Clustering Longitudinal Ordinal Data

Francesco Amato, Julien Jacques

▶ **To cite this version:**

Francesco Amato, Julien Jacques. Clustering Longitudinal Ordinal Data. StatLearn, Apr 2022, Cargèse, France. hal-03657085

HAL Id: hal-03657085

https://hal.univ-lyon2.fr/hal-03657085

Submitted on 2 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Clustering Longitudinal Ordinal Data

F. Amato[1], J. Jacques[1]

[1] Univ Lyon, Univ Lyon 2, ERIC, Lyon.

Context

- Longitudinal ordinal data $y_{i,j,t}$ whose levels are coded $\{1,\dots,C_j\}$: the observation of the j-th variable for the i-th unit at time t $(i=1,\dots,N; j=1,\dots,J$ and $t=1,...,T)$.
- We want to **cluster units accounting for the temporal behavior**
- ⇒ 💡 Idea: rewrite them in a three-way format and use **latent underlying continous matrix-variate distributions**!
- We organize our data in a random-matrix form such that:

$$Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T} \end{pmatrix}$$

From continuous...

- Matrix-variate Normal: $Z \sim \mathcal{MN}_{(J\times T)}(M,\Phi,\Sigma)$, where
  - $M \in \mathbb{R}^{J\times T}$ is the matrix of means
  - $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix between the T occasions
  - $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix of the J variables

  The matrix-normal probability density function (pdf) is given by

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

- Mixtures of Matrix-Normals (MMN) were introduced by Viroli [3]

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

  where
  - K : number of mixture components
  - $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$: vector of mixing proportions, $\sum_{k=1}^K \pi_k = 1$
  - $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^K$: set of component-specific parameters $\Theta_k = \{M_k,\Phi_k,\Sigma_k\}$

$\Longrightarrow$ **Advantages**: offers a parsimonious and easily interpretable way to include the time dimension in the clustering.

...to ordinal data!

In the `clustMD` framework [2] cross-sectional mixed data are assumed to be all manifestation of underlying multivariate normals, and a Gaussian mixture model operating on the underlying normal variable is used to cluster them.

As for the classical `clustMD`, we can assume that each observed ordinal matrix Y is indeed the manifestation of a latent random matrix Z, which follows a matrix-normal distribution.

$$Z_i = \begin{pmatrix} z_{i,1,1} & \cdots & z_{i,1,t} & \cdots & z_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ z_{i,j,1} & \cdots & z_{i,j,t} & \cdots & z_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ z_{i,J,1} & \cdots & z_{i,J,t} & \cdots & z_{i,J,T} \end{pmatrix} \longrightarrow Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T} \end{pmatrix}$$

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

A key point is of course the choice of the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^J$. In [1], thresholds are fixed in advance to avoid identifiability and computational complexity issues. Also in [2], for ordinal variables they are fixed such that $\gamma_{j,c} = \varphi^{-1}(\delta_c)$, where δ_c is the proportion of variable J which are less than or equal to level c and φ is the normal cumulative distribution function.

Model

The model relies on the following hypotheses:

- $\boldsymbol{\ell}_i \in \{0,1\}^K$ is the latent allocation variable such that $\ell_{ik}=1$ if the i-th unit belongs to the k-th cluster.
- $\mathbf{Y}_i^R = (Y_{i1}^R,...,Y_{iR}^R) \in \mathbb{R}^R$ indicate the observed response pattern for the i-th unit.
- $\boldsymbol{\ell}_i \sim \mathcal{M}(1,\boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_1,...,\pi_K)$
- $\mathbf{Z}_i|\boldsymbol{\ell}_{ik}=1 \sim \mathcal{MN}_{(J\times T)}(\mathbf{Z}_i|\Theta_k)$, $\Theta_k = \{M_k,\Phi_k,\Sigma_k\}$
- $\mathbf{Y}_i^R|\mathbf{Z}_i,\boldsymbol{\ell}_{ik}=1 \sim \mathcal{M}(1,\boldsymbol{\xi}_i^R)$, $\boldsymbol{\xi}_i^R = (\mathbf{1}_{\Omega_1}(\mathbf{Z}_i),\dots,\mathbf{1}_{\Omega_R}(\mathbf{Z}_i))$

where $\mathcal{M}$ indicate the multinomial distribution, Ω_r is the portion of the $J\times T$-space which determines the the r-th pattern, and $\mathbf{1}_{\Omega_r}(\mathbf{Z}_i)$ is the indicator function that equals 1 when the elements in $\mathbf{Z}_i$ have values that determine the r-th pattern. Of course, when $\mathbf{Z}_i$ is given, the the value of $\mathbf{Y}_i$ is no more random.

We can derive the joint density of $\mathbf{Z}_i, \mathbf{Y}_i^R, \boldsymbol{\ell}_i$ as:

$$f(\mathbf{Y}_i^R,\mathbf{Z}_i,\boldsymbol{\ell}_i) = f(\mathbf{Y}_i^R|\mathbf{Z}_i,\boldsymbol{\ell}_i)f(\mathbf{Z}_i|\boldsymbol{\ell}_i)f(\boldsymbol{\ell}_i),$$

where:

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

Model Inference

Due to the presence of latent variables, the maximization of the likelihood cannot be done in "close form", and we must then use an EM algorithm, which maximizes a lower limit of the log-likelihood: the complete log-likelihood. We can write the complete log-likelihood as

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

- E-step: we compute the expectation of the complete log-likelihood with respect to the latent data Z and the cluster labels ℓ. For each response pattern r, **we can approximate the value of $\mathbf{Z}_i|\boldsymbol{\ell}_i$ as the expected value of the truncated multivariate normals (using properties of matrix-variate normals)**, given the parameters Θ_k of the assigned cluster.

  The latent variable $\boldsymbol{\ell}_i$ can be computed by means of Bayes' theorem as:

$$\mathbb{E}(\ell_{ik}^{(s)}|Y_i^R = r,\boldsymbol{\Theta}^{(s-1)},\boldsymbol{\pi}^{(s-1)}) = \frac{\pi_k^{(s-1)}\int_{\Omega_r} f(Z|\Theta_k^{(s-1)})dZ}{\sum_{k=1}^K \pi_k^{(s-1)}\int_{\Omega_r} f(Z|\Theta_k^{(s-1)})dZ},$$

  which would require Monte-Carlo approximation.
- M-step: the parameter updates are given by:

$$\hat{\Sigma}_k^{(s)} = \frac{\sum_{i=1}^N \ell_{ik}^{(s)}(Z_i-\hat{M}_k^{(s)})\hat{\Phi}_k^{-1(s-1)}(Z_i-\hat{M}_k^{(s)})^\intercal}{T\sum_{i=1}^N \ell_{ik}^{(s)}} \quad , \quad \hat{M}_k^{(s)} = \frac{\sum_{i=1}^N \ell_{ik}^{(s)} Z_i}{\sum_{i=1}^N \ell_{ik}^{(s)}}$$

$$\hat{\Phi}_k^{(s)} = \frac{\sum_{i=1}^N \ell_{ik}^{(s)}(Z_i-\hat{M}_k^{(s)})^\intercal\hat{\Sigma}_k^{-1(s)}(Z_i-\hat{M}_k^{(s)})}{J\sum_{i=1}^N \ell_{ik}^{(s)}} \quad , \quad \hat{\pi}_k^{(s)} = \frac{\sum_{i=1}^N \ell_{ik}^{(s)}}{N}$$

The E and M step are iterated until convergence of the log-likelihood.

A longitudinal `clustMD`? 🧐

This is the first step of a broader project, aiming at extending this framework to account for mixed data (continuous, ordinal, nominal, count) in order to cluster mixed longitudinal dataset.

References

[1] M. Corneli, C. Bouveyron, and P. Latouche. Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. *Journal of Computational and Graphical Statistics*, 2020.

[2] D. McParland and I. C. Gormley. Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10(2):155–169, 2016.

[3] C. Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4):511–522, 2011.