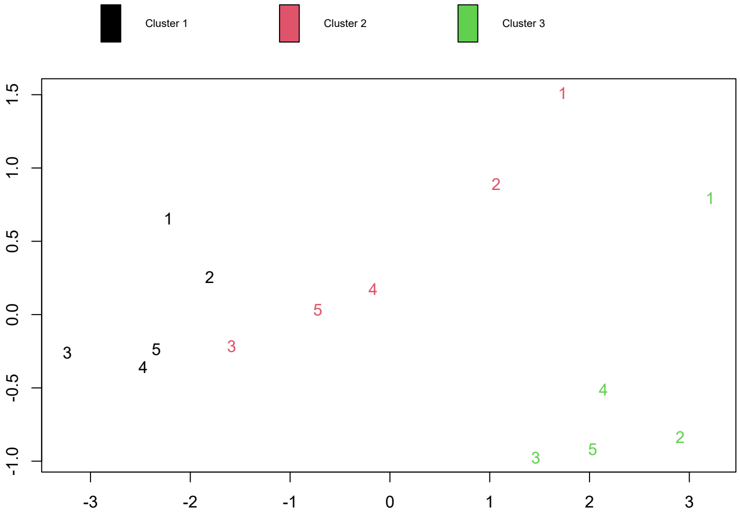
BIC selects $K = 3$ clusters (among $1, \ldots, 6$).

Representation of the 337 individuals (using isoMDS):



Results

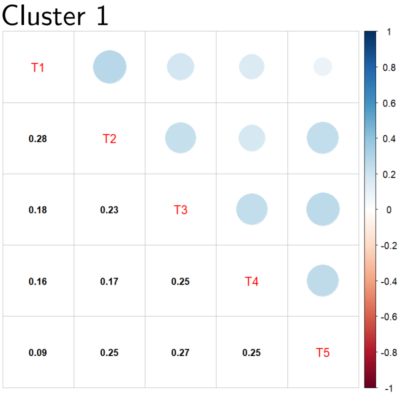
Time evolution of clusters means (isoMDS)

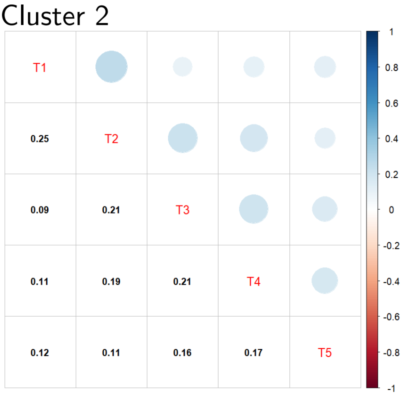


Cluster interpretation

- Cluster 1:
  - 124 units,
  - overall neutrality-level and stable means
  - **lockdown has no effect on them**
- Cluster 3:
  - 64 units,
  - overall neutrality-level for Q5 and Q8 macro-groups
  - high level for Q12 macro group ("rethinking-way-of-life" questions)
  - **lockdown has definitevely had an impact on them**
- Cluster 2:
  - 149 units,
  - intermediate between the Cluster 1 and Cluster 3,
  - close to Cluster 3 at the begining, and close to Cluster 1 at the end
  - **they want to change their life when they are on lockdown, but they quickly return to their usual habits once the confinement is over**

Time covariance matrices





Correlation between distant times is indeed higher for Cluster 1 than for Cluster 2.

Conclusion and future works

Conclusions and future works

Conclusions

- model-based **longitudinal clustering algorithm** for **ordinal data**
- respect the true nature of ordinal data
- parsimonious modelling
- nice interpretation properties (time and variable covariance matrice)
- R package under development
- preprint : https://hal.science/hal-04105669

Future works

- investigated more parsimonious models through covariance matrix reparametrization
- add non ordinal data to be able to cluster longitudinal mixed-type data