CLUSTERING LONGITUDINAL ORDINAL DATA

Julien Jacques [1] and Francesco Amato [1]

[1] Univ Lyon, Univ Lyon 2, ERIC, Lyon, (e-mail: julien.jacques@univ-lyon2.fr, francesco.amato@univ-lyon2.fr)

ABSTRACT: In social sciences, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model-based clustering algorithm for such longitudinal data. Assuming that an ordinal variable is the discretization of a underlying latent continuous variable, the model relies on a mixture of matrix-variate normal distributions, accounting simultaneously for within- and between-time dependence structures. An EM algorithm is considered for parameter estimation. An evaluation of the model through synthetic data show its estimation abilities and its advantages when compared to competitors. A real-world application concerning preferences for grocery shopping during the Covid-19 pandemic period in France will be presented.

KEYWORDS: Ordinal data, longitudinal data, clustering, matrix variate distribution, EM algorithm

1 The data

Let denote by $y_{i,j,t}$ the observation of the j-th ordinal variable for the i-th unit at time t ($i = 1,...,N$; $j = 1,..,J$ and $t = 1,...,T$). The categories of the j-th ordinal variable are quoted by 1 to C_j. The data are organized in a random-matrix form such that $\mathbf{Y} = \{Y_i\}_{i=1}^N$ is a sample of $J \times T$-variate matrix observations:

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

2 Latent Gaussian distribution for ordinal variable

We assume that each variable $y_{i,j,t}$ is the manifestation of an underlying latent continuous variable $z_{i,j,t}$ which follows a Gaussian distribution. At this point, we can assume that each observed ordinal matrix Y_i is indeed the manifestation of a latent continuous random matrix $Z_i \in \mathbb{R}^{J\times T}$, which follows a matrix-normal distribution $\mathcal{MN}_{(J\times T)}(M,\Phi,\Sigma)$, where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix 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

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

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

3 Model-based clustering

When data are heterogeneous, mixture model is an efficient way to perform clustering. In the present case, we consider Mixture of Matrix-Normals (MMN, Viroli, 2011). As usually for mixture models, parameter estimation is done using an EM algorithm. The number of cluster is selected using the BIC criterion.

4 Applications

An evaluation of the model through synthetic data show its estimation abilities and its advantages when compared to competitors. A real-world application concerning preferences for grocery shopping during the Covid-19 pandemic period in France will be presented.

References

VIROLI, CINZIA. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.