Sparse matrix-variate model-based clustering via penalized estimation

Working Group on Model-Based Clustering

Alessandro Casa

Joint work with: Andrea Cappozzo & Michael Fop

School of Mathematics and Statistics

University College Dublin

alessandro.casa@ucd.ie

Athens - 26th October 2021

Framework

- Matrix Gaussian Mixture Model (MGMM) provides a probabilistic approach to cluster matrix-variate (three-way) data
- Let $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be a set of n matrices with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$. MGMM expresses the marginal density for each $\mathbf{X}_i$ as

$$f(\mathbf{X}_i, \Theta) = \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; \mathbf{M}_k, \Omega_k, \Gamma_k)$$

  - $\phi_{p\times q}(\cdot, \mathbf{M}_k, \Omega_k, \Gamma_k)$, $p \times q$ matrix normal distribution
  - τ_k's, mixing proportions $\tau_k > 0, k = 1, \ldots, K, \sum_k \tau_k = 1$
  - $\mathbf{M}_k \in \mathbb{R}^{p\times q}$ mean matrix, $\Omega_k \in \mathbb{R}^{p\times p}$ and $\Gamma_k \in \mathbb{R}^{q\times q}$ rows and columns precision matrices

Limitation

$|\Theta|$ scales quadratically with both p and q leading to overparameterization even in moderate dimensions

Possible solutions

- Solutions proposed introduce a rigid way to induce parsimony
  $\longrightarrow$ association structures are constant across groups
- We adopt a penalized likelihood approach by maximizing

$$\ell(\Theta, \mathbf{X}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \tau_k \phi_{p \times q}(\mathbf{X}_i, \mathbf{M}_k, \Omega_k, \Gamma_k) - p_\lambda(\Theta)$$

  - $p_\lambda(\Theta)$, penalty term to be defined
  - $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, vector of penalty coefficients
- Main assumption: the matrices involved possess their own cluster-dependent degrees of sparsity

$$\downarrow$$

reduced number of parameters, cluster-wise conditional independence patterns eases the interpretation

Proposal and future directions

- Two different specifications for $p_\lambda(\Theta)$

$$1)\quad \sum_{k=1}^{K} \lambda_1 ||\mathbf{P}_1 * \mathbf{M}_k||_1 + \sum_{k=1}^{K} \lambda_2 ||\mathbf{P}_2 * \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||\mathbf{P}_3 * \Gamma_k||_1$$

$$2)\quad \sum_{k=1}^{K} \lambda_1 \sum_{r=1}^{p} ||m_{r\cdot,k}||_2 + \sum_{k=1}^{K} \lambda_2 ||\mathbf{P}_2 * \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||\mathbf{P}_3 * \Gamma_k||_1$$

  with $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ matrices with non-negative entries, $m_{r\cdot,k}$ the r-th row of $\mathbf{M}_k$, $||A||_1 = \sum_{jh} |A_{jh}|$ and $||\cdot||_2$ the Euclidean norm
- Group lasso type penalty in 2) allows to perform variable selection in a matrix-variate framework
- **Open problem** $\longrightarrow$ we need to select $\lambda_1, \lambda_2, \lambda_3$ and K
  - Exhaustive search is computationally unfeasible
  - E-MS algorithm seems promising but any suggestion is more than welcome