Penalized model-based clustering for three-way data structures

Clustering penalizzato basato sul modello per dati con struttura tridimensionale

Andrea Cappozzo, Alessandro Casa, and Michael Fop

Abstract Recently, there has been an increasing interest in developing statistical methods able to find groups in matrix-valued data. To this extent, matrix Gaussian mixture models (MGMM) provide a natural extension to the popular model-based clustering based on Normal mixtures. Unfortunately, the overparametrization issue, already affecting the vector-variate framework, is further exacerbated when it comes to MGMM, since the number of parameters scales quadratically with both row and column dimensions. In order to overcome this limitation, the present paper introduces a sparse model-based clustering approach for three-way data structures. By means of penalized estimation, our methodology shrinks the estimates towards zero, achieving more stable and parsimonious clustering in high dimensional scenarios. An application to satellite images underlines the benefits of the proposed method.
Abstract *Ad oggi, c'è un sempre crescente interesse nello sviluppo di metodi statistici in grado di identificare gruppi in dati matriciali. I modelli mistura con kernel Gaussiano matriciale (MGMM) forniscono un'estensione naturale al clustering basato su modelli con misture normali multivariate. Sfortunatamente, il problema dell'eccessiva parametrizzazione, che già interessa il framework a due dimensioni, è particolarmente evidente nei MGMM, poiché il numero di parametri cresce all'aumentare sia del numero di righe che di colonne. Al fine di superare questa limitazione, l'articolo introduce un approccio di clustering basato su modelli sparsi per strutture di dati matriciali. Tramite la stima penalizzata, la nostra metodologia riduce il valore delle stime ottenendo un clustering più stabile e parsimonioso in scenari ad alta dimensione. Un'applicazione a immagini satellitari evidenzia i vantaggi del metodo proposto.*

Andrea Cappozzo,
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: andrea.cappozzo@unimib.it

Alessandro Casa, Michael Fop
School of Mathematics & Statistics, University College Dublin e-mail: alessandro.casa@ucd.ie, michael.fop@ucd.ie

Key words: Model based clustering, Matrix distribution, EM-algorithm, penalized likelihood, Sparse matrix estimation

1 Introduction and motivation

Model-based clustering is a mathematical-based approach to account for heterogeneity in a population, useful for discovering subgroups in data [2]. This framework assumes that each cluster corresponds to a different component of a finite mixture, with the Gaussian distribution being the standard choice when dealing with continuous data [4]. Nonetheless, the ever-increasing complexity of real-world datasets is jeopardizing the usage of standard Gaussian Mixture Models (GMM), as they tend to be over-parametrized in high-dimensional spaces [1]. To this extent, parameters regularization by means of penalized estimation has been proven useful in performing model-based clustering and variable selection in such scenarios [9].

The aforementioned problem complicates even further when dealing with three-way data structures, where for each statistical unit variables are repeatedly measured over different occasions. These increasingly common situations lead to a complex statistical framework, for which the observations are assumed to be realizations of some matrix-variate distribution. In details, for a given sample of n standardized matrices $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$, the GMM extension to the three-way data context is provided by the matrix normal mixture model (MGMM [7]), in which the marginal density of each $\mathbf{X}_i$ reads:

$$f(\mathbf{X}_i;\boldsymbol{\Theta}) = \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i;\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k). \tag{1}$$

The number of mixture components is denoted by K, τ_k's are the mixing proportions with $\tau_k > 0, \forall k = 1,\dots,K$, $\sum_{k=1}^{K} \tau_k = 1$ and $\phi_{p\times q}(\cdot;\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k)$ is the k-th component density of a $p \times q$ matrix normal distribution:

$$\begin{aligned}\phi_{p\times q}(\mathbf{X}_i;\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) = {}& (2\pi)^{-\frac{pq}{2}} |\boldsymbol{\Omega}_k|^{\frac{q}{2}} |\boldsymbol{\Gamma}_k|^{\frac{p}{2}} \\ & \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}_k(\mathbf{X}_i - \mathbf{M}_k)\boldsymbol{\Gamma}_k(\mathbf{X}_i - \mathbf{M}_k)')\right\},\end{aligned}$$

with $\mathbf{M}_k$ representing the mean matrix and $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$ are the rows and columns precision matrices with dimensions $p \times p$ and $q \times q$, respectively. The previously mentioned over-parametrization issue deeply affects model in (1), since the number of parameters $\boldsymbol{\Theta} = \{\tau_k, \mathbf{M}_k, \boldsymbol{\Psi}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ scales quadratically with both dimensions p and q. Motivated by this, the present manuscript proposes a matrix-variate extension to the two-way penalized model-based clustering framework introduced in [9], assuming that $\mathbf{M}_k$, $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$, $k = 1,\dots,K$, possess some degree of sparsity. The resulting model flexibly accounts for cluster-wise conditional independence patterns, providing a unified way for jointly reducing the number of estimated parameters and eliminating redundant variables, combining advantages of state-of-the-art procedures for matrix-variate clustering [5, 8].

The remainder of the paper proceeds as follows: in Section 2 we introduce our new proposal and we discuss its main methodological aspects. An application to soil classification by means of satellite images is reported in Section 3. Section 4 summarizes the novel contributions and highlights future research directions.

2 Penalized matrix-variate mixture model

A penalized likelihood approach is introduced for parameter estimation. The resulting objective function to be maximized with respect to $\boldsymbol{\Theta}$ is:

$$\ell(\boldsymbol{\Theta};\mathbf{X}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i;\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) \right\} - p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k), \quad (2)$$

with the penalization term $p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k)$ being equal to

$$p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) = \sum_{k=1}^{K} \lambda_1 ||\mathbf{P}_1 * \mathbf{M}_k||_1 + \sum_{k=1}^{K} \lambda_2 ||\mathbf{P}_2 * \boldsymbol{\Omega}_k||_1 + \sum_{k=1}^{K} \lambda_3 ||\mathbf{P}_3 * \boldsymbol{\Gamma}_k||_1.$$

With $*$ we denote element-wise product, $\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3$ are matrices with non-negative entries, λ_1,λ_2 and λ_3 are penalty coefficients and $||\mathbf{A}||_1 = \sum_{jh}|A_{jh}|$. A dedicated EM-algorithm is devised for inference by firstly defining a suitable *penalized complete log-likelihood* for model (2):

$$\begin{aligned}\ell_C(\boldsymbol{\Theta};\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \Big[&\log \tau_k - \frac{pq}{2}\log 2\pi + \frac{q}{2}\log|\boldsymbol{\Omega}_k| + \frac{p}{2}\log|\boldsymbol{\Gamma}_k| + \\ &-\frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}_k(\mathbf{X}_i-\mathbf{M}_k)\boldsymbol{\Gamma}_k(\mathbf{X}_i-\mathbf{M}_k)'\right\}\Big] - p_{\lambda_1,\lambda_2,\lambda_3}(\mathbf{M}_k,\boldsymbol{\Omega}_k,\boldsymbol{\Gamma}_k) \quad (3)\end{aligned}$$

where as usual for mixture models $z_{ik} = 1$ if observation $\mathbf{X}_i$ belongs to the k-th component, and 0 otherwise. The E-step at the t-th iteration requires computing the estimated a posteriori probabilities of class membership $\hat{z}_{ik}^{(t)}$, achieved via the standard updating formula. On the other hand, the M-step involves a partial optimization strategy. Let us denote with m_{lsk}, x_{lsi}, ω_{lsk}, γ_{lsk} and p_{ls1} the element in the l-th row and s-th column of matrices $\mathbf{M}_k$, $\mathbf{X}_i$, $\boldsymbol{\Omega}_k$, $\boldsymbol{\Gamma}_k$ and $\mathbf{P}_1$. The sparse estimation of $\mathbf{M}_k$ is achieved via a cell-wise coordinate ascent algorithm, where $\hat{m}_{lsk}^{(t)} = 0$ if

$$\begin{aligned}
&\left|\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\sum_{\substack{r=1\\ r\neq l}}^{p} \hat{\omega}_{lrk}^{(t-1)} \left(\sum_{c=1}^{q} \left(x_{rci}-\hat{m}_{rck}^{(t)}\right)\hat{\gamma}_{csk}^{(t-1)}\right)+\right.\right.\\
&\qquad\left.\left. + \ \hat{\omega}_{llk}^{(t-1)} \left(\sum_{\substack{c=1\\ c\neq s}}^{q} \left(x_{lci}-\hat{m}_{lck}^{(t)}\right)\hat{\gamma}_{csk}^{(t-1)}\right) + \hat{\omega}_{llk}^{(t-1)} x_{lsi}\hat{\gamma}_{ssk}^{(t-1)}\right]\right| \leq \lambda_1 p_{ls1}, \qquad (4)
\end{aligned}$$

otherwise, $\hat{m}_{lsk}^{(t)}$ is obtained by solving

$$\begin{aligned}
\hat{n}_k^{(t)}\,\hat{\omega}_{llk}^{(t-1)}\hat{m}_{lsk}^{(t)}\hat{\gamma}_{ssk}^{(t-1)} + \lambda_1 p_{ls1}\,\mathrm{sign}\left(\hat{m}_{lsk}^{(t)}\right) &= \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}\sum_{r=1}^{p}\sum_{c=1}^{q}\hat{\omega}_{lrk}^{(t-1)}x_{rci}\hat{\gamma}_{csk}^{(t-1)}+\\
&\quad - \hat{n}_k^{(t)}\left(\sum_{\substack{r=1\\ r\neq l}}^{p}\sum_{\substack{c=1\\ c\neq s}}^{q}\hat{\omega}_{lrk}^{(t-1)}\hat{m}_{rck}^{(t)}\hat{\gamma}_{csk}^{(t-1)}\right) \qquad (5)
\end{aligned}$$

with respect to $\hat{m}_{lsk}^{(t)}$, where $\hat{n}_k^{(t)} = \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}$. Lastly, expressions for estimating sparse precision matrices $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$ rely on dedicated modifications of the the coordinate descent graphical LASSO [3].

3 Application to Satellite Data

The considered data encompass 397, 211 and 237 satellite images of respectively grey soil, damp grey soil and soil with vegetation stubble. Each scene (represented by $p = 9$ pixels) is recorded $q = 4$ times with different spectral bands, resulting in $n = 845$ samples of 4×9 matrices. The methodology described in Section 2 is employed to perform clustering on this three-way dataset, mimicking the analyses performed in [5, 7]. Table 1 reports the classification error rate and number of estimated parameters for our method and two competing procedures, namely constrained MGMM [5] and standard GMM (the original three-way samples are unfolded in order to recast the problem into a two-way framework). Our model not only succeeds in better retrieving the true underlying data partition, but it is also the most parsimonious, displaying the lowest number of non-zero estimated parameters. The resulting sparse structures retrieved by our proposal are showcased in

Table 1 Misclassification errors and number of free estimated parameters for three clustering procedures, Satellite Data. *Sparsemixmat* denotes the proposal introduced in the present paper.

	Sparsemixmat	Sarkar et al. [5]	Mclust [6]
Misclassification Error	0.0793	0.0828	0.3053
# of free parameters	218	275	850

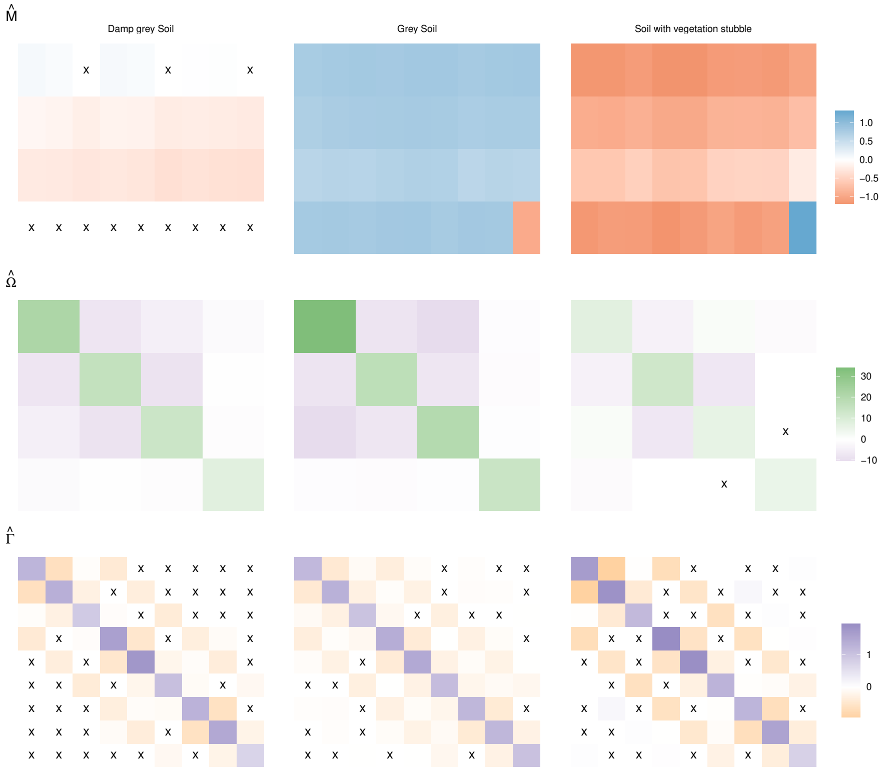

Fig. 1 Estimated sparse mean matrices (upper plots), row-precision matrices (middle plots) and column-precision matrices (lower plots) for the three clusters, satellite data. Matrix entries that are shrunk to 0 by the penalized estimator are highlighted with an $\times$.

Figure 1, where estimated parameters for the three soil types are displayed. Matrix entries that are shrunk to 0 by the penalized estimator are highlighted with an $\times$ in the plots. As expected, the clustering is mainly driven by the different patterns in the mean matrices, while the column-precision matrices $\hat{\boldsymbol{\Gamma}}_k$, $k = 1, \ldots, 3$ possess the highest level of sparsity.

4 Conclusion

The present work has introduced a novel penalized matrix-variate mixture model, able to capture heterogeneity and redundancy in three-way data structures. By means of sparse estimation, we are able to overcome the over-parametrization issue occuring in MGMM when either the row or the column dimensions increase.

Future research directions aim at deriving an efficient procedure for performing model selection. Jointly determining the best values for the penalty coefficients, as well as the number of mixture components define a challenging computational problem: feasible solutions are currently being investigated.

References

[1] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, 71:52–78, 2014.

[2] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science*, volume 50. Cambridge University Press, jul 2019.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, jul 2008.

[4] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004.

[5] S. Sarkar, X. Zhu, V. Melnykov, and S. Ingrassia. On parsimonious models for modeling matrix data. *Computational Statistics and Data Analysis*, 142:106822, 2020.

[6] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289–317, 2016.

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

[8] Y. Wang and V. Melnykov. On variable selection in matrix mixture modelling. *Stat*, 9(1):1–11, dec 2020.

[9] H. Zhou, W. Pan, and X. Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009.